CrossMark

Annual Review of Applied Linguistics, **36** (2016), pp. 73–97.
© Cambridge University Press, 2016
doi: 10.1017/S0267190516000015

Task-Based Learner Production: A Substantive and Methodological Review

Luke Plonsky
Institute of Education, University College London
l.plonsky@ucl.ac.uk

YouJin Kim
Georgia State University
ykim39@gsu.edu

ABSTRACT

Tasks are frequently used to elicit learner language in second language (L2) research. The purposes for doing so, however, vary widely, covering a range of theoretical models, designs, and analyses. For example, task-based researchers have examined a range of linguistic and interactional features (e.g., accuracy, language-related episodes) that are found in learner production and that vary as a function of task conditions (e.g., +/− complex), modes (oral, written, computer-mediated), and settings (second vs. foreign language). This article presents a synthesis of substantive interests and methodological practices in this area. We first collected a sample of 85 primary studies of task-based language production published from 2006 to 2015. Each study was then coded for the target features it analyzed as well as other contextual and demographic variables. We also coded for methodological features related to study designs, sampling, analyses, and reporting practices. The results indicate a strong preference toward analyses of grammar, vocabulary, accuracy, and different features of L2 interaction, and very little interest in task-induced pronunciation, pragmatics, and the quality of task performance. More fundamentally, this domain may be hindered by a lack of theoretical and operational consistency. The data also point to a number of concerns related to research and reporting practices (e.g., low statistical power; missing data). Based on our findings, we outline a number of pointed recommendations for future research in this domain.

INTRODUCTION

Analyses of learner language have always been central to second language acquisition (SLA) research. Indeed, Corder's (1967) and Selinker's (1972) seminal works, argued by many scholars to symbolize the genesis of the field (e.g., Ortega, 2009; but cf. Thomas, 2013), both regard learner production as fundamental. In the last two decades, tasks have been used by numerous researchers to elicit learner production. The interests of such studies range from testing theoretical

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

models (e.g., the cognition hypothesis) to exploring instructional treatments and developing novel assessments.

The goal of the current article is to review studies that analyze task-based learner language. We present a systematic review of the substantive foci addressed in this line of research. Specifically, among other issues and practices, we consider the extent to which different linguistic and interactional features of task-based learner production have been examined. We also assess a number of methodological features (i.e., research designs, data analyses, and reporting practices) found in the domain. In doing so, we connect our work to the growing body of methodological syntheses (e.g., Plonsky, 2013) and to the broader methodological reform movement taking place in applied linguistics (Byrnes, 2013).

LITERATURE REVIEW

Since Long (1985), tasks have been used in diverse empirical and pedagogical domains, serving two main purposes: as (a) a primary unit of instruction and (b) a tool for eliciting language production. The growing interest in this area has largely been motivated by the fact that tasks provide contexts for meaningful second language (L2) learning and use. Accordingly, tasks have become essential to both L2 research and pedagogy, and have linked the two domains (Mackey, 2007).

Tasks have been used as a tool to elicit learner production in two major types of designs. In some instances, the tasks and the language produced in carrying them out function as a treatment or intervention, the effects of which (i.e., L2 development) are made known on a posttest. We refer in the present study to this major design type as "task-based learning outcome" or "task-as-treatment" studies. Oftentimes, however, researchers collect samples of task-based learner production in order to understand the impact of task design or condition on individual features or more global assessments of learner language. We refer to this type of research as "task-based learner performance" studies.

These two major design types serve as the organizing principle for the first two of three sections of the literature review that follows. We focus throughout on the learner language induced in the context of tasks as defined by Bygate, Skehan, and Swain (2001): "A task is an activity which requires learners to use language, with emphasis on meaning, to attain an objective (p. 11)". In the third section, we describe previous reviews studies that, like ours, have engaged in methodological synthesis.

Task-Based Learner Performance Studies

Many studies of task-based production are interested in testing theoretical claims or hypotheses related to task design (e.g., task complexity, task types, task modality) and implementation (e.g., task planning, task repetition) as well as learner characteristics (e.g., proficiency, anxiety, task familiarity) during task performance. Such studies employ both monologic and dialogic tasks, and many different features of learner language and interaction are analyzed. For instance, a number of studies

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

have tested Robinson's (2001) cognition hypothesis and Skehan's (2009) trade-off hypothesis, both of which focus on the effect of task complexity on learners' linguistic performance and interactional competence (see Jackson & Suethanapornkul, 2013, for a review).

Whether they involve monologic or dialogic performance, most task-based learner production research has examined linguistic features of learner language. Some studies analyze individual grammatical and/or lexical items. For example, Michel (2013) was interested in the use of conjunctions by students who performed tasks at two levels of complexity. Other studies analyze broader performance-based measures within the complexity, accuracy, and fluency (CAF) framework (Housen, Kuiken, & Vedder, 2012; Michel, Kuiken, & Vedder, 2012). Complexity traditionally indicates L2 learners' ability to use a range of sophisticated structures and vocabulary in the L2 (Bulté & Housen, 2012; Ellis & Barkhuizen, 2005). Accuracy refers to learners' ability to produce target-like language (see Foster & Wigglesworth, this volume). And fluency is characterized by the rate and/or amount of production in a given time (Segalowitz, 2010). According to Housen et al. (2012), "CAF emerge as the primary epiphenomena of the psycholinguistic processes and mechanisms underlying the acquisition, representation and processing of L2 system" (p. 2). It is not surprising then that these constructs appear frequently in SLA, language assessment, and other areas of applied linguistics.

Despite their frequent application and intuitive appeal, there is little consensus surrounding measures of CAF. In line with recent discussions of the importance of including various types and levels of these linguistic constructs (Norris & Ortega, 2009), numerous submeasures have been developed for CAF. Michel, Kuiken, and Vedder (2007), for example, included five measures of accuracy (total number of errors per AS-unit, number of lexical errors per AS-unit, number of omissions per AS-unit, percentage of self-repairs, and ratio of self-repairs to errors), four measures of complexity (total number of clauses per AS-unit, subordination index, percentage of lexical words, and Guirad's Index), and three fluency measures (the ratio of syllables per minute in unpruned speech, the ratio of syllables per minute in pruned speech, and number of filled pauses per 100 words). Their study is not atypical. Both within and across studies, many definitions and measures of CAF and of other features observed in task-based production can be found. One of the primary goals of the present study is therefore to arrive at a better understanding of the extent to which different features of learner production have been examined in the context of task performance.

Looking across one subset of task-based learner performance studies, Jackson and Suethanapornkul (2013) meta-analyzed investigations into Robinson's cognition hypothesis. Their findings indicate that resource-directing tasks are associated with (slightly) higher accuracy and (slightly) less fluency (see Sasayama, Malicka, & Norris, forthcoming, for a book-length synthesis and meta-analysis on the relationship between task complexity and learner output). However, little is known about precisely which kinds of learner production data have been examined in recent task-based studies.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

Task performance in both written and oral modes has also been compared in terms of interactional features (Payant & Kim, 2015) such as feedback moves. Interaction-based studies of task-induced production often examine the presence of interactional and nonlinguistic features believed to promote L2 development. For instance, the occurrence of learners' attention to linguistic forms such as corrective feedback, negotiation of meaning, language-related episodes (LREs, i.e., when learners talk about, question, and/or self-or-other correct language use; see Swain & Lapkin, 1998), and form-focused episodes (FFEs, i.e., learners' attention to linguistic forms during meaning-oriented tasks) have all been identified. Such studies exhibit methodological and theoretical parallels to the first phase of research in the interactionist tradition of SLA; see commentaries in the introduction to Mackey's (2007) volume and in Plonsky and Gass's (2011) methodological synthesis, both in the domain of interaction. Both task-based and interactionist researchers have examined in some depth the features of learner dialog (e.g., clarification requests, repairs) in different contexts (e.g., lab vs. classroom) and participant configurations (e.g., equal vs. unequal proficiency, number of interlocutors) (e.g., see Gass & Varonis, 1985).

Some scholars have gone beyond linguistic- and interaction-based targets to study the content of learner production as a function of task features or conditions. Among other dependent variables, Sample and Michel (2014) examined participants' ability to complete a task, which the authors argued to be "more important than any other feature of task performance" (p. 30). Such efforts address the necessity to confirm the quality or adequacy of task performance (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012; Ortega, 2005) as well as the assumptions researchers make about task features. A given task may be complex from researcher's or teacher's perspective, but not from that of the participants (see Révész & Gurzynski-Weiss, this volume). More fundamentally, examinations of task-induced learner language that fail to consider the semantic content therein run counter to many of the theoretical and practical principles underlying a task-based approach to L2 learning and teaching (Pallotti, 2009).

Situational features of tasks have been manipulated as well to examine their impact on linguistic and interactional features of learner production (e.g., Philp, Adams, & Iwashita, 2014). Conditions vary, for example, as to whether the task is completed in the oral or written mode (e.g., Révész, 2009) and whether the task is completed face-to-face with an interlocutor or via computer-mediated communication (CMC) (e.g., Gurzynski-Weiss & Baralt, 2014; see Ziegler, this volume, for a review). Although lab vs. classroom and foreign- vs. second language effects are rarely examined in primary studies, there is theoretical and empirical evidence to support the idea that these settings may moderate certain types of findings (Fujii & Mackey, 2009; Plonsky & Oswald, 2014; Ziegler, in press; Ziegler, this volume).

The present study brings together these various strands of research to understand the extent to which and means by which different designs, task conditions, and features of learner language have been examined.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

Task-Based Learning Outcome Studies

The second group of studies we introduce focuses on the role of tasks in L2 development. This line of research generally employs a pretest–posttest (task-as-treatment) design. Similar to the studies described thus far, developmental or interventionist studies often manipulate task features, situational/contextual, and/or participant variables. Though not mutually exclusive, the two types of designs differ in that this line of task-based research also examines how different task features and conditions impact subsequent language learning. This type of research runs parallel in many ways to the second phase in the interactionist tradition of SLA (Plonsky & Gass, 2011).

By implementing an interventionist approach, researchers are able to estimate the effects of different task types or conditions as well as task-based production on subsequent development in L2 lexis and grammatical forms (e.g., Kim, 2012). The treatment groups in Révész (2009), for example, completed a picture description task in one of four conditions: +/−task support (access to photos while describing them), +/−interactional feedback in the form of recasts. The results of the study presented both the number of recasts provided during the tasks (i.e., an interactional feature) and a comparison of the gains in performance on the targeted linguistic feature (past progressive) across the four task conditions. Similarly, Kim and Tracy-Ventura (2013) compared the effects of different task repetition conditions (task vs. procedural repetition) on learners' ability to produce language that is complex, accurate, and fluent. Among other results, the study found that both conditions led to a small increase in fluency on posttests, but that procedural repetition (and not task repetition) induced a boost in syntactic complexity.

Recently, the development of pragmatic features in a task-as-treatment approach has begun to receive attention as well. For instance, Kim and Taguchi (2015) examined the role of task complexity in the learning of request-making expressions by using collaborative writing tasks with Korean junior high school learners. Their findings indicated that both simple and complex tasks were equally beneficial for immediate learning outcomes, but only learners who performed more complex tasks that required reasoning processes maintained their learning four weeks after task performance.

Several meta-analyses have addressed the effects of tasks on L2 development. Keck, Iberri-Shea, Tracy-Ventura, and Wa-Mbaleka (2006) found strong evidence of learning resulting from task-based interaction, especially when tasks were designed in a way that required use of the target feature(s). Cobb (2010) replicated and expanded on Keck et al. by including more recent studies, but also restricted her domain to studies focused on gains in L2 grammar only. Both of these meta-analyses could be considered to have addressed subdomains within the scopes of Mackey and Goo's (2007) meta-analysis on the effects of L2 interaction and Plonsky and Gass's (2011) synthesis of interactionist research methods and outcomes.

Complementary to these meta-analyses, which have largely examined the effects of task-based instruction and interaction, the current study reviews the linguistic and interaction features analyzed in this body of research.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

Methodological Synthesis in L2 Research

As we state in the title and introduction to this article, we are not only interested in the target features and analyses of task-based learner production but in the methodological practices of this domain as well. Syntheses of research practices, part of the broader conceptual and methodological tradition of meta-research across the sciences (see Ioannidis, Fanelli, Dunne, & Goodman, 2015), seek to both describe and evaluate individual domains of inquiry. Certain methodological features are examined that are not necessarily associated with higher or lower quality but that help illuminate the domain and perhaps point to areas in need of further investigation (e.g., the change over time in lab- vs. classroom-based research found in Plonsky & Gass, 2011). Other features are indeed concerned with methodological rigor, validity, and transparency (i.e., study quality; see Plonsky, 2013) such as the use of pretests in quasi-experiments and full reporting of data.

Some methodological syntheses review a particular feature or technique across subdomains. Derrick (2015), for example, synthesized reporting practices related to the development and use of data collection instruments in L2 research. Plonsky and Gonulal (2015) reviewed the use of factor analysis in applied linguistics. Other methodological syntheses, including the present study, consist of methodological reviews of a particular substantive domain. For example, Liu and Brown (2015) synthesized research and reporting practices in studies of written corrective feedback. Even closer to the focus of the present study, Ziegler (2016) reviewed methodological features of interaction-based research comparing synchronous CMC (SCMC) and face-to-face contexts.

Results in this small but growing body of research are mixed but do not generally paint a very favorable picture of L2 research methods. A thorough review of findings would not be appropriate here. However, one trend observed across several domains is a lack of transparency. L2 researchers often fail to report critical data such as standard deviations (along with their means), test statistics, reliability estimates, and so forth. The field of L2 research has also been weakened by dogmatic adherence to and overuse of null hypothesis significance testing (see Norris, 2015; Plonsky, 2015b), a problem compounded by the small samples common to the field (Plonsky, 2014).

The Study

There are two motives behind the current study. The first is an interest in providing the field of applied linguistics with a better understanding of the features of learner production that are examined in research on task-based teaching and learning. Our second motive is an interest in describing and evaluating the research and reporting practices in this domain. One motive, corresponding to Research Question 1, is chiefly substantive in nature, and the other, which corresponds to Research Questions 2 and 3, is methodological. Our research questions are as follows:

1. To what extent and by which means have different features of learner production been analyzed in task-based L2 research?

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

2. To what extent have different designs, analyses, and reporting practices been employed in task-based L2 research?
3. How reliable (psychometrically speaking) are researchers' analyses of task-based production?

METHOD

The procedures used to carry out this study follow what is considered best practices in the realm of synthetic or meta-analytic research in applied linguistics (see Plonsky & Oswald, 2015). The focus of the current study, however, required a somewhat unique approach. Whereas meta-analyses of L2 research examine the effect sizes resulting from a treatment or intervention, the current study addressed substantive and methodological features (not findings) in studies of task-based learner production.

Inclusion Criteria and Search Techniques

The first step in this study involved defining the parameters by which primary studies would be selected. The resulting set of inclusion criteria were developed and applied, leading to a total sample of 85 unique reports.

The first and main criterion was an analysis of task-induced learner production. Within this broad parameter, all theoretical approaches, language/interactional features, modes, designs, and contexts were considered. Some studies identified in the search included treatments involving task-based activities and/or production but did not analyze what the learners produced during the tasks. These studies were excluded (e.g., de la Fuente, 2006). Studies that only examined teacher or learner perceptions of task-based teaching and learning were also excluded (e.g., Chen & Brown, 2012; Zheng & Borg, 2014). In the interest of synthesizing and assessing current research in this area, only studies published in the last 10 years, 2006–2015, were eligible.

Although an inclusive approach is generally preferred (Plonsky & Brown, 2015; Plonsky & Oswald, 2015), the current study included only published studies. This choice was informed by three considerations. First, the study was not necessarily interested in the entire population of research on task-based learner production but, rather, those that have likely had an influence on theory and practice in the domain. Second, published studies, which have gone through peer and/or editorial review, are likely to be of higher quality than unpublished ones. Third, excluding unpublished studies often leads to an overestimate of meta-analytic effects (i.e., publication bias; see Oswald & Plonsky, 2010), which were not considered in the present study. Although dissertations and conference presentations may be of high quality, access to these types of outlets is often restricted, thus reducing their influence on theory and practice.

A number of search techniques were applied to identify primary studies that met the criteria just described. The first and most productive searches were conducted via three databases: Linguistics and Language Behavior Abstracts (LLBA),

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

TABLE 1. *Studies of Task-Based Production From 2006 to 2015*

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	TOTAL
k	5	7	5	15	9	7	14	9	12	2	85
%	6	8	6	18	11	8	16	11	14	2	100

Note. k = number of studies.

TABLE 2. *Studies of Task-Based Production Across Journals*

Journal	*k* (%)	Journal	*k* (%)
System	8 (9%)	*Language Awareness*	4 (5%)
Studies in Second Lang. Acquisition	7 (8%)	*Language Learning & Technology*	4 (5%)
International Review of App. Ling.	6 (7%)	*Journal of Second Language Writing*	3 (4%)
Modern Language Journal	6 (7%)	*Innovation in LL & Teaching*	3 (4%)
Language Learning	5 (6%)	*Asian EFL Journal*	3 (4%)
Language Teaching Research	4 (5%)	*Applied Language Learning*	2 (2%)
Applied Linguistics	4 (5%)	*CALL*	2 (2%)
CALICO Journal	4 (5%)	*TESOL Quarterly*	2 (2%)

Note. The remaining journals each contributed one study to the sample: *Annual Review of Cognitive Linguistics; ELT Journal*; *EUROSLA Yearbook; Heritage Language Journal*; *International Journal of Applied Linguistics*; *International Journal of Research Studies in Language Learning*; *Journal of Applied Linguistics*; *Journal of Asia TEFL*; *Language Assessment Quarterly*; *Language Learning Journal*; *Latin American Journal of Content & Language Integrated Learning*; *Pazhuhesh-e Zabanha-ye Khareji*; *Southeast Asian Journal of English Language Studies*; *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*; *ReCALL; RESLA*; and *TESL Canada*.

Google Scholar, and Google (keywords: task-based, TBLT, second language, L2). In addition, the samples of two meta-analyses with scopes closely related to that of the current study were reviewed to identify additional studies (Cobb, 2010; Jackson & Suethanapornkul, 2013). Two special issues, one in *Applied Linguistics* (Housen & Kuiken, 2009b) and one in *International Review of Applied Linguistics* (Robinson & Gilabert, 2007) were also examined for studies that might meet the inclusion criteria.

The final sample consisted of 85 unique reports, published from 2006 to 2015 (see Table 1). Full references for the sample will be made available by the authors upon request. Further searches could have been applied using other techniques (e.g., forward citations, email list requests) and extensions on those already employed (e.g., additional database searches). The sample should not be considered exhaustive. Nevertheless, we have no reason to believe that the sample obtained for the present study is not representative of the domain of interest.

In addition to chronological dispersion, a range of different journals are represented in the sample. As shown in Table 2, a small number of reports were published in a variety of prominent applied linguistics journals with both general and specialized orientations. And although all studies were published in English, the journals represented are quite international in scope. This domain is clearly of interest to scholars from a broad range of countries and contexts around the world.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

TABLE 3. *Coding Scheme Categories and Items*

Category	Items
Study Identification	author(s)*; year*; journal
Design and Setting	class; laboratory; random assignment; first language*; target language*; second language setting; foreign language setting; heritage learners; proficiency (as stated by authors): beginner, intermediate, advanced; educational context: elementary, high school, university, language institute; sample size*; design (task-based learner performance vs. task-based learning outcome / task-as-treatment)
Production Targets and Features	mode: oral, written, face-to-face, CMC; skills: grammar, vocabulary, pronunciation, pragmatics; interactional features; production features: complexity, accuracy, lexis, fluency; first language use; quality/appropriateness of learner production
Coding Process	corpus/computer-based; manual
Statistical Analyses	frequency/percentage, correlation, chi-square, *t* test, ANOVA, post hoc test, ANCOVA, MANOVA, MANCOVA, factor analysis, regression, structural equation modeling, hierarchical linear modeling, cluster analysis, discriminant function analysis, Rasch analysis, nonparametric test, other*, number of statistical tests*
Transparency	reliability reported; reliability index: percent agreement, alpha, kappa, correlation; observed reliability*; assumptions checked; missing data: standard deviation, test statistic, other; effect size; confidence interval; visual display of data; research questions; task available

Note. All items coded dichotomously (i.e., 0 = absent, 1 = present) unless marked with * to indicate items with an open-ended response.

Data Collection

A coding scheme was developed to extract substantive and methodological information from each study in the sample. Six categories of data were collected: (a) study identification (e.g., author, journal), (b) study design and setting (e.g., class vs. lab, target language), (c) production targets and features (e.g., mode, grammar, complexity), (d) coding process (e.g., corpus-based, by hand), (e) statistical analyses, and (f) reporting practices associated with transparency (e.g., reliability, effect sizes, confidence intervals). The complete set of items appears in Table 3, and our data collection instrument will be made available on the IRIS (data collection instruments for research into second language learning and teaching) database (Marsden, Mackey, & Plonsky, 2015).

After an initial version of the coding scheme was developed, Plonsky coded three studies. In doing so, the operational definitions of a small number of items were refined. He then trained a second rater, his research assistant (RA), who has taken graduate courses in both task-based teaching and learning and in meta-analysis. After discussing the goals of the study and each item in the coding scheme, a step that led to the refinement of several operational definitions, the RA

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

and Plonsky independently coded the remainder of the sample. Plonsky coded a total of 44 studies, and the RA coded 41.

Once all 85 studies were coded, Plonsky, recoded 10% of the sample (nine studies) that had been coded previously by the RA to estimate the interrater reliability of our data collection process. The overall agreement rate was 95%. Because our instrument was composed almost entirely of categorical items, which are susceptible to chance agreement and inflated reliability estimates (see Plonsky & Derrick, in press), we also calculated Cohen's kappa ($\kappa = .91$). Both indices demonstrated an overall high level of interrater agreement. However, upon closer inspection, we were concerned about four individual items. The two raters disagreed four out of nine times on the analyses of learners' grammar and vocabulary, respectively, three times on the total number of statistical tests conducted, and three times on study design. Plonsky then recoded these four variables for the 41 studies that had been coded by the RA. The revised values were used in all subsequent analyses.

Analysis

The analyses involved in the current study were very straightforward. Research Questions 1 and 2 were addressed by calculating frequencies and percentages of the (categorical) substantive and methodological features found in the sample. Measures of central tendency were calculated for the few continuously measured variables such as sample size. Research Question 3 was addressed by calculating the mean, standard deviation, and 95% confidence intervals of the reliability coefficients associated with coding of task-based production as reported in individual studies. A small number of studies reported reliability coefficients for multiple target features (e.g., number of LREs + accuracy). In such cases, the estimates were combined (averaged) in order to avoid data dependencies at the meta-analytic level (Lipsey & Wilson, 2001). Reliability estimates that are reported may be higher than those that go unreported or that are never calculated in the first place. Consequently, the meta-analytic estimate provided in our results may be somewhat inflated.

RESULTS

Demographics, Contexts, and Research Designs

Frequencies and percentages of different demographics, research designs, and contexts were calculated for the sample of studies examining task-based learner production. Table 4 presents the results for two demographic features: learner proficiency and age. The majority of the studies in this area have collected data from what authors have referred to as intermediate learners. In nearly all cases, this label should be considered impressionistic; very few authors provide any justification (e.g., TOEFL score) for their assessment of proficiency. The data also reveal a very strong tendency toward research on adult learners.

Several design features in this domain were examined. The first, sample size, is closely related to the demographics just described. Samples ranged from 1 to 968.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

TABLE 4. *Learner Demographics*

Learner types	*k*	%*
Proficiency		
Beginner	27	32
Intermediate	53	62
Advanced	14	16
Multiple	20	24
Age		
Children (<13)	10	12
Teen (13-17)	9	11
Adult (>17)	64	75
Multiple	1	1
Not reported	1	1
Heritage	3	4

*Totals do not equal 100 because some studies included participants at multiple levels.

TABLE 5. *Research Contexts, Settings, and Institution Types*

Descriptor	*k*	%
Context		
Class	51	60
Lab	34	40
Setting		
Second language	19	22
Foreign language	66	78
Institution		
Elementary	10	12
High school	7	8
University	57	67
Language institute	11	13

Most, however, were closer to the median of 36 (interquartile range = 28). Studies were also coded for two major design types: (a) task-based learner performance ($k = 57$, 67%) vs. (b) task-as-treatment ($k = 28$, 33%).

The study included three variables related to research context or setting, the results of which are found in Table 5. Again, several clear patterns emerge. Studies of task-based learner production are more classroom- than lab-based, and the vast majority of research in the domain is conducted in foreign-language rather than L2 settings. Most research in this area is also conducted with university learners, a finding that reflects the result above related to participant age.

This review also examined the different first languages (L1s) and target languages (TLs) being studied. As shown in Figures 1 and 2, this domain has relied on very few L1s and TLs, with English being the clear frontrunner in both categories. In fact, 86% of the studies in the sample are based on data from users of English as an L1 or TL.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

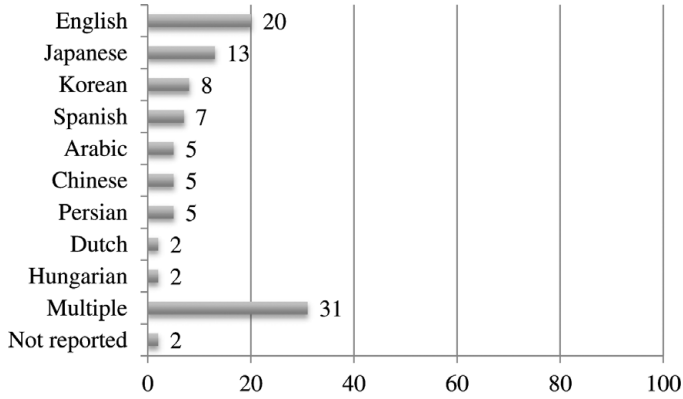


FIGURE 1. Participants' First Languages (%)

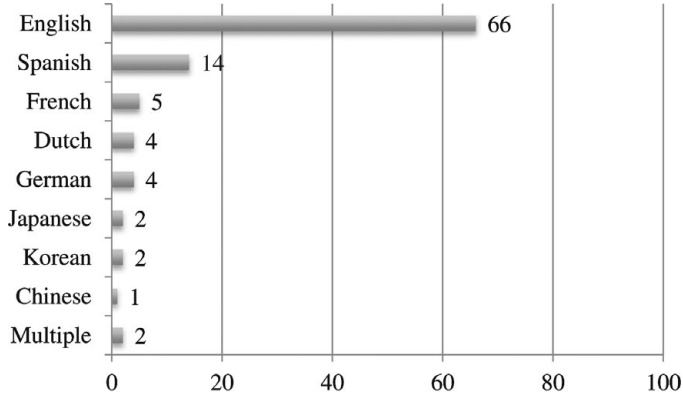


FIGURE 2. Participants' Target Languages (%)

Analyses of Learner Production Data

In order to first understand the data collection procedures of research on task-based production, studies were coded for task modality (oral, written), as well as whether tasks were carried out via face-to-face or computer-mediated communication. The results for this aspect of the analyses appear in Table 6.

One of the main goals of this study was to better understand the features that have been studied in task-based production. Each study in the sample was coded for the different targets of analysis. These included linguistic domains such as grammar and pronunciation, the well-known CAF framework, use of participant L1, interaction features (e.g., LREs), and the quality of their task completion (i.e., was the task completed successfully? see Sample & Michel, 2014).

The results for this phase of the analysis appear in Figures 3 and 4 and are presented as the percentage of studies in the sample that analyzed different features of learner production and interaction. We see first of all a variety of target features.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

TABLE 6. *Modality of Task-Based Learner Production*

Mode*	*k*	%
Oral	69	81
Written	24	28
Face-to-face	67	79
Computer-mediated	18	21

*Categories are not mutually exclusive.

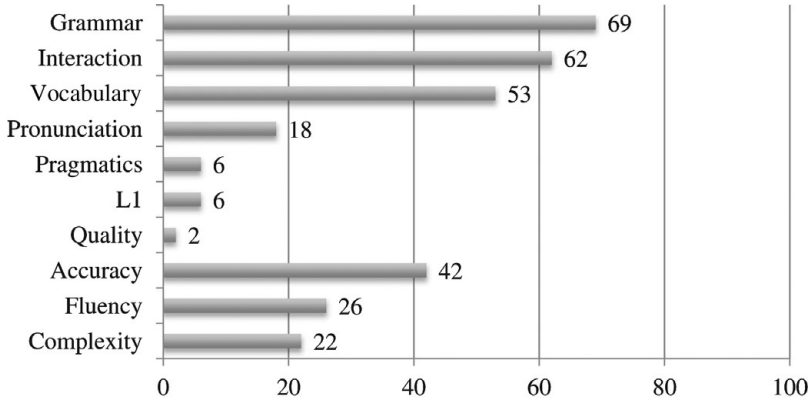


FIGURE 3. Target Features in Task-Based Production (%)

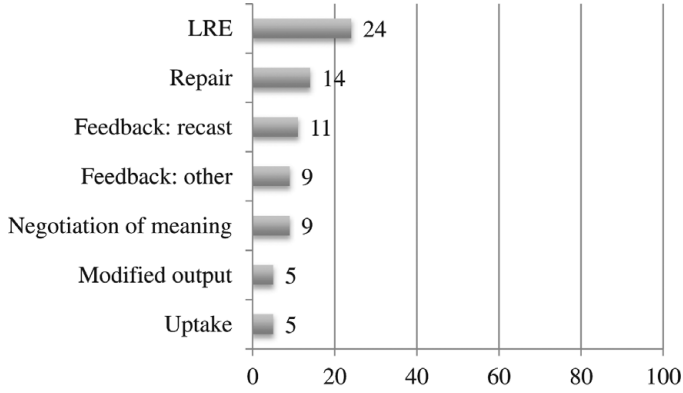


FIGURE 4. Interactional Features Examined (%)

It is also clear from the data that interest in learner production resulting from task-based research has focused most heavily on grammar, interactional features, and vocabulary; among CAF, the most often studied feature is accuracy. Considerably less attention has also been paid to pronunciation, fluency, and complexity, and

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

TABLE 7. *Techniques for Coding Learner Production*

Coding	*k*	%
Manual	78	92
Corpus/computer-based	4	5
Both	3	4

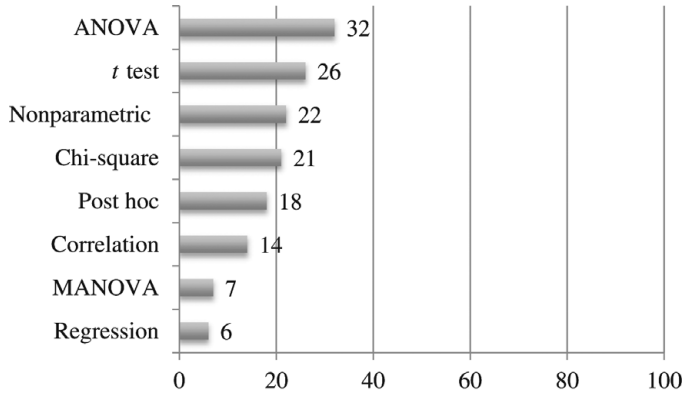


FIGURE 5. Statistical Analyses in Studies of Task-Based Production (%)

very few studies have examined pragmatics and use of the L1. The quality of the task content produced by the participants (e.g., task completion) has been rarely studied.

In order to better understand this domain, we also coded for subtypes of interactional features, a broad and widely studied aspect of this domain. Figure 4 presents the results for this phase of the analysis. Approximately a quarter of the entire sample studied LREs. Learner repair, feedback types, and the presence of negotiation for meaning have also been studied with some regularity. A variety of other interactional features were observed but in less than 5% of the sample. These include humor (Collentine, 2009), communication strategies (Khamis, 2010), and turn taking (Butler & Zeng, 2014).

In addition to the features of L2 production, we were also interested in the tools and techniques employed to analyze them. This phase of the study involved coding primary studies, first, for whether learner production was analyzed by hand or using automatic, computer-based tools such as those developed by corpus linguists. We also recorded the statistical analyses applied by researchers in this area. These results appear in Table 7 and Figure 5, respectively. Nearly all coding of target features in learner production was carried out manually. Only six studies applied automatic, corpus-based computer techniques whether alone or in coordination with hand coding.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

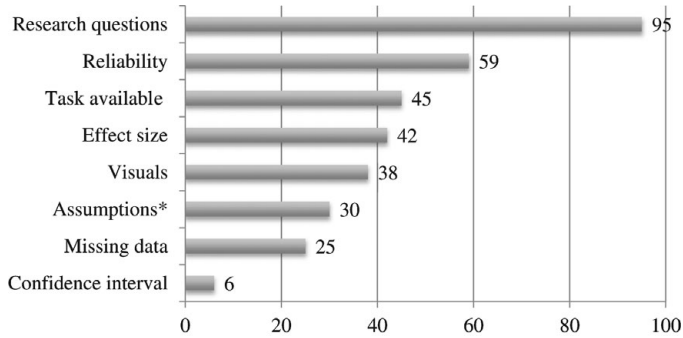


FIGURE 6. Features Associated with Researcher Transparency (%)
*Calculated by dividing the number of studies that reported to have checked the assumption of their statistical tests (21) by the number of studies that conducted one or more statistical tests (70)

There are relatively few unique statistics observed in this domain. As in other areas of applied linguistics, most studies rely on tests of mean differences between or among groups. Not only are the highest percentages found for ANOVA and *t* test, but also all post hoc tests found in the sample were associated with ANOVAs, and all nonparametric tests were conceptually (if not statistically) addressing differences between or within groups. The following analyses were found in a single study within the sample: ANCOVA, factor analysis, and Rasch analysis.

In addition to understanding the variety of unique analyses being carried out in our sample, we were also interested in how they are being used in combination as well as the total number of statistical tests applied per study. The latter issue is a concern due to the small samples typical in the field and their debilitating effect on statistical power. Whereas a small number of studies ($k = 15$, 18%) presented only descriptive statistics and ran no statistical tests, twice as many ($k = 36$, 42%) conducted a single type of statistical analysis (e.g., one or more *t* tests), and just about as many ($k = 34$, 40%) conducted multiple unique analyses (e.g., one or more *t* tests and correlations). Of the studies that conducted one or more statistical tests, the number of tests of statistical significance per study ranged from 2 to 109 (median = 12.5; interquartile range = 27.25).

Transparency

Eight features associated with researcher transparency were coded for, the results of which are presented in Figure 6. Research questions are consistently found in primary reports of research in this domain. Other practices are found only sporadically. Interrater reliability, for example, which is necessary given the judgment required to code for features of learner production (e.g., errors, T-units, LREs), was found in just over half of the sample. A somewhat smaller portion of the studies in our sample provided the tasks used in their studies. One quarter of the

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

TABLE 8. *Descriptive Statistics for Meta-Analysis of Reliability Coefficients*

M	*SD*	95% CI	Min–Max
.94	.05	[.93, .95]	.80–1.00

sample also failed to report a statistic—usually one or more standard deviations—required of the type of data being presented.

Reliability

One finding in the previous section pertains to the presence of reliability estimates in studies of task-based production. The results shown in Table 8 are concerned with the actual (aggregated) reliability as observed in the sample. Reliability estimates in this domain are generally quite high and consistent.

DISCUSSION

This study applied a synthetic approach to examine research on task-based L2 learner production. Our interests were both substantive and methodological. We were concerned not only with what constructs and features had been studied but also with the methods that had been used to study them. It is somewhat artificial to consider these issues in isolation as "study results are determined conjointly by the nature of the substantive phenomenon under investigation and the nature of the methods used to study it" (Lipsey, 2009, p. 150). In our discussion, we therefore consider the relationships between these two types of findings when and as appropriate. We also use our results to make empirically grounded suggestions for building on and improving primary research in this domain of applied linguistics.

The current study revealed that interactional features are second only to grammar in the frequency with which they are examined. This pattern may be attributed to two reasons. First, many interactionist SLA researchers who use tasks to elicit participant interaction have shown that task features can also affect the presence and amount of different types of interactional features (Kim, 2009; Révész, 2011). Second, as the educational value of learner-learner interaction has been studied extensively in recent years, researchers have sought to understand how learning opportunities during interaction might be facilitated (Philp et al., 2014).

We also analyzed the range of interactional features examined in the sample, the most frequent of which was LREs. Studies in this domain have investigated several other moves as well including feedback types (e.g., recasts), modified output, and uptake. These results as well as those discussed in the previous paragraph provide evidence of the strong theoretical and methodological connection between research on L2 interaction and task-based teaching and learning.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

In terms of linguistic features that were analyzed in recent task-based studies, our results show a heavy focus on learners' grammar and vocabulary production (69% and 53% of the sample, respectively). Moreover, it was common to observe studies interested in LREs, the most frequently analyzed interactional features, subdividing them further as grammatical or lexical (Kim, 2009; Swain & Lapkin, 1998). This finding is not particularly surprising given the frequent interest in encouraging learners' attention to linguistic forms during meaning-oriented tasks. This result also mirrors classroom-based research indicating the frequency of teacher feedback in response to learners' grammar and vocabulary errors (Brown, in press).

Although there are certainly many compelling reasons for the focus on learner-produced grammar and vocabulary, many of the theoretical and practical rationales for studying these targets would also apply to pronunciation and pragmatics. Outside of the realm of task-based teaching and learning, empirical efforts on L2 pronunciation and pragmatics are thriving (e.g., Culpeper, Taguchi, & Mackey, in preparation; Lee, Jang, & Plonsky, 2015). We would encourage task-oriented researchers to collaborate with colleagues who work in L2 pronunciation and/or pragmatics, and vice versa, to add to this body of research and to build on the momentum in these areas (e.g., Gurzynski-Weiss, Long, & Solon, in preparation; Taguchi & Kim, under contract).

The CAF framework has received substantial theoretical and empirical attention in applied linguistics (Housen & Kuiken, 2009a; Housen et al., 2012; Norris & Ortega, 2009; Pallotti, 2009). As attested by the results of the current study, task-based researchers have certainly benefited from and contributed to such discussions. A substantial portion of the studies in our sample examined one or more of the CAF features. A number of these studies stem from early task-based research that predominantly adopted monologic tasks (e.g., picture description). For instance, early studies of task complexity testing the cognition hypothesis or trade-off hypothesis (see Jackson & Suethanapornkul, 2013) and task planning time studies (Ellis, 2005, 2009) often focused their analyses on CAF. These and other efforts have led to significant developments in terms of both methods and theoretical underpinnings of various submeasures of each area (e.g., clausal complexity, phrasal complexity). However, in coding our sample of studies we observed a great deal of within- and between-study variety in terms of how these constructs are defined and measured.

The heterogeneity that results from the number and variety of unique operationalizations of CAF yields problems at two levels of analysis. First, a lack of consistency across studies impedes efforts such as the present study and Sasayama et al. (forthcoming) to synthesize findings and draw stable conclusions (see related comments in Foster & Wigglesworth, this volume). Theoretical plurality is certainly valuable, but there is a need for greater definitional and operational clarity of CAF (Norris & Ortega, 2009; Pallotti, 2009). Based on a synthesis of accuracy measures and their associated weaknesses, Foster and Wigglesworth (this volume) present and illustrate a well-reasoned, alternative index. Their arguments are both substantively and methodological informed. We cannot predict whether their index will be adopted, but we applaud their effort toward consolidation. Similarly,

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

Bulté and Housen's (2012) review found myriad definitions, operationalizations, and measures associated with linguistic complexity. The different types of CAF measures used in task-based studies need to be theoretically motivated and used in a way that allows for comparability across studies.

The second problem resulting from a lack of operational consistency is more statistical in nature. It is not uncommon to find studies of task-based teaching and learning with multiple independent and dependent variables. Furthermore, many designs include multiple measures of a single dependent variable such as linguistic complexity. Task-based researchers have occasionally applied multivariate techniques such as MANOVA in such situations (e.g., Ishikawa, 2008). As shown in our results, however, more often than not, numerous univariate and bivariate statistical tests are conducted. This practice, which is particularly problematic in light of the small samples typical in this body of research and throughout applied linguistics, is irresponsible and poses a serious threat to internal validity at the primary and secondary levels. A combination of (a) theoretical and operational consolidation and (b) a reduction in statistical testing would likely lead to marked improvements in our understanding of task-based learner language. In some situations, given the inherently multivariate nature of language learning and use, multivariate statistics can and should be utilized (Brown, 2015; Plonsky & Oswald, under review).

Designs, Analyses, and Reporting Practices

Many of our results for research and reporting practices in this domain reflect those observed elsewhere in recent years in quantitative applied linguistics (e.g., Plonsky, 2013). Some of these findings are associated with higher or lower quality. Others, however, simply provide a better understanding of this body of research.

With respect to designs, our results indicate that the majority of the sample was interested in describing linguistic and interactional features as a function of task and/or situational variables; only 33% of our sample tested the effects of a task-based treatment on subsequent learning. This might be due to the nature of traditional task-based studies, which compared learners' task performance (e.g., CAF) between tasks with different design and implementation conditions. Over time the link between task-based research and SLA has grown, and more recent studies have adopted a pretest–posttest design to shed light on task-based language learning.

There are several findings related to sampling, another aspect of study design, worth highlighting as well. Samples in task-based research are limited both in size and diversity. Almost everything we know about tasks and learner language is based on highly educated young adult users of English. The main cause for such a restricted range of sample demographics is clear and well known: convenience. But the consequence of systematic oversight is potentially severe: massive limitations to both theoretical advancement and to the potential of our research to inform practice.

The analyses in this domain were also found to be less than optimal, though not in any immediate sense. The main issue is one that has surfaced several times in

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

recent years in broader methodological and statistical discussions: an overreliance on null hypothesis significance testing (Norris, 2015; Plonsky, 2015b). Statistical tests in research on task-based learner production are simply overused and underpowered (i.e., lacking in the ability to detect true statistical relationships). Although we did not include this feature in our data collection instrument, with two notable exceptions (Alwi, Adams, & Newton, 2012; Baralt, 2013), we observed little to no evidence of concern for power. A related issue, noted by Pallotti (2009), is the practice of equating (statistically significant) mean differences with importance and, conversely, a lack of such differences with unimportance. Assuming principled design and data collection, the outcome of a theoretically motivated comparison of mean scores should be of interest regardless of whether or not it yields a statistically significant or large difference. In fact, when a difference is predicted by theory, some might argue a lack of evidence for that difference to be even more worthy of attention (see Godfroid & Spino, 2015; Plonsky, 2015a).

Looking ahead, we recommend a move toward larger samples and fewer contrasts. This approach will lead to a smaller but more stable set of results, which can then be aggregated at the synthetic/meta-analytic level. More generally, we would also suggest that researchers divest their attention away from statistical significance and toward a closer consideration and interpretation of descriptive statistics. Standard deviations are a prime example. At the primary study level, these values provide critical information regarding the extent to which participants perform uniformly (or not) on a given task or in response to a given treatment (see Pallotti, 2009). At the meta-analytic level, they are used to calculate effect sizes. Nevertheless, standard deviations are almost never interpreted and they are often left out of primary reports.

Our results also show that automatized (corpus-based) analyses of learner language are almost never employed. Corpus techniques are not appropriate in all situations, and they can certainly present challenges. For example, corpus tools can be less accurate when used to analyze productive data that contain errors or other irregularities. Another challenge is that not all applied linguists have programming skills. We have found in our experience, however, that corpus researchers are often willing to engage with research problems outside their typical areas of study. We have also found that programs needed to assist with text analysis can often be developed in a reasonable amount of time, particularly in the case of lexical targets. In addition to greater efficiency and the obvious logistical benefits, there are at least two main advantages of corpus techniques over traditional, hand-coded analyses. First, corpus analyses can vastly improve replicability. The culture of openness and transparency currently developing in applied linguistics through IRIS and other initiatives has long since been present among corpus linguists, and both corpora and corpus tools are often shared freely (see, e.g., Kristopher Kyle and Scott Crossley's tools for assessing lexical sophistication and syntactic complexity: http://www.kristopherkyle.com/). This approach would not only allow for others to reexamine an original set of linguistic analyses, but also to replicate the analyses and generalize the results to other learners, tasks, and so forth (Kyle & Crossley, 2015). Another potential advantage of corpus analyses for task-based

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

learner production is the reliability inherent in the technique. Despite the generally high overall (meta-analytic) reliability reported above ($M = .94$, $SD = .05$), human rating is always susceptible to error. By contrast, computer programs always follow the rules written into them.

Related to the notion of transparency, the present study also coded for a number of reporting practices. Research questions were found in almost every study in the sample. Most studies (59%) reported an estimate of (interrater) reliability, which is higher than in most domains of L2 research (see Plonsky & Derrick, in press), but still less than ideal. Also concerning was the lack of interpretation surrounding reliability coefficients. As with effect sizes (42%), authors appear to treat the reporting of reliability estimates as a perfunctory act. They are almost never considered beyond an over simplified dichotomy of acceptable/unacceptable. Visual displays of data are found regularly in this domain (38% of the sample). However, the types of graphic techniques employed were rarely effective (see Hudson, 2015; Larson-Hall & Plonsky, 2015). For example, most were simple bar graphs with no indication of dispersion. Finally, we note that 45% of the sample made their tasks available to readers and other researchers. This figure could and should be higher, and we would like to see more researchers upload their materials and instruments to the IRIS database, but we view this result as generally positive and indicative of a degree of transparency on the part of the authors in our sample.

Recommendations for Future Research

Our focus throughout much of this synthesis has been retrospective. However, we are not only interested in describing and evaluating previous research in this domain. We are quite concerned in fact about future analyses of task-based production. Therefore, based on the findings of the current study, we present in this section a set of recommendations for areas in need of attention in future efforts. Many of these comments echo our colleagues' and collaborators' suggestions made as part of broader methodological discussions taking place in the field (see Norris, Plonsky, Ross, & Schoonen, 2015). For the sake of clarity and economy, our recommendations appear in bulleted form.

1. Substantive areas in need of further attention
    - Production targets and features: pronunciation, pragmatics, L1 use, and quality of task performance
    - Nonlinguistic and interactional features: modified output, uptake, and feedback types
    - The role of learner background (e.g., individual differences, L1) as a moderator of production
    - Face validity of task types
2. Designs
    - Larger samples
    - Greater variety of learner demographics (L1s, TL, ages, and proficiency levels)

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

- Aptitude-treatment interaction (ATI) studies, to investigate the role of individual differences as a moderator of the learning that results from different task conditions

3. Analyses
   - Greater use of corpus techniques for analyzing learner production
   - Fewer tests of statistical significance
   - Greater use and interpretation of descriptive statistics, including effect sizes and confidence intervals
   - Use of multivariate statistics when appropriate
4. Data reporting and transparency
   - More thorough reporting of descriptive statistics including standard deviations, effect sizes, and confidence intervals
   - Greater reporting and interpretations of reliability estimates
   - Greater task and instrument availability (e.g., via IRIS)
   - Improved techniques for visual displays of data

CONCLUSION

This study, a review of research on task-based learner production, has found empirical efforts in this area to be both abundant and diverse in their interests. Pursuing a wide range of linguistic and interactional features of learner language is certainly beneficial for advancing our theoretical understanding of tasks and of L2 development. However, the combination of a wide range of constructs and of definitions and measures of those constructs has led to a necessarily uneven body of results. In response to this issue and based on the results of our study, we suggest a number of substantive features that merit additional attention. We also join others in recommending an effort toward enhanced theoretical integration and consistency. Complementary to the findings concerning the linguistic and interactional targets, we have also examined a range of methodological practices in this domain. We are concerned about several of our findings including sampling practices, data analyses, and reporting practices. However, we are also encouraged by the productivity and persistence of scholars in this area who will no doubt continue their efforts to advance our understanding of the role of tasks in language learning and teaching.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0267190516000015.

REFERENCES

Alwi, N. A. N. M., Adams, R., & Newton, J. (2012). Writing to learn via text chat: Task implementation and focus on form. *Journal of Second Language Writing*, *21*, 23–39.

Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in Second Language Acquisition*, *35*, 689–725.

Brown, D. (in press). The type and linguistic foci of oral corrective feedback in the L2 classroom: A meta-analysis. *Language Teaching Research*.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

Brown, J. D. (2015). Why bother learning advanced quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 9–20). New York, NY: Routledge.

Bulté, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). Philadelphia, PA: John Benjamins.

Butler, Y. G., & Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessments. *Language Assessment Quarterly, 11*, 45–75.

Bygate, M., Skehan, P., & Swain, M. (2001). Introduction. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 1–20). Essex, UK: Pearson Education.

Byrnes, H. (2013). Notes from the editor. *Modern Language Journal, 97*, 825–827.

Chen, J. C. C., & Brown, K. L. (2012). The effects of authentic audience on English as a second language (ESL) writers: A task-based, computer-mediated approach. *Computer Assisted Language Learning, 25*, 435–454.

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, California.

Collentine, K. (2009). Learner use of holistic language units in multimodal, task-based synchronous computer-mediated communication. *Language Learning & Technology, 13*, 68–87

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics, 5*, 161–170.

Culpeper, J., Taguchi, N., & Mackey, A. (in preparation). *Methods in second language pragmatics*. New York, NY: NY: Routledge.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2012). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and nonnative speakers. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Investigating complexity, accuracy and fluency in SLA* (pp. 121–142). Philadelphia, PA: John Benjamins.

de la Fuente, M. J. (2006). Classroom L2 vocabulary acquisition: Investigating the role of pedagogical tasks and form-focused instruction. *Language Teaching Research, 10*, 263–295.

Derrick, D. J. (2015). Instrument reporting practices in second language research. *TESOL Quarterly, 50*, 132–153.

Ellis, R. (Ed). (2005). *Planning and task performance in a second language*. Amsterdam, The Netherlands: John Benjamins.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics, 30*, 474–509.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford, UK: Oxford University Press.

Fujii, A., & Mackey, A. (2009). Interactional feedback in learner-learner interactions in a task-based EFL classroom. *International Review of Applied Linguistics, 47*, 267–301.

Gass, S., & Varonis, E. M. (1985). Task variation and nonnative/nonnative negotiation of meaning. In S. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 149–161). Rowley, MA: Newbury House.

Godfroid, A., & Spino, L. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence. *Language Learning, 65*, 896–928.

Gurzynski-Weiss, L., & Baralt, M. (2014). Exploring learner perception and use of task-based interactional feedback in FTF and CMC modes. *Studies in Second Language Acquisition*, *36*, 1–37.

Gurzynski-Weiss, L., Long, A. Y., & Solon, M. (Eds.). (in preparation). TBLT and L2 pronunciation: Do the benefits of tasks extend beyond grammar and lexis? [Special issue]. *Studies in Second Language Acquisition*.

Housen, A., & Kuiken, F. (2009a). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*, 461–473.

Housen, A., & Kuiken, F. (Eds.). (2009b). Complexity, accuracy, and fluency (CAF) in second language acquisition research [Special issue]. *Applied Linguistics, 30*.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. *32*). Philadelphia, PA: John Benjamins.

Hudson, T. (2015). Presenting quantitative data visually. In L. Plonsky (Ed.), *Advancing quantitative methods in second research* (pp. 78–105). New York, NY: Routledge.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biology, 13*, e1002264.
Ishikawa, T. (2008). The effect of task demands of intentional reasoning on L2 speech performance. *The Journal of Asia TEFL, 5*, 29–63.
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., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Philadelphia, PA: John Benjamins.
Khamis, H. (2010). Communication strategies in computer-mediated communication: An Egyptian EFL context. *CALICO Journal, 28*, 35–48.
Kim, Y. (2009). The effects of task complexity on learner–learner interaction. *System*, *37*, 254–268.
Kim, Y. (2012). Task complexity, learning opportunities, and Korean EFL learners' question development. *Studies in Second Language Acquisition*, *34*, 627.
Kim, Y., & Taguchi, N. (2015). Promoting task-based pragmatics instruction in EFL classroom contexts: The role of task complexity. *Modern Language Journal, 99*, 656–677.
Kim, Y., & Tracy-Ventura, N. (2013). The role of task repetition in L2 performance development: What needs to be repeated during task-based interaction? *System*, *41*, 829–840.
Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*, 757–786.
Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*, 345–366.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing, 30*, 66–81.
Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*(Supp. 1), 127–159.
Lipsey, M. W. (2009). Identifying interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 147–158). New York, NY: Russell Sage Foundation.
Long, M. (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, UK: Multilingual Matters.
Mackey, A. (2007). Introduction. The role of conversational interaction in second language acquisition. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 1–26). Oxford, UK: Oxford University Press.
Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–449). Oxford, UK: Oxford University Press.
Marsden, E., Mackey, A., & Plonsky, L. (2016). Breadth and depth: The IRIS repository. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). New York, NY: Routledge.
Michel, M. C. (2013). The use of conjunctions in cognitively simple versus complex oral L2 tasks. *Modern Language Journal*, *97*, 178–195.
Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, *45*, 241–259.
Michel, M. C., Kuiken, F., & Vedder, I. (2012). Task complexity and interaction: (Combined) effects on task-based performance in Dutch as a second language. *EUROSLA Yearbook*, *12*, 164–190.
Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning, 65*(Supp. 1), 97–126.
Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*, 555–578.
Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning, 65*, 470–476.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

Ortega, L. (2005). What do learners plan? Learner-driven attention to form during pre-task planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 77–109). Amsterdam, The Netherlands: John Benjamins.
Ortega, L. (2009). *Understanding second language acquisition*. New York, NY: Hodder Education.
Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics, 30*, 85–110.
Pallotti, G. (2009). CAF: Defining, refining and differentiation constructs. *Applied Linguistics, 30*, 590–601.
Payant, C., & Kim, Y. (2015). Language mediation in an L3 classroom: The role of task modalities and task types. *Foreign Language Annals, 48*, 706–729.
Philp, J., Adams, R., & Iwashita, N. (2014). *Peer interaction and second language learning*. New York, NY: Routledge.
Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *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. (2015a). Quantitative considerations for improving replicability in CALL and applied linguistics. *CALICO Journal, 32*, 232–244.
Plonsky, L. (2015b). Statistical power, *p* values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23–45). New York, NY: Routledge.
Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*, 267–278.
Plonsky, L., & Derrick, D. J. (in press). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*.
Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366.
Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning, 65*(Suppl. 1), 9–36.
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, NY: Routledge.
Plonsky, L., & Oswald, F. L. (under review). *Multiple regression as a flexible alternative to ANOVA in L2 research*.
Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition*, *31*, 437–470.
Révész, A. (2011). Task complexity, focus on L2 constructions, and individual differences: A classroom based study. *Modern Language Journal*, *95*, 162–181.
Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*, 27–57.
Robinson, P., & Gilabert, R. (Eds.) (2007). Task complexity, the cognition hypothesis and second language instruction [Special issue]. *International Review of Applied Linguistics in Language Teaching (IRAL), 45*.
Sample, E., & Michel, M. (2014). An exploratory study into trade-off effects of complexity, accuracy, and fluency on young learners' oral task repetition. *TESL Canada Journal, 31*, 23–46.
Sasayama, S., Malicka, A., & Norris, J. M. (forthcoming). *Cognitive task complexity: A research synthesis and meta-analysis*. Philadelphia, PA: John Benjamins.
Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge: New York
Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics, 10*, 209–31.
Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532.
Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *Modern Language Journal, 82*, 320–337.
Taguchi, N., & Kim, Y. (Eds.). (under contract). *Task-based approaches to pragmatics teaching and assessment*. Philadelphia, PA: John Benjamins.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press

Thomas, M. (2013). History of the study of second language acquisition. In J. Herschensohn & M. Young-Scholten (Eds.), *The Cambridge handbook of second language acquisition* (pp. 26–45). Cambridge, UK: Cambridge University Press.

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

Ziegler, N. (2016). Methodological practices in interaction in synchronous computer mediated communication: A synthetic approach. In A. Mackey & E. Marsden (Eds.), *Instruments for research into second languages: Empirical studies advancing methodology* (pp. 197–223). New York, NY: Routledge.

Ziegler, N. (in press). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*.

https://doi.org/10.1017/S0267190516000015 Published online by Cambridge University Press