SYSTEMATIC REVIEW ARTICLE

The Effectiveness of Second Language Strategy Instruction: A Meta-analysis

Luke Plonsky
Northern Arizona University

Research on the effects of second language strategy instruction (SI) has been extensive yet inconclusive. This meta-analysis, therefore, aims to provide a reliable, quantitative measure of the effect of SI as well as a description of the relationship between SI and the variables that moderate its effectiveness (i.e., different learning contexts, treatments, and outcome variables). A comprehensive search was conducted to collect the population of SI studies. Effect sizes were calculated for 61 primary studies, contributing a total of 95 unique samples, all of which were coded for potential moderators. The findings indicate a small to medium overall effect of SI ($d = 0.49$). Variables found to moderate its effectiveness include type and number of strategies, learning context (second vs. foreign language), and length of intervention. Following a contextualized interpretation of the results, the article concludes with a discussion of theoretical, practical, and methodological implications.

Keywords strategy instruction; language learning strategies; second language acquisition; meta-analysis

Introduction

Second language (L2) learning strategies have been the object of a large body of research over the last 30 years. Dozens of reports have examined
the types of strategies learners use as well as how they are used in conjunction with other variables such as gender, proficiency, style, motivation, autonomy, and epistemological beliefs (Green & Oxford, 1995; Ikeda & Takeuchi, 2006; Jiménez Catalán, 2003; Mogagwe & Oliver, 2007; O’Malley, Chamot, Stewner-Manzano, Russo, & Küpper, 1985; Oxford & Nyikos, 1989; Yamamorio, Isoda, Hiromori, & Oxford, 2003). Flanking this line of predominantly descriptive and correlational inquiry, experimental research on L2 strategy instruction (SI) has been recognized as a source of significant theoretical and practical potential (Chamot, 2001, 2005; Cohen, Weaver, & Li, 1998; Dörnyei, 1995; R. Ellis, 2005; Grenfell, 2007; McDonough, 1995). Defined as explicit instruction on specific practices or techniques that can be employed autonomously to improve one’s L2 learning and/or use (Y. Chen, 2007; Ellis & Sinclair, 1989; Taylor, Stevens, & Asher, 2006; Tudor, 1996), SI (also known as learner training [Reiss, 1981; Tudor, 1996] and strategies-based instruction [Cohen, 1998]) ultimately aims to “empower students by allowing them to take control of the language learning process” (Cohen, 1998, p. 70).

In addition to intuitive appeal and popularity among L2 teachers, empirical support for SI has been found across many learning contexts (e.g., foreign and second language), treatments (e.g., number and type of strategies, length of intervention), and outcome variables, such as reading and writing (e.g., El-Koumy, 1999; Lawson & Hogben, 1998). The overall results, however, are hardly conclusive. Studies of SI have also produced negative and mixed results across many of the same contexts, treatments, and outcome variables (e.g., Bimmel, van den Bergh, & Oostdam, 2001; Maxim, 2002). Moreover, doubts about the effectiveness of SI have been raised due to methodological flaws in previous research (e.g., small sample sizes, nonrandom group assignment, exclusion of comparison groups), the complexity of variables that affect L2 strategy use, uncertainty of long-term effects, cost/benefit ratio concerns, nonempirically justified strategies, a lack of valid and reliable instruments, and the absence of a comprehensive theory (Bimmel et al., 2001; Chamot, 2005; Dörnyei, 1995; Green & Oxford, 1995; Gu, 1996; Hassan et al., 2005; Kimura, Masuhara, Fukada, & Takeuchi, 1993; Macaro, 2006; Macaro & Erler, 2007; Manchón, Roca de Larios, & Murphy, 2007; McDonough, 1995; Nakatani, 2005; Nakatani & Goh, 2007; Oxford, 1993a; Purpura, 1998; Rees-Miller, 1993; Thompson & Rubin, 1996; Walters, 2006). Furthermore, although many articles and book chapters have provided suggestions for and critical accounts of SI (Chen, 2007; Cohen et al., 1998; Dörnyei, 1995; Hassan et al., 2005; McDonough, 1995; Rubin, Chamot, Harris, & Anderson, 2007; Thompson & Rubin, 1996), no comprehensive meta-analysis of the research in this area has
been carried out to date. The first goal of this article is, therefore, to obtain a reliable, quantitative measure of the effectiveness of SI. Second, this study seeks to determine the relationship between the effectiveness of SI and different L2 contexts, treatment characteristics, and outcome variables.

**Literature Review**

Learning strategies research began with the study of “good” language learners. Researchers in this area (Naiman, Fröhlich, Stern, & Todesco, 1996; Rubin, 1975; Rubin & Thompson, 1994; Stern, 1975), initially concerned with explaining the variability in success among L2 learners, sought to describe the characteristics and practices of successful language learners in the hopes of understanding them and passing them on to less successful learners. However, the qualities of “good” language learners were found to vary, which prevented the formulation of a precise prescription for how to learn a language successfully (Gan, Humphreys, & Hamp-Lyons, 2004; Politzer & McGroarty, 1985; Stevick, 1989).

The focus shifted, then, from “good” language learning to L2 strategies. Defined as the steps taken on the part of language learners to improve their L2 learning and/or use1 (Oxford, 1990), strategies proved a more worthwhile pursuit because of the complexity involved in L2 learning and the broad applicability of strategies to different learners, tasks, and contexts (Chamot, 1993; Gillette, 1987; Griffiths & Parr, 2001; Hong-Nam & Leavell, 2006). Research on L2 strategies, including SI, took on an individualized conceptualization of language learning, which was welcomed as part of a movement in education toward learner-centeredness (see Nunan, 1988; Nyikos, 1996; Tudor, 1996; Wenden, 2002). In addition to connecting to education through a common concern with learners on an individual level, strategies received increased attention in second language acquisition (SLA) as models of communicative competence began to include a component of strategic competence (Bachman, 1990; Bachman & Palmer, 1996; Canale, 1983; Canale & Swain, 1980; Dörnyei & Thurrell, 1991). Additionally, although strategic competence originally referred only to communication (i.e., listening and speaking) strategies employed to compensate for linguistic deficiencies (Canale & Swain, 1980), proponents of SI perceived this line of inquiry as theoretical backing for their more pragmatic agendas relating to many different types, uses, and applications of strategies. With support from these and other areas (e.g., self-regulatory learning in educational psychology) (cf. Tseng, Dörnyei, & Schmitt, 2006; Wenden, 1998),
teachers and researchers alike have used SI to both practical and theoretical ends.

As stated earlier, the effects of SI have been tested widely with different learning contexts, treatments, and outcome variables. However, secondary research on SI has presented accounts that lack comprehensiveness, only describe SI across a single dimension (e.g., outcome variables), and/or are systematic but somewhat coarsely grained compared to meta-analytic methods. Despite the over 100 articles and books chapters providing reviews and methodological recommendations for this type of intervention, no solid models have emerged, and the accumulated findings have been largely mixed. Therefore, in line with this study’s dual purpose of describing the overall effectiveness of SI research as well as determining the relationships between SI and other variables that may moderate its effectiveness, the remainder of this literature review will present the breadth of SI research across the following dimensions: learning contexts, treatments, and outcomes (i.e., dependent variables).

**Contexts**

As Carrell (1998) pointed out, successful strategy use and, consequently, strategy instruction are context-dependent. In other words, effective SI may depend on the context in which strategies are taught and used. With this in mind, this subsection describes several contextual variables hypothesized to moderate the effectiveness of SI.

**Proficiency**

In order to understand how proficiency might affect strategy instruction, it is also helpful to consider how it affects strategy use. A significant amount of research has studied the two, and the results have generally described their relationship as positive and linear (Cohen, 1998; Green & Oxford, 1995; Park, 1997; Wharton, 2000). However, some studies have found a curvilinear relationship between the two, characterized by more frequent use of strategies by intermediate learners than low- or high-proficiency learners (Corrales, 1989; Hong-Nam & Leavell, 2006; Phillips, 1991). Oxford (1999) explained this finding by suggesting that after attaining a certain level of competence in the L2, more advanced learners’ use of strategies becomes increasingly automatic and therefore less available to conscious inspection. Mixed results such as these underscore the difficulty in predicting the effectiveness of SI at different levels. A related challenge is the lack of research that has controlled for proficiency. Most studies have investigated the effects of SI with beginner (e.g., Morin, 2003), intermediate (e.g., Cohen, Weaver, & Li, 1996), or, to a lesser extent, advanced and mixed-proficiency groups (e.g., Linan-Thompson, Vaughn,
Hickman-Davis, & Kouzekanani, 2003; Walters, 2006). Ikeda and Takeuchi (2003) and Moore and Surber (1992) are among the few who have tested the relative effects of SI at different levels, as recommended by Ikeda (2002). In both studies, the higher proficiency group was able to benefit more from the treatment than the lower proficiency group. Ikeda and Takeuchi explained that this finding may be due to lower and higher level processing. Chularut and DeBacker (2004), who also compared learners at different proficiencies, however, did not find an advantage for SI in more advanced learners (see also Maxim, 2002; Vandergrift & Tafaghodtari, 2010). The synthetic nature of the present study allows for comparisons to be made between different proficiency levels despite the lack of control for this variable in primary studies.

Second Versus Foreign Language
In addition to various proficiency levels, SI has also been tested in both second-language (e.g., Fraser, 1999; O’Malley et al., 1985) and foreign language environments (e.g., Barnett 1988a, 1988b; El-Koumy, 1999). Due to logistical constraints, primary research has paid only minimal attention to differences between second and foreign language strategy use and the comparative effects of SI in both contexts. Riley and Harsch (1999), in a study unique in its consideration of L2 students in both learning environments, found no effect for their treatment, although pretreatment differences did exist between the two groups.

Age, Level of Education, and Classroom Versus Laboratory
To round out this discussion of context, it should be mentioned that learners’ age, level of education, and experimental context (i.e., classroom vs. laboratory) may also influence the effectiveness of SI. It may be of interest, for example, to determine whether greater (meta)cognitive capacity of adults offers an advantage over children or whether the control afforded by conducting research in a lab enhances the effects of SI. Again, the primary research to date has produced results across all educational levels with children (e.g., Macaro & Erler, 2007), adolescents (e.g., Rodriguez & Sadoski, 2000), and adults (e.g., Song, 1997) in both classroom (Maleki, 2007) and laboratory settings (Liu, Zhang, Wang, Deng, & Niu, 2007). With the exception of Feyten, Flaitz, and LaRocca (1999), however, very few individual studies have controlled for any of these variables.

Treatments
Although there is widespread agreement among language researchers and practitioners on the benefits of L2 strategies, there is much less agreement on how
to foster a student’s ability to employ strategies for language learning and use. To be sure, there are many treatment-related variables to consider, yet very few studies have investigated the relative effects of different methods of SI (Walters, 2004, 2006). Unfortunately, the lack of theory in this area has left researchers and practitioners to design studies of SI based largely on convenience, intuition, and/or some level of idiosyncrasy.

Whether the objective is empirical or pedagogical, one important step in designing a program of SI is deciding which and how many strategies to teach. The variety of strategies and strategy classifications available presents one of the greatest opportunities to researchers interested in SI. That same variety, however, also presents one of the greatest challenges to theoretical and synthetic accounts of this area, as there is no consensus on a particular classification scheme for sensibly and appropriately summarizing results (Chamot, 1993; Chaudron, 2006; Oxford, 1994; Wong, 2005).

Common to several classification schemes, however, is a distinction among cognitive, metacognitive, and social/socioaffective strategies (e.g., Gan et al., 2004; Purpura, 1997). Cognitive strategies, employed while engaging directly with the L2, include activities such as rehearsal, organization, and elaboration (O’Malley & Chamot, 1990; Oxford, 1990). Some examples of cognitive strategies taught in studies of SI are the keyword method (Avila & Sadoski, 1996) and developing questions while reading (Bimmel et al., 2001). Metacognitive strategies, which involve preparation prior to or reflection following L2 contact or use, have been suggested in previous reviews as an essential component to improving L2 performance through SI (Carrell, 1998; Macaro, Graham, & Vanderplank, 2007; Vandergrift, 2003). Two of the metacognitive strategies taught in Dreyer and Nel (2003), for example, include setting goals for reading and planning how to read. Social strategies generally relate to interaction with others. Clarification requests and cooperation with peers are two examples, the latter of which was selected as part of O’Malley et al.’s (1985) early study of SI.

Stemming from educational psychology, the delineation of strategies described above has been used widely in L2 strategies research (e.g., O’Malley & Chamot, 1990). Additional schemes have been developed, however, with more finely and coarsely grained distinctions. Bialystok (1990) and Dörnyei (1995), for example, grouped strategies based on whether their purpose is to assist in L2 learning or use. The skills that certain strategies are used with (e.g., reading, speaking) have also served to sort L2 strategies as taught in studies of SI (Chamot, 2005; Cohen, 1990; Cohen & Macaro, 2007; McDonough, 1995; Oxford, 1994); such a system is convenient for comparisons across studies of
SI with skill-specific outcome variables. Finally, regardless of which strategy classification is adopted, many researchers in this area suggest that only strategies with empirical support for their effectiveness should be taught and that the rationale for choosing them should be communicated to the learners (e.g., Chamot & Rubin, 1994; Rivera-Mills & Plonsky, 2007).

In addition to different types of strategies, the effectiveness of SI may also be moderated by the length of the treatment (Carrell, 1998; Manchón, 2007; Nyikos & Fan, 2007). This issue has raised concerns about the cost/benefit ratio of SI. Bialystok (1990) argued that “what one must teach students of a language is not strategy, but language” (p. 147) (see also Kellerman, 1991). Most researchers take a more moderate position, suggesting that class time given to SI must be justified as a worthwhile departure away from other L2 tasks that focus more on the language itself (Feyten et al., 1999; Oxford, 1993a). Meanwhile, for SI to be effective, it must be thorough, allowing time for students to fully understand how and when the strategies can be used, with plenty of opportunities for guided practice (Chamot, Barnhardt, El-Dinary, & Robbins, 1999; Chamot & O’Malley, 1996; Hosenfeld, Arnold, Kirchofer, Laciura, & Wilson, 1981; Nyikos & Fan, 2007; Oxford & Leaver, 1996). Wide variability in this dimension exists across studies. Interventions range from shorter, intensive programs of 5 hours in 2 weeks (Raymond, 1993), for example, to treatment periods that are spread out over longer periods, as in Macaro and Erler’s (2007) 14-month longitudinal study. However long or short, Macaro (2001) suggested that materials be developed with appropriate time allocated to different activities such as awareness-raising, practice, scaffolding, and evaluation (see Carrell, 1998).

Several other judgments must be made relating to how SI is conducted. One concern is whether the strategies are taught as an embedded or separate component of the curriculum, with most experts recommending the former (Chamot & Küpper, 1989; O’Malley & Chamot, 1990; Oxford, 1993b; Oxford et al., 1990; Walters, 2006). Additional considerations include the language of instruction, particularly for foreign language contexts (e.g., Chamot, 2005) and among beginner-level learners (Chamot, 1994), as well as whether the teacher or researcher should impart the intervention. Although researcher-led SI may improve the balance of instruction among experimental groups, familiarity of strategies, and full disclosure of treatment procedures in the written report, the more common and convenient practice of teacher- and teacher/researcher-led interventions (e.g., Y. Chen, 2007; Hsu, 2003; Rodriguez & Sadoski, 2000) may benefit from an established rapport with students and a greater understanding of their backgrounds, needs, and preferences.
Outcome Variables
The discussion so far has been centered on the independent variables of the population of studies in question. In this subsection, the review turns to the dependent/outcome variables that have been used to measure the effects of SI.

The skill most frequently measured as an outcome or dependent variable in studies of SI is reading. Much of the work in this area draws on evidence of successful reading SI from first-language (L1) contexts (see Dymock, 2007; Walters, 2004). Taylor et al.’s (2006) meta-analysis of reading SI, replicated by a subset of the data in the present study, yielded overall positive evidence in favor of interventions involving L2 reading SI as well (Hedges’s $g = 0.54$).

Second language writing has also been the object of interventions aimed at improving students’ strategies and abilities. Sengupta (2000), for example, trained foreign language high school students on revision strategies. Although somewhat scarce (Manchón et al., 2007), studies that have focused on SI for writing include Bishop (2001) and Ching (2002).

Several studies have also sought to improve learners’ verbal communication by teaching them strategies for speaking and listening. It has been suggested that the real-time constraints inherent to these skills may limit the teachability of strategies for this mode (Chamot, 2005; Farrell & Mallard, 2006). Furthermore, the social/interactive nature of tasks that involve more than one interlocutor can obscure the outcomes of SI for speaking and listening (Nakatani & Goh, 2007). Despite these challenges, an active line of research has attempted, with varying degrees of success, to train students on strategies that can be employed with these skills (e.g., Y. Chen, 2005; Lam, 2006; Rubin, 1990; Scullen & Jourdain, 2000; Thompson & Rubin, 1996).

A number of studies have also sought to develop learners’ ability to acquire new words. Some of these strategies are meant to help learners resist behaviors such as rote memorization that stem from beliefs that reduce language learning to habit formation (Abraham & Vann, 1987; Fan, 2003; Graham, 1997; Griffiths, 2007; Horwitz, 1987, 1988; Kern, 1995; Mori, 1999; Schulz, 2001). Drawing on current models of vocabulary acquisition (Pavičić, 2008; Schmitt, 2000), many studies have aimed to improve students’ efforts in this area through strategies that promote deeper processing (e.g., Avila & Sadoski, 1996; Brown & Perry, 1991; Fraser, 1999).

In addition to the skill-specific outcomes discussed above, many studies have included measures of other dependent variables, such as (a) students’ beliefs, attitudes, and awareness of strategies themselves or of language learning more generally (e.g., Bull & Ma, 2001; Chamot, 1993; Laoire, 2007; Sengupta, 2000); (b) autonomy (Y. Chen, 2007); (c) frequency and variety of strategy use
(Chamot, 1993; Cohen et al., 1996; Flaitz & Feyten, 1996; Ikeda & Takeuchi, 2003; Macaro & Erler, 2007); (d) overall proficiency (Feyten et al., 1999); (e) grammatical accuracy (Ayaduray & Jacobs, 1997; Cadierno-Lopez, 1992); and (f) pronunciation (Hazan, Sennema, Iba, & Faulkner, 2005; Maleki, 2007).

Research Questions

The studies cataloged above portray the multidimensional nature of instruction on L2 strategies. Although advances have been made toward understanding the complexity of this type of intervention, no secondary studies have employed the methodological and theoretical rigor needed to locate reliable patterns and suggest a model of SI (e.g., Chamot, 2005). Furthermore, although literature reviews can be helpful in determining trends and gaps in the research, they are often unable to provide answers to empirical questions relating to complex constructs and processes (Lipsey & Wilson, 2001). It is beneficial for the field, therefore, to assess the cumulative findings of mature domains such as SI—an active area of SLA research since the early 1980s and the focus of over 400 empirical, theoretical, and review articles (see Plonsky, 2010)—by means of quantitative, secondary analyses (Norris & Ortega, 2006).

A synthetic investigation of SI by means of meta-analysis can supply the empirical motivation for a working model that explains the extent to which strategies have been taught successfully and the strength of the relationship between SI and different contexts, treatments, outcome variables, and research methods. Taking into consideration these goals as well as recommendations for synthetic research within the field as a whole, the present meta-analysis seeks to answer the following two research questions:

1. How effective is L2 strategy instruction?
2. What is the relationship between the effectiveness of SI and different learning contexts, treatments, outcome variables, and research methods?

Methods

Data Collection and Coding

Prior to collecting data for meta-analysis, a population of studies must be defined and representatively, if not exhaustively, sampled (Lipsey & Wilson, 2001). The criteria that define the body of research to be meta-analyzed must be broad enough to produce robust findings across a number of studies worthy of
Plonsky

Meta-analysis of L2 Strategy Instruction

meta-analysis, yet conceptually (if not operationally) narrow enough to avoid inappropriate aggregation of findings (i.e., the apples and oranges problem; see Cooper, 2009; Plonsky & Oswald, in press). Taking these principles into consideration, this meta-analysis included all studies that met the following eligibility criteria: (a) participants studying an L2, (b) treatment involving instruction on one or more L2 strategies, (c) data collected and compared in a control-experimental (between groups) design, (d) a quantitative measure of the effect of SI as a dependent variable, and (e) a reported effect size (or sufficient data reported to extract an effect size).

The search for studies that met these criteria began with keyword searches. Four academic databases—Linguistics and Language Behavior Abstracts (LLBA), the Educational Information Resource Center (ERIC), PsycInfo, and Google Scholar—were searched using combinations of the following words: (a) learner training, (b) strategies-based instruction, (c) learning to learn, (d) strategies instruction, (e) strategy instruction, (f) second language, and (g) strategies. In addition, forward citations were retrieved for seminal papers (e.g., O’Malley et al., 1985) using Web of Science. These searches produced thousands of studies, reviews, books, and Web pages of researchers, and references from relevant sources were then examined for additional studies. The cycle of locating studies, reference consultation, and study retrieval continued until the search was believed to be exhaustive (i.e., only producing duplicate studies) within the parameters of the inclusion criteria.

Once collected, the studies were organized into a database and coded to determine which moderators and study effects would be numerous enough to allow for meta-analysis. The coding scheme (see Table 1), designed to reflect the dimensions of SI discussed in the review of the literature, was developed through an iterative process that considered previous reviews of SI (e.g., Hassan et al., 2005; McDonough, 1999), recommendations for best practices of SI (e.g., Harris et al., 2001; Oxford et al., 1990), methodological and theoretical commonalities and idiosyncrasies within studies of SI, and other related but previously understudied or unstudied phenomena. To ensure that data were obtained reliably, a subset of 10 studies (approximately 16% of the total 61) was coded by a second rater, who was trained on the conventions and operationalizations of the coding scheme. The overall agreement between the two raters was 96%. Item-by-item agreement was also calculated to determine whether greater definitional clarity was needed in any particular area (Orwin & Vevea, 2009), and agreement was found to be at or above 90% on all items.
Table 1 Data coded from primary studies

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification</td>
<td></td>
</tr>
<tr>
<td>Author</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td></td>
</tr>
<tr>
<td>Year of publication</td>
<td></td>
</tr>
<tr>
<td>Type of publication</td>
<td>Article</td>
</tr>
<tr>
<td>Context</td>
<td></td>
</tr>
<tr>
<td>Second or foreign language</td>
<td>Second</td>
</tr>
<tr>
<td>Participants’ age (average)</td>
<td></td>
</tr>
<tr>
<td>Educational institution</td>
<td>Elementary</td>
</tr>
<tr>
<td>Classroom or laboratory</td>
<td>Class</td>
</tr>
<tr>
<td>Country</td>
<td></td>
</tr>
<tr>
<td>Participants’ L1</td>
<td></td>
</tr>
<tr>
<td>Participants’ L2</td>
<td></td>
</tr>
<tr>
<td>Proficiency level</td>
<td>Beginner</td>
</tr>
<tr>
<td>Method</td>
<td></td>
</tr>
<tr>
<td>Pretest?</td>
<td>Yes</td>
</tr>
<tr>
<td>Control group?</td>
<td>Yes</td>
</tr>
<tr>
<td>Random assignment?</td>
<td>Yes</td>
</tr>
<tr>
<td>Delayed posttest?</td>
<td>Yes</td>
</tr>
<tr>
<td>Groups equal before treatment?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Variablea</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td></td>
</tr>
<tr>
<td>Strategies taught (general)</td>
<td>Cognitive</td>
</tr>
<tr>
<td>Strategies taught (specific)</td>
<td></td>
</tr>
<tr>
<td>Number of strategies taught</td>
<td></td>
</tr>
<tr>
<td>Empirical justification of strategies?</td>
<td>Yes</td>
</tr>
<tr>
<td>Rationale explained to Ss</td>
<td>Yes</td>
</tr>
<tr>
<td>Length of treatment (days/weeks)</td>
<td></td>
</tr>
<tr>
<td>Length of treatment (hours)</td>
<td></td>
</tr>
<tr>
<td>Treatment provider</td>
<td>Teacher</td>
</tr>
<tr>
<td>Experience/training of teacher</td>
<td></td>
</tr>
<tr>
<td>Embedded into curriculum</td>
<td>Yes</td>
</tr>
<tr>
<td>Language of treatment</td>
<td>L1</td>
</tr>
<tr>
<td>Opportunities for practice</td>
<td>Yes</td>
</tr>
<tr>
<td>Outcome(s)</td>
<td></td>
</tr>
<tr>
<td>N-size control group</td>
<td></td>
</tr>
<tr>
<td>N-size treatment group(s)</td>
<td></td>
</tr>
<tr>
<td>Dependent variableb</td>
<td>R</td>
</tr>
<tr>
<td>Measure/instrument</td>
<td></td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
</tr>
<tr>
<td>Statistical tests used</td>
<td></td>
</tr>
<tr>
<td>Means</td>
<td></td>
</tr>
<tr>
<td>Standard deviations</td>
<td></td>
</tr>
<tr>
<td>Effect sizes</td>
<td></td>
</tr>
</tbody>
</table>

aVariables without labeled values are continuous, noncategorical, or open-ended.

bR = reading, W = writing, L = listening, S = speaking, V = vocabulary, P = pronunciation; O = overall L2 ability, U = strategy use, A = attitude.
Analysis

As is customary in meta-analysis, effect sizes (Cohen’s $d$) served as the basis for all quantitative analyses. Although eight studies in the final sample reported an effect size statistic, none reported a $d$-value, and effect sizes used in the analysis were calculated from the reported means, standard deviations, and sample sizes. Thirty-six studies lacked one or more of these statistics, and requests were sent to 15 authors, only 5 of whom responded with the missing data. The remaining studies were excluded from the analysis. In total, 218 studies of SI were identified, 157 of which were excluded due to one or more of the following reasons: missing data (e.g., Carrell, Pharis, & Liberto, 1989; Nunan, 1997b; Paige, Cohen, & Shively, 2004; Sasaki, 2000; Seo, 2005), lack of control/comparison group (e.g., Hagaman & Reid, 2008; Salataci & Akyel, 2002; Shen & Huang, 2007; Wyra, Lawson, & Hungi, 2007), qualitative outcomes (e.g., Cotterall, 1990; Coyle, 2007; Lam & Wong, 2000; Nunan, 1996, 2002), strategies taught for purposes other than L2 learning or use (e.g., Amer, 1993; Min, 2005), treatment conditions that focused on language learning (as opposed to the learning of strategies) (e.g., Cheung, 1997; Sagarra & Alba, 2006), L2 learner data combined with L1 data (Gaskill & Murphy, 2004; Lovett et al., 2008; Muñiz-Swicegood, 1994), strategies used to teach the L2 but not explicitly taught to participants (e.g., Bahr & Dansereau, 2005; Barcroft, 2002; Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Desrochers, Gelinas, & Wieland, 1989; Desrochers, Wieland, & Coté, 1991; Sanchez, 2004), and duplicate data reported elsewhere (Cohen, Weaver, & Li, 1995, 1998; Flaitz & Feyten, 1996; Graham & Macaro, 2008; Nunan, 1997a). Thirty-two dissertations including SI were also identified but excluded because they were not freely accessible (e.g., Aninao, 1993; Holunga, 1994; Kohler, 2002; Ozeki, 2000). A total of 61 individual reports were included in the final sample.

Multiple Effect Sizes

From a meta-analytic perspective, it is straightforward when each primary study produces a single effect size. After collecting and coding the data, however, it was apparent that many of the studies to be aggregated had produced more than one measure of the treatment effect. Referred to as stochastic dependency, this condition is present in meta-analysis when studies include multiple outcome measures taken by the same participants (see Gleser & Olkin, 2009). Several well-defined procedures are available for dealing with stochastically dependent data. One approach randomly selects a single effect size per study but, in doing so, systematically discards valuable data. A second approach involves calculating an average of the effect sizes for each study (Lipsey & Wilson, 2001); this
technique was applied when multiple effect sizes came from the same group of participants. Thus, the average effect from each sample/experimental group was treated as an independent data point, separate from effect sizes from different groups of participants within an individual study (see Lipsey & Wilson, 2001, pp. 112–113). In total, effect sizes were obtained from 95 unique samples and 6,791 participants.

**Weighting of Effect Sizes**

As is often the case in meta-analytic research, the coding process brought to light several limitations of SI research (e.g., unreported reliability of measures). Whereas some meta-analysts advocate exclusion of studies that fail to meet certain methodological standards (e.g., Slavin, 1986; see also Valentine, 2009), all studies that met the criteria were included, thus avoiding a necessarily subjective determination of quality while gaining access to a greater portion of the accumulated research. That is not to say, however, that the issue of quality has been ignored. Absorbing into the analysis different methods and weaknesses that are inherent to primary studies presents a problem that can be addressed empirically and/or mitigated by the volume of the sample (Lipsey & Wilson, 2001; Norris & Ortega, 2006; Oswald & Plonsky, 2010; Plonsky, 2011; Valentine, 2009). Hunter and Schmidt (2004), for instance, devised a set of techniques for weighting effect sizes based on minimally subjective assessments of quality, such as measurement reliability and range restriction. Accounting for these and other statistical artifacts would likely show that the effects of SI have been somewhat misrepresented by understated and inconsistent results (Oswald & McCloy, 2003). Due to the lack of reporting on reliability in primary studies (25 of the 61 studies), however, weighting of the effect sizes in this study was based solely on individual sample sizes. By employing this straightforward adjustment, the analysis presents an appropriate reflection of the statistical power of larger sample sizes while preserving interpretability and clarity of findings.

**Averages**

Once effect sizes were coded, combined, and weighted, the research questions were addressed. The first step was to find the weighted mean effect size and confidence intervals for the population of studies included in the meta-analysis; this simple calculation was used to answer research question 1, which pertained to the overall effectiveness of SI. Research question 2 sought to describe any systematic variability among studies of SI. Average effect sizes and confidence intervals were calculated again for subgroups formed on the basis of moderating variables identified a priori. In order to group and compare average effects, four continuous variables were collapsed into dichotomous categories: (a) age
Figure 1 Funnel plot of unweighted effect sizes ($M = 0.58; k = 95$) and sample sizes. ($<12$, $\geq 12$; this break was chosen to reflect assumptions about differences in how children and adults learn languages); (b) proficiency level (beginner, intermediate/advanced); (c) number of strategies taught ($\leq 8$, $> 8$; most studies taught either a small number of strategies up to about 8 or much larger set of strategies); and (d) length of treatment ($\leq 2$ weeks, $> 2$ weeks; this cutoff reflects and tests the difference between approximately half of the studies that were carried out over a few days and those that were carried out over several weeks or months).

Publication Bias
As described earlier, this study intended to include a representative if not exhaustive population of studies. Several factors, however, may have contributed to a bias in observed values. The symmetry of a funnel plot provides a general indication of whether the sampled studies reflect publication bias resulting from suppression of nonstatistically significant findings on the part of journals and/or individual researchers (Rosenthal, 1979; Rothstein, Sutton, & Borenstein, 2005; Vevea & Woods, 2005). Further motivation for an analysis of the file-drawer problem stems from the positive orientation toward SI expressed either implicitly or explicitly by many researchers.4

The funnel plot in Figure 1 displays a slightly higher number of effect sizes to the right of the aggregated (unweighted) average ($d = 0.58$). This distribution indicates some evidence for publication bias in the population of primary studies. If no bias were present, the plots would be more equally weighted on both sides of the mean. Nevertheless, it is important to remember that publication bias is only one of the variables that may contribute irregularities...
Table 2  Overall effects of SI ($d$)

<table>
<thead>
<tr>
<th></th>
<th>$M$</th>
<th>$K^a$</th>
<th>$SE$</th>
<th>95% Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control vs. Experimental group$^b$</td>
<td>0.49</td>
<td>95</td>
<td>0.02</td>
<td>0.44 to 0.53</td>
</tr>
</tbody>
</table>

$^a$Unique samples.

$^b$Sixty-one unique studies; 6,791 individual participants.

(e.g., asymmetry) in a funnel plot. Others include heterogeneity in the sample (i.e., the presence of moderators) and issues associated with the validity and reliability of outcome measures (Sterne, Becker, & Egger, 2005). In addition to assessing the presence of publication bias, the funnel plot also provides an indication of the proximity of the observed mean to the true population mean. As sample size increases, standard error is expected to decrease, thus yielding a more accurate measure of the treatment effect. Likewise, as we see in this plot, studies with smaller samples exhibit greater variance in effect sizes, whereas larger sample sizes (i.e., those higher up on the plot) cluster closer to the mean.

Results

The first research question addresses the effectiveness of SI by aggregating effects from the population of primary studies. The overall weighted mean ($d = 0.49$; see Table 2) represents a medium effect on Cohen’s (1988) scale but would be characterized as small to medium with respect to Oswald and Plonsky’s (2010) proposed standards for interpreting effect sizes in SLA (see also Wolf, 1986). Benchmarks aside, the overall weighted mean tells us that, on average, the experimental participants scored approximately 0.5 standard deviations above control group participants on the outcome measures. (See the appendix for mean effect sizes from all 95 contributing samples.)

Hall, Tickle-Degnen, Rosenthal, and Mosteller (1994) told us that the purpose of meta-analysis is to “summarize and add new knowledge” (p. 24–25, emphasis added). Therefore, in addition to obtaining a measure of the overall effect, this study was concerned with locating systematic differences in the effectiveness of SI across certain substantive and methodological features. To address this question, average effect sizes were calculated for subgroups formed by the study characteristics that had been coded (see Table 1). Moderator analyses were carried out for those items discussed in previous research to potentially affect the outcome of SI. Table 3 presents the weighted moderator analyses.
### Table 3 Effectiveness of SI by subgroups

<table>
<thead>
<tr>
<th>Group</th>
<th>Subgroup</th>
<th>Value</th>
<th>95% Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td><strong>Context</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Setting</td>
<td>L2</td>
<td>0.84</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>FL</td>
<td>0.46</td>
<td>0.41</td>
</tr>
<tr>
<td>Age</td>
<td>&lt;12</td>
<td>1.29</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>≥12</td>
<td>0.47</td>
<td>0.42</td>
</tr>
<tr>
<td>Institution</td>
<td>Elementary</td>
<td>0.78</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>High school</td>
<td>0.23</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>University</td>
<td>0.55</td>
<td>0.49</td>
</tr>
<tr>
<td>Class</td>
<td>Class</td>
<td>0.43</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>Lab</td>
<td>0.79</td>
<td>0.70</td>
</tr>
<tr>
<td>Proficiency</td>
<td>Low</td>
<td>0.35</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Intermediate/advanced</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td><strong>Treatment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategy type</td>
<td>Cognitive</td>
<td>0.48</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Metacognitive</td>
<td>0.24</td>
<td>0.14</td>
</tr>
<tr>
<td>No. of strategies</td>
<td>≤8</td>
<td>0.69</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>&gt;8</td>
<td>0.28</td>
<td>0.15</td>
</tr>
<tr>
<td>Treatment length</td>
<td>≤2 weeks</td>
<td>0.42</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>&gt;2 weeks</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td><strong>Outcomes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target skills</td>
<td>Reading</td>
<td>0.74</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>0.42</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Listening</td>
<td>0.06</td>
<td>−0.26</td>
</tr>
<tr>
<td></td>
<td>Speaking</td>
<td>0.97</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>0.64</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>Pronunciation</td>
<td>0.70</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Grammar</td>
<td>0.01</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>General</td>
<td>−0.14</td>
<td>−0.41</td>
</tr>
<tr>
<td>Strategy use</td>
<td>1.11</td>
<td>1.01</td>
<td>1.20</td>
</tr>
<tr>
<td>Attitude</td>
<td>0.27</td>
<td>−0.02</td>
<td>0.57</td>
</tr>
<tr>
<td><strong>Method</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>Yes</td>
<td>0.54</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0.39</td>
<td>0.31</td>
</tr>
<tr>
<td>Random groups</td>
<td>Yes</td>
<td>0.65</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0.42</td>
<td>0.33</td>
</tr>
<tr>
<td>Reliability reported</td>
<td>Yes</td>
<td>0.65</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0.42</td>
<td>0.36</td>
</tr>
</tbody>
</table>

*aExperimental participants only.*
Five contextual variables were assessed with respect to their relationship with the effectiveness of SI (see Figure 2). Larger effects were obtained in second language settings and with younger learners over foreign language settings and older learners, respectively. These data, however, must be interpreted with caution due to an outlier \((d = 3.92)\) among a relatively small sample of studies in the L2 context. The effect of studies conducted in postsecondary institutions \((d = 0.55)\) is also more than twice as large as in high school studies \((d = 0.23)\). Finally, SI in laboratories \((d = 0.79)\) produced much larger effects than classroom studies \((d = 0.43)\).

With respect to treatment-related variables (see Figure 3), there is a clear difference in outcomes following instruction on cognitive \((d = 0.48)\) versus metacognitive \((d = 0.24)\) strategies. There is also an advantage for longer interventions as well as those that focus on only a few strategies.

Turning to the effects of SI with different outcome variables, the data indicate that SI is much more effective for certain skills than for others (Figure 4). Overall, medium to large effects were obtained for treatment groups over comparison groups in reading, speaking, vocabulary, pronunciation, and strategy use; more modest effects were found for writing and attitude toward language learning; negligible effects resulted from studies with listening, grammar, and general language ability as dependent variables.

Finally, as mentioned earlier, meta-analysis enables an empirical assessment of the relationship between study quality and obtained effects (see Plonsky,
To that end, average effect sizes were calculated for subgroups based on certain aspects of research design and reporting practices in the population of studies (see Figure 5). Studies that pretested for pretreatment differences between treatment and comparison groups produced a larger effect, on average, than studies that did not pretest ($d = 0.54$ and $d = 0.40$, respectively). Random group assignment, another measure of methodological quality taken to ensure equality between groups, was also found to relate to larger effects. Finally, studies that reported reliability yielded much larger effects ($d = 0.65$) than those that failed to do so ($d = 0.42$).
Discussion

The previous section presented results for the two research questions addressed in this study. The first research question was answered with an estimate of the overall effect of SI (weighted $d = 0.49$; unweighted $d = 0.58$; $K = 95$; $N = 6,791$). As mentioned earlier, the magnitude of this result could generally be described as a small to medium effect. Cohen (1988) and others (e.g., Hedges, 2008), however, cautioned that effect sizes are most accurately understood in relation to other effect sizes. To that end, it is worth noting the strength of this finding compared to the results of Hattie’s (1987, 1992) summary of 304 meta-analyses of a wide variety of topics in educational research, which found an overall average effect of $d = 0.40$. Lipsey and Wilson (1993) also combined the effect sizes from 80 meta-analyses of diverse areas within the domain of educational research in which the comparison groups, similar to most SI studies, received an alternate or traditional treatment. The overall mean from that body of research, $d = 0.34$, was actually somewhat smaller than from L2 SI. Moving closer to the focus of this study, it is interesting to note that Hattie, Biggs, and Purdie’s (1996) meta-analysis of SI in L1 education yielded an overall $d$-value that is nearly identical to that of this study: 0.45.

As a relatively young field, the comparisons that can be made between meta-analytic findings from L2 research are few but growing. Overall effects of meta-analyses in SLA have generally been larger than the effects of SI (see Oswald & Plonsky, 2010). For example, in one of several meta-analyses concerned with corrective feedback, Russell and Spada (2006) found an overall effect of $d = 1.16$ (cf. Li, 2010, which found an overall $d$-value of 0.63). Larger
effects were also found in recent meta-analyses of research on task-based interaction by Keck, Iberri-Shea, Tracy-Ventura, and Wa-Mbaleka (2006) and on CALL glosses by Abraham (2008) \((d = 0.92 \text{ and } d = 0.73, \text{ respectively})\). However, not all effects from SLA have been so large. The results of Spada and Tomita’s (2010) four separate meta-analyses of the relationship between types of instruction and types of L2 features found \(d\)-values ranging from 0.33 to 0.88, and Lee and Huang’s (2008) synthesis of input enhancement found an overall effect of \(d = 0.22\). To summarize, although the overall finding of the present study appears consistent with moderate effects found from other, related disciplines within the social sciences, it may be viewed as modest when compared to other bodies of L2 research that have undergone meta-analysis. As the field of SLA continues to accumulate data and additional effects and relationships are meta-analyzed, a broader base of aggregated findings will be available to compare and interpret results. (See also Plonsky, 2011, for data and discussion on the magnitude of effects found across the field of SLA.)

In addition to assessing the magnitude of its effects, the main finding of this study can also be understood with respect to its relevance for language learning and teaching from both theoretical and practical standpoints. Although some theories of L2 acquisition consider the importance of strategies (e.g., R. Ellis, 1995; McLaughlin, 1987; Young & Perkins, 1995), there has been much less concern with the role of SI beyond general assertions that cognitive processes cannot be taught (e.g., Bialystok, 1990; Poulisse, 1993). Future models of language learning, particularly within instructed SLA, should therefore account for an accelerated rate of acquisition when learners are taught to self-regulate using strategies. This study also carries important implications for language practitioners, such as an empirical justification for integrating learner training programs into L2 curricula. Scholars in this area, however, emphasize the importance of context-specific teacher training for SI (Chamot & Küpper, 1989; Chamot & Rubin, 1994). Concomitantly, caution must be exercised in designing SI programs due to the range of variables found to relate to L2 strategy use and the effectiveness of SI (e.g., Lujan-Ortega & Clark-Carter, 2000; Rossiter, 2005; Yamamorio et al., 2003).

Results from the subgroup analyses indicate a surprisingly clear picture of how different contexts, treatments, outcomes, and aspects of study quality relate to the effectiveness of SI. With respect to increased effects of SI for students of intermediate/advanced over beginner levels of proficiency, it should first be acknowledged that L2 proficiency is a highly complex construct (Bachman & Palmer, 1996), and identifying levels for the purposes of subgroup analyses is
somewhat arbitrary. Furthermore, removing two outliers from the intermediate/advanced subgroup approximates the means and confidence intervals of the two groups. That is not to say that differences in the effects of SI do not exist at different levels (cf. Taylor et al., 2006; Vandergrift & Tafaghodtari, 2010). As discussed in the literature review, several studies have found intermediate and advanced language learners to use strategies more often and more effectively than novices (Wharton, 2000; cf. Shmais, 2003). Similarly, SI appears to yield outcomes for language learning and use in which more advanced students are enabled to more accurately incorporate novel strategies into their strategic repertoire (i.e., the rich get richer). On the other hand, although not addressed specifically by this study, ceiling effects may come into play when strategies are taught to learners who use strategies effectively prior to any formal intervention (S.-Q. Chen, 1990; Yoshida-Morise, 1998). Taken together, these findings highlight the importance of selecting strategies for SI that are both level appropriate and that are based on pretreatment measures of strategy use (Chamot & Rubin, 1994; Grenfell & Harris, 1999; Harris, 2003).

The advantage found for SI in second language over foreign language contexts is perhaps not surprising (see Taylor et al., 2006). In addition to a different set of motivations, second language learners are aided by increased exposure to the target language and greater opportunities to practice their strategies and skills (Dörnyei, 2005). It is also worth noting the substantial difference in the effects of SI obtained in secondary ($d = 0.23$) compared to postsecondary ($d = 0.55$) institutions. This result may be partially explained by range restriction; that is, college students are likely a more homogenous group, thus producing less group variance, which yields larger effect sizes. The large difference between effects from laboratory and classroom studies—the last of the contextual variables to be discussed—can be interpreted in a couple of ways. The data could be viewed in opposition to the claim that SI is most effective when integrated as part of the L2 curriculum (e.g., O’Malley & Chamot, 1990). However, the advantage of laboratory settings is more likely due to inherently greater control over the treatment and other potentially intervening variables. Similar findings have also been observed in previous meta-analyses of L2 research (e.g., Li, 2010; Mackey & Goo, 2007).

Results from aggregated group effects across different treatment-related variables emphasize once again the importance of strategy selection in SI. Although instruction on both cognitive and metacognitive strategies is beneficial, the general finding is that SI is most effective when a less-is-more approach is adopted (Chamot, 1994). Moderator analyses also revealed larger effects for longer treatments (see Hattie et al., 1996, for a discussion of the effects of
length for SI in L1 contexts). From a pedagogical standpoint, these findings may prompt some to question whether the benefits of SI outweigh the costs. Several scholars have leveled such critiques, claiming that time in the classroom is better spent learning and using the language rather than learning how to learn and use the language (Kellerman, 1991; Rees-Miller, 1993). Alternately, this finding reinforces the notion that the greatest gains result from SI when learners are allowed to develop their use of strategies over time (see, e.g., Chamot & O’Malley, 1994).

With medium to large effect sizes for 7 out of 10 dependent variables, the results for subgroups based on outcome variables are mixed but mostly in favor of SI. Although further research may be needed to increase the consistency of results for some outcome variables (e.g., writing, grammar, pronunciation), the lack of instruments tested for reliability may be a greater cause for concern (Dörnyei, 2005; Dörnyei & Skehan, 2003; for examples of how more reliable measures for strategies research can be developed, see Gao, 2005; Pavičić, 2008, ch. 4; Tseng et al., 2006). In addition to the relatively small number of studies reporting estimates of measurement reliability, there is evidence for a relationship between (the reporting of) reliability and the effectiveness of SI (i.e., $d$-values of 0.65 and 0.42 were found for studies that did and did not report reliability of outcome measures, respectively). Interestingly, an identical pattern of effects was also found in subgroups of studies that employed ($d = 0.65$) and did not employ ($d = 0.42$) random group assignment. It is not unreasonable to attribute a portion of these differences in effects to the researchers who carried out and reported on these studies. In other words, it may be that researchers who insist on adhering to superior research practices are more likely to produce studies that are based more accurately on relevant theory and that use more rigorous methods, both of which may contribute to larger effects. Along with concerns about instrument reliability and other indications of quality in primary research, this study has also revealed a number of measures of questionable validity. Maleki (2007), for example, employed a written test to assess the effect of SI for oral communication. Finally, although opportunities to practice strategies were present in nearly all studies, the similarity between practice exercises and dependent measures ranged from identical to starkly different. Unfortunately, it was not possible to code for the degree to which the intervention prepared participants for the specific skill or task measured as the outcome variable. Future reviews might consider training-task similarity as an additional moderator of the effects of SI (see Perkins & Salomon, 1989, for a discussion of this issue as it relates to L1 SI).
Suggestions for Future Research

This article has attempted to provide an indication of the theoretical and practical potential of SI. Like many secondary reviews, however, the completeness of the present study is contingent on a population of primary studies that has yet to provide answers to many questions relating to SI. The burden to fill those gaps rests largely on the community of L2 researchers. However, progress in this area—and in the field of SLA as a whole—will be obstructed unless more researchers choose to adopt a synthetically minded approach to their work (see Norris & Ortega, 2006). Although some decisions about study design in classroom research are necessarily based on convenience, other aspects of carrying out and reporting on studies are entirely within the control of individual researchers. With these sentiments in mind, this article will conclude by offering the following suggestions for SI research, many of which echo previous calls for more general reform of L2 research practices (see, e.g., N. C. Ellis; 2006; Norris & Ortega, 2000, 2006; Plonsky & Gass, 2011):

1. The value of SI depends greatly on whether and to what extent its effects last over time, yet only eight of the studies contributing to this meta-analysis included delayed posttests. Therefore, additional measurements of the persisting effects of SI are needed.

2. Our cumulative knowledge of the effects of SI would also benefit from more thorough reporting of data, including basic descriptive statistics needed to aggregate and compare findings across studies, such as standard deviations, sample sizes, means, confidence intervals, effect sizes, and measurement reliability. In addition to aiding future reviews of SI, improved reporting practices will also enable researchers and consumers of primary studies to contextualize and interpret findings more accurately.

3. More detailed explanations of treatment procedures are needed for findings to be replicated and compared to other studies (see Carrell, 1998). For example, 57 of the 61 studies reported providing opportunities for students to practice the strategies they were taught, yet many of the qualitative and quantitative aspects of that practice were unclear or omitted.

4. In addition to more complete reporting of procedures and results, the entire community of language professionals has much to gain from an increased availability of published and unpublished data. Indeed, the community of L2 researchers might consider taking steps toward dealing with this issue from both the top-down (e.g., by implementing stricter standards for publication) and the bottom-up (e.g., by greater willingness of individuals
to grant access to synthetically minded researchers for the purposes of additional or secondary analyses).

5. Although the body of research on SI is extensive, several gaps in the empirical literature remain. In terms of learner populations, additional studies are needed with second (as opposed to foreign) language, preadolescent, and advanced learners. This study has also revealed an imbalance among outcome variables/skills, with further research needed on the effects of SI on writing (see Plonsky, 2009b), listening, pronunciation, grammar, and attitudes. Although the effects of instruction on certain strategies have been studied extensively (e.g., the keyword method), scores of strategies remain untested (e.g., very few socioaffective strategies have been taught in studies of SI; for an example of socioaffective SI in the L1 context, see Pintrich & de Groot, 1990). Additional studies, particularly in these underresearched areas, will not only clarify the role of potential moderators, but they will also allow us to examine interactions between moderators. Although not yet predicted by theory in this area, it is not unreasonable to expect the effectiveness of SI to be influenced by multiple variables simultaneously. Consider, for example, the difference in literacy skills between child and adult L2 learners and how it might relate to unique effects of SI for reading and writing versus listening and speaking. The larger issue here, however, is one of intentionality. In order to reliably determine relationships (via meta-analysis) between the effectiveness of SI and those variables hypothesized to moderate its effectiveness, decisions about learners, outcomes, strategies, and so forth need to be made deliberately, with due consideration given not only to the predictions of relevant models but also to the extent to which the variables of interest have been studied or perhaps overlooked.

6. Finally, reliance on p-values for determining the significance of an effect presents a major barrier to progress in SI research. This practice, the object of over five decades of debate and controversy in other social sciences (e.g., Abelson, 1997; Balluerka, Gómez, & Hidalgo, 2005; J. Cohen, 1994; Lykken, 1968), has been characterized as an uninformative, unreliable, and arbitrary means of interpreting data (e.g., Norris & Ortega, 2000; Oswald & Plonsky, 2010; Plonsky, 2009a; Schmidt, 1996). Null hypothesis significance testing has surely done the domain of SI research a disservice by ignoring the magnitude of effects and distilling continuous data into a crude yes/no dichotomy. For example, a vote-count of studies of SI—a type of synthesis that tallies and compares the number of studies finding statistically significant results to those that did not (e.g., Cooper, 1998)—would grossly overestimate its effectiveness. Moreover, several studies
contributed practically meaningful d-values from comparisons that failed to reach statistical significance, whereas others that found a statistically significant difference (usually due to larger sample sizes) contributed very modest effect sizes (e.g., Kusiak, 2001).

**Conclusion**

This article presents a meta-analysis of 95 samples from 61 studies of the effectiveness of L2 SI. Several variables were found to moderate that effect, including context, age, proficiency, educational level, setting, type and number of strategies taught, outcome variable, and duration of SI. The study also produced evidence to support claims of a relationship between certain methodological characteristics of primary studies (pretesting, random group assignment, reporting of reliability) and the effects of SI they produce.

In closing, I would like to mention that the intent behind this article is not to provide what would surely be a premature closure on a lasting and productive line of L2 inquiry. On the contrary, the retrospective value of the current study is perhaps no greater than its prospective value. In addition to summary results and interpretations of hypotheses both tested and untested in primary studies, this article has shown the interplay of variables that affect SI to be complex and in need of further study in order to fully understand the many implications of SI for both theory and practice.

Revised version accepted 5 May 2010

**Notes**

1. The definition of a strategy has been the object of much debate. Disagreements exist over (a) whether strategies should be characterized as internal or external mechanisms (Wenden, 1987), (b) whether a strategy is a plan of action or the action itself (Phakiti, 2003), and (c) the definitional equivalency of terms often used to describe or define strategies (e.g., tactic, process, action, technique) (see Oxford, 1990). See Macaro (2006) for a thorough discussion of this issue.

2. This topic also highlights another challenge to secondary accounts of SI—the subjectivity with which individual researchers define the proficiency of their participants.

3. Some claim that communication strategies need not be taught, assuming transfer from the L1 (Kellerman, 1991). Others (e.g., Wolfersberger, 2003) have argued that although transfer of strategies is possible, students need assistance identifying their L1 strategies and applying them to L2 tasks. An anonymous reviewer pointed out, however, that listening strategies may represent an exception to this tendency because they are rarely taught in the L1 or L2.
It should be noted that unpublished studies were not excluded; on the contrary, I was active in seeking out relevant, unpublished research. Nevertheless, citations of fugitive literature are relatively uncommon, and all but a few (two or three) of the studies were peer-reviewed and/or published.

Valentine and Cooper (2003) cautioned that these rules of thumb are meant for the social sciences in general and that J. Cohen (1988) himself recommended that effect sizes are most useful when understood in relation to other, field-specific effects. For example, it would not be appropriate or particularly useful to interpret the magnitude of effects from education compared to economics or physiological psychology. Oswald and Plonsky (2010), likewise, insisted that the same principle can be applied to findings from different domains of SLA research, which must develop their own standards for practical significance.

In order to facilitate replication and/or reanalysis, the dataset used in this study will be made available upon request.

When the outlier is removed, the average effect size for second language and foreign language settings is roughly equal ($d = 0.61$ and 0.46, respectively), with overlapping confidence intervals.

For practical reasons, most SLA research involves the use of intact classes. Although convenient for classroom-based research, intact classes can restrict the research design by impeding random assignment to experimental conditions (Hatch & Lazaraton, 1990). One of the aims of this article was to measure the relationship between the effectiveness of SI and certain aspects of quality. Therefore, for the purposes of this study, random group assignment was designated to any study that randomly assigned participants, whether individually or as a class, to a treatment or control condition.

References

References marked with an asterisk indicate studies included in the meta-analysis.


Plonsky

Meta-analysis of L2 Strategy Instruction


### Appendix

Effects and $n$ sizes from All 95 Contributing Samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>$n$</th>
<th>$d$</th>
<th>Sample</th>
<th>$n$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atay &amp; Ozbulgan (2007)</td>
<td>25</td>
<td>0.97</td>
<td>Macaro &amp; Erler (2007)</td>
<td>62</td>
<td>0.62</td>
</tr>
<tr>
<td>Avila &amp; Sadoski (1996)</td>
<td>18</td>
<td>0.72</td>
<td>Maleki (2007)</td>
<td>30</td>
<td>1.22</td>
</tr>
<tr>
<td>Ayaduray &amp; Jacobs (1997)</td>
<td>16</td>
<td>3.56</td>
<td>Moore &amp; Surber (1992)</td>
<td>25</td>
<td>0.78</td>
</tr>
<tr>
<td>Barnett (1988a)</td>
<td>108</td>
<td>0.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barnett (1988b)</td>
<td>51</td>
<td>0.20</td>
<td></td>
<td>13</td>
<td>0.80</td>
</tr>
<tr>
<td>Bejarano et al. (1997)</td>
<td>15</td>
<td>0.38</td>
<td></td>
<td>28</td>
<td>0.84</td>
</tr>
<tr>
<td>Bimmel et al. (2001)</td>
<td>12</td>
<td>0.26</td>
<td></td>
<td>11</td>
<td>−0.17</td>
</tr>
<tr>
<td>Bouvet &amp; Close (2006)</td>
<td>5</td>
<td>1.33</td>
<td></td>
<td>7</td>
<td>0.03</td>
</tr>
<tr>
<td>Campos et al. (2004)</td>
<td>91</td>
<td>0.57</td>
<td>Morin (2003)</td>
<td>15</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>91</td>
<td>0.76</td>
<td>Morin (2006)</td>
<td>14</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>91</td>
<td>0.66</td>
<td>Morin &amp; Goebel (2001)</td>
<td>18</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>0.45</td>
<td></td>
<td>31</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>0.39</td>
<td>Najar (1999)</td>
<td>135</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>0.60</td>
<td>Nakatani (2005)</td>
<td>28</td>
<td>1.09</td>
</tr>
</tbody>
</table>

*Continued*
Appendix
Continued

<table>
<thead>
<tr>
<th>Sample</th>
<th>n</th>
<th>d</th>
<th>Sample</th>
<th>n</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrell (1985)</td>
<td>14</td>
<td>1.54</td>
<td>Naughton (2006)</td>
<td>24</td>
<td>1.56</td>
</tr>
<tr>
<td>Chularut &amp; DeBacker (2004)</td>
<td>20</td>
<td>1.33</td>
<td>O’Malley et al. (1985)</td>
<td>27</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>3.92</td>
<td></td>
<td>26</td>
<td>0.27</td>
</tr>
<tr>
<td>Cross (2009)</td>
<td>7</td>
<td>−0.76</td>
<td>Pappa et al. (2003)</td>
<td>30</td>
<td>0.27</td>
</tr>
<tr>
<td>Dornyei (1995)</td>
<td>53</td>
<td>0.36</td>
<td></td>
<td>29</td>
<td>0.52</td>
</tr>
<tr>
<td>Dreyer &amp; Nel (2003)</td>
<td>89</td>
<td>0.78</td>
<td>Pressley et al. (1980)</td>
<td>12</td>
<td>2.47</td>
</tr>
<tr>
<td>El-Koumy (1999)</td>
<td>62</td>
<td>1.74</td>
<td></td>
<td>12</td>
<td>1.33</td>
</tr>
<tr>
<td>Feyten et al. (1999)</td>
<td>84</td>
<td>−1.05</td>
<td></td>
<td>22</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>−1.61</td>
<td></td>
<td>15</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td>126</td>
<td>0.22</td>
<td>Rasekh &amp; Ranjbar (2003)</td>
<td>27</td>
<td>0.98</td>
</tr>
<tr>
<td>Flaitz et al. (1995)</td>
<td>130</td>
<td>0.58</td>
<td>Raymond (1993)</td>
<td>21</td>
<td>0.55</td>
</tr>
<tr>
<td>Gladwin &amp; Stepp-Greany (2008)</td>
<td>8</td>
<td>−0.20</td>
<td>Rao (2007)</td>
<td>39</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>−0.03</td>
<td></td>
<td>39</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>−0.23</td>
<td>Rico (2008)</td>
<td>10</td>
<td>0.75</td>
</tr>
<tr>
<td>Graham (2007)</td>
<td>32</td>
<td>0.02</td>
<td>Rubin (1990)</td>
<td>294</td>
<td>−0.25</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.10</td>
<td>Saif (2008)</td>
<td>36</td>
<td>1.18</td>
</tr>
<tr>
<td>Graham &amp; Macaro (2008)</td>
<td>29</td>
<td>0.52</td>
<td>Schueller (2009)</td>
<td>30</td>
<td>1.47</td>
</tr>
<tr>
<td>Hamp-Lyons &amp; Proulx (1982)</td>
<td>39</td>
<td>0.08</td>
<td></td>
<td>24</td>
<td>0.75</td>
</tr>
<tr>
<td>Hogben &amp; Lawson (1997)</td>
<td>11</td>
<td>0.96</td>
<td></td>
<td>14</td>
<td>−1.47</td>
</tr>
<tr>
<td>Huang &amp; Ma (2007)</td>
<td>18</td>
<td>0.67</td>
<td>Sengupta (2000)</td>
<td>35</td>
<td>−0.27</td>
</tr>
<tr>
<td>Ikeda &amp; Takeuchi (2003)</td>
<td>65</td>
<td>0.65</td>
<td>Soleimani (2008)</td>
<td>48</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0.47</td>
<td>Thomas &amp; Wang (1996)</td>
<td>13</td>
<td>−0.07</td>
</tr>
</tbody>
</table>

Continued
## Appendix

### Continued

<table>
<thead>
<tr>
<th>Sample</th>
<th>$n$</th>
<th>$d$</th>
<th>Sample</th>
<th>$n$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jin (2002)</td>
<td>23</td>
<td>−0.25</td>
<td></td>
<td>13</td>
<td>1.27</td>
</tr>
<tr>
<td>Kern (1989)</td>
<td>26</td>
<td>0.56</td>
<td></td>
<td>14</td>
<td>0.41</td>
</tr>
<tr>
<td>Kern (2005)</td>
<td>26</td>
<td>0.65</td>
<td></td>
<td>14</td>
<td>0.39</td>
</tr>
<tr>
<td>Kimura et al. (1993)</td>
<td>29</td>
<td>0.010</td>
<td>Walters (2006)</td>
<td>14</td>
<td>−0.79</td>
</tr>
<tr>
<td>Kitajima (1997)</td>
<td>13</td>
<td>1.00</td>
<td></td>
<td>11</td>
<td>−0.24</td>
</tr>
<tr>
<td>Kusiak (2001)</td>
<td>78</td>
<td>0.18</td>
<td>Vandergrift &amp; Tafaghodtari (2010)</td>
<td>59</td>
<td>0.57</td>
</tr>
<tr>
<td>Lam (2006)</td>
<td>20</td>
<td>−0.02</td>
<td>Viswat &amp; Jackson (1994)</td>
<td>150</td>
<td>0.22</td>
</tr>
<tr>
<td>Lam (2009)</td>
<td>20</td>
<td>0.55</td>
<td>Zhang (2008)</td>
<td>50</td>
<td>2.71</td>
</tr>
<tr>
<td>Lawson &amp; Hogben (1998)</td>
<td>14</td>
<td>1.39</td>
<td>Zhicheng (1992)</td>
<td>15</td>
<td>0.93</td>
</tr>
<tr>
<td>Liu et al. (2007)</td>
<td>30</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Each row represents a unique sample.