The Journal of English Language Pedagogy and Practice
Vol.10, No.21, pp.184-225, Fall & Winter 2017

The Impressibility of Speaking Accuracy/Fluency among EFL Undergraduates: A Meta-Analysis

Arman Toni [1], Jaleh Hassaskhah [2*], Parviz Birjandi [3]

1, 2, 3. Department of English, Science and Research Branch, Islamic Azad University, Tehran, Iran

*Corresponding author: jaleh_hassaskhah@yahoo.com

Received: 2017.4.3
Accepted: 2017.8.12
Online publication: 2017.12.10

Abstract

This meta-analysis is an investigation into the impressibility of two dimensions of the speaking skill, namely accuracy and fluency, in relation to the experimented treatments among Iranian EFL undergraduates. Having surveyed a collected bank of 74 research reports, the relationships among the variables in the 14 included studies were examined. More specifically, the analysis involved a statistical review of 67 effect sizes (at 95% CI) calculated from studies conducted between 2006 and 2016, including 890 participants. The analysis indicated that in 77% of the experimented treatments, the students performed as well as the students in the regular programs with no significant improvement in oral accuracy. The analysis also demonstrated that 63% of the treatments did not yield significant improvement in oral fluency in comparison with the regular instruction. Moreover, the synthesis of the effects of the contextual factors showed that low-level (i.e., elementary) learners experienced a better improvement in speaking performance. The analysis also revealed that, among the experimented treatments, dialogic tasks were most effective on oral accuracy while interviews were influential in promoting the students' both oral accuracy and fluency. Finally, the quality of the study reports was analyzed and some directions for further research were suggested.

Keywords: speaking, accuracy, fluency, EFL undergraduates, meta-analysis

Introduction

Learning to speak a foreign language is commonly difficult, especially at the initial and intermediate stages of learning, so that learners often resort to thinking-for-speaking patterns, code-switching and other ways of retaining and repairing their speech and avoiding communication breakdowns in the form of deviations from form or meaning or both / accuracy or fluency or both (Robinson & Ellis, 2008). These deviations, referred to as errors (Pienenamm & KeBler, 2011; Ellis, 1994), result in anomalous and *effortful* instances of foreign/second language production.

However, in the course of history, teachers, depending on their beliefs in what the learning of a language is, have treated these deviations differently. As Ferris (2004) put it, while in its early years, the main focus of teachers' errors treatment was learners' linguistic accuracy, the focus shifted in the 1970s to the process of language acquisition rather than the product, and hence a complete neglect of accuracy. This view was opposed by some researchers such as Horowitz (1986) who proposed that the total neglect of form was counterproductive and instead argued that learning merely the process of language acquisition is not enough to familiarize students with the rules of language. Consequently, the interest in the study of errors treatment was renewed (e.g. Semke, 1984), and as such, studies that explored the effectiveness of error treatment were reintroduced.

In this vein of investigation, taking the many aspects of speaking proficiency such as accuracy, fluency, complexity, vocabulary, pronunciation, etc. into consideration, accuracy and fluency have been the focus of prominent interest, assessed as dependent variables in many research studies (Foster & Skehan, 1996; Skehan & Foster, 1997; Mehnert, 1998; Iwashita, Elder & McNamara, 2001; Larsen-Freeman, 2006, among others). One may find two major approaches with respect to the development of speaking proficiency: the view which focuses on the correctness of language, that is, *accuracy-oriented* approach, and the one which considers speech as successful as long as the learner makes oneself understood no matter how incorrect the language, that is, *fluency-oriented*.

Accuracy: The Diverse Concept

The emphasis on accuracy accounts for the production of correct instances of language. On the contrary, inaccuracy is an indication of erroneousness and results in structurally wrong sentences, which endangers the goals of any language curriculum. However, one may often hear the word 'grammar' in various combinations, implying that there has not been any one shared explicit definition of grammar, and hence accuracy. For most linguists, for example, grammar is the system of a language - the language patterns that indicate relationships among words in sentences. As Rivers (1981, p. 68) put it, it is "the rules of a language set out in a terminology which is hard to remember, with many exceptions appended to each rule". At its core, the term grammar refers to either the integral structure of words (i.e., morphology) and sentences (i.e., syntax) in a language, or to the study of this structure published as grammar/structure rules in books.

From a psychological perspective, grammar is the subconscious mental rules which speakers follow to produce language. With a unique use of the word *grammar*, Chomsky (1965) asserts that a child, who has acquired a language, has developed a representation of a system of rules that govern how sentences are to be formed, produced and comprehended. Thus, to Chomsky, grammar is not a property of a language, but of mind-tacit knowledge about what establishes the native language and how it works (Johnson & Johnson, 1999).

From applied linguistics' perspective, as pointed out by Richards and Schmidt (2010), grammar is an account of the system in which linguistic units are combined to create sentences in the language. "The grammatical rules of a language do not tell us what to do. Rather, they tell us how to respond correctly within the structural system of the language" (Pollock, 1997, p. vii). They take into account the meanings and functions of sentences in the overall system of a given language. Therefore, speaking and writing in a language calls for an *accurate* in-depth knowledge of form, sentence structure and grammar system.

Among the different approaches toward accuracy, the applied linguistics' viewpoint has received much attention by Second Language Acquisition researchers and practitioners as well. This might be due to, as Kumaravadivelu (2006) puts it, the "practicality" of the definition of grammar in applied

linguistics as well as its operational definition. In an attempt to explore some practical methods, previous research have advised to tap into accuracy by measuring the percentage of target-like use of plurals (Crookes, 1989), target-like use of vocabulary (Skehan & Foster, 1997), error-free speech (Foster & Skehan, 1996), error-free T-units (Robinson, 1995; Ortega, 1999), error-free AS-units (Lambert & Engler, 2007), and the number of errors per T-unit (Bygate, 2001), to name some.

Fluency: The Controversial Myth

As acknowledged by Freed (2000, cited in Tavakoli, 2011), the notion of fluency is also hard to define, and even though the term is constantly applied within SLA research, there is no general consensus about what is perceived as fluency (Chambers, 1997). Binder (1996), as a case in point, defines fluency as the fluid combination of accuracy and speed which characterizes competent performance. According to Binder (1988, 1996), fluency describes proficient, expert, and automatic performances. In addition, Binder, Haughton and Bateman (2002) posit that fluency is a combination of "quality plus pace". Thus, the notion ranges on the continuum from incompetent performance, lacking accuracy and speed, to total mastery, characterizing perfect quality and pace.

Having explored native and non-native teachers' perceptions of fluency, Kormos and Denes (2004) maintain that fluency is best conceived of as 'skillful performance'. The results of their study suggest that fluency is "primarily a temporal and intonational phenomenon" (Kormos & Denes, 2004, p.158) and that features such as number and length of pauses and speech rate affect fluency judgments. Riggenbach (1991, cited in Tavakoli, 2011) reported that the frequency of unfilled pauses is an indicator of "dysfluency", yet stressed on the differentiation of such pauses based on their function and the place of occurrence.

By and large, what almost all fluency-oriented approaches toward speaking have in common is the belief that meaningful communication is the key in developing the speaking skill. As opposed to accuracy-oriented approaches, the proponents of this viewpoint maintain that grammatical errors are trivial, especially in the initial stages of learning. Moreover, too much emphasis on

error correction is considered harmful, for it may result in extreme monitor in the mind, impeding the natural and normal acquisition of spoken skills (Ebsworth, 1998); as such, speaking a language is not about saying the words in the correct manner, but about achieving a useful pace of performance (Binder et al, 2002).

Empirical Research on Oral Performance in Iran

In the Iranian EFL context, where research on the effects, if any, of various kinds of treatments on learners' accuracy and fluency in speaking abounds, results are also diverse on the part of the examined independent variables. Among the studies, some took an accuracy-oriented approach towards examining the effects of the desirable treatments; some took a fluency-oriented approach, while others looked at the problem from both perspectives.

As for accuracy, in almost all of the studies, it was looked upon from the applied linguistics' point of view, yet from different aspects. Some researchers tended to assess the development of different linguistic forms (e.g., Ansarian & Chehrazad, 2015), some varied the rubrics they employed (e.g., Hazrativand, 2012) and still some others focused on the effects of various types of tasks on the accuracy of Iranian EFL learners' oral production (e.g. Rafie, Rahmany & Sadeqi, 2015). In most of these studies though, the level of accuracy was measured by identifying the number of error-free clauses, which was divided by the total number of clauses produced. The clause in which there was no error in syntax, morphology or word order was counted as an error-free clause. Also, errors in lexis were considered only if a word was nonexistent in English, or if a word was indisputably inappropriate. Thus, it could be argued that the researchers following the accuracy approach based their definition of accuracy on the applied linguistics' point of view. However, as is clear, such studies did not clearly address any specific target structure, nor did they take into account the learners' fluency in speaking.

As for research into fluency or studies with an accuracy-fluency orientation, due to the diverse number of operational definitions and the employed assessment checklists, the variety of methodologies has also been added. As a case in point, Askari and Langroudi (2014) investigated the impacts of the implications of Ur's model on Iranian EFL learners' accuracy and fluency in speaking ability. Out of the five components of Ur's Model, however, the

researchers did not specify as to how they implemented the proposed components and what the treatment exactly included. Moreover, the study did not justify its claim as to how practicing Ur's Model in an EFL class would "lead learners from accuracy to fluency" (p. 84). Shiriyan and Nejadansari (2014) also investigated the implementation of another treatment in the hope of leading to more accurate and fluent oral performances. Based on the findings from the study, deductions were made that exposure to literature-based activities would lead to more accurate as well as fluent L2 oral production. However, the study does not take into account test effects in using a similar speaking ability test.

Of the various number of conducted research into the Iranian EFL learners' oral performance in general and their speaking accuracy and fluency in particular, the results show that there are diverse-even contradictory-conclusions. Not underestimating the contributions of the previous studies, the following remarks could be outlined:

- As for the conceptual framework to define the notion of "grammar", most studies took the applied linguistics' approach, though with a focus on linguistic forms and syntax.
- Almost all study reports applied the term "accuracy" to refer to "grammatical accuracy".
- Different studies tended to base their assessment of accuracy on different rubrics.
- Due to the diverse definitions of fluency, the different studies looked upon fluency from different perspectives, hence different assessment procedures.
- In some studies, no established conceptual framework for fluency was introduced, nor was a fluency rubric presented. This has resulted in heterogeneous conclusions on the part of the examined treatments, which at times calls the validity of the interpretations into question.
- Due to the short interval between the pre-test and the post-test, some studies could not control for the test effect.
- Despite the fact that grammatical accuracy was introduced as the theoretical framework of most studies, some did not address any specific

target structure, nor did they clarify as to what linguistic forms were considered in the course of the study.

• It has well been documented that the learning context, being it the EFL or the ESL, has remarkable differential impacts on the learning outcomes (Oxford & Burry-Stock, 1995; Widdowson, 1997; Nayar, 1997; Nation & Newton, 2009 among others), especially for the speaking skill. Nonetheless, very few studies considered an inclusive review of literature with regard to previous studies conducted in EFL contexts in general and the Iranian EFL context in particular.

• As for study design, the reviewed studies mostly followed the quantitative experimental or the causal-comparative design.

• Finally, in studies with reference to both accuracy and fluency, no specification of the plausibility of the simultaneous facilitation of accuracy and fluency under the conditions of the independent variables was reported.

Regarding the specific problem which triggered the present study, it should be mentioned that similar to other fields of SLA inquiry in Iran, research into speaking has witnessed diverse peaks and troughs among the rising community of EFL learners at language schools and other educational institutions. The concern gets even more critical when it comes to English language departments at universities, where would-be language teachers, studying their bachelor's degree, are required to reach proficiency level to such an extent that they have a satisfactory oral performance since they would be expected to be both accurate and fluent. Moreover, the divergent results (even contradictory in some cases) obtained from previous studies as well as the question of whether their discussions and proposed resolutions would have practical contributions to language teaching led the present researchers to seek to shed light on earlier findings and to conduct an analytical review of the effectiveness of the treatments having been implemented as independent variables.

To the best of our knowledge, no comprehensive reviews, more specifically in the form of meta-analyses, have ever been conducted that address the impacts of different techniques and practices on oral accuracy as well as fluency. To fill this gap, the present meta-analytic review of the effectiveness of

English Language Teaching (ELT) practices on accuracy and fluency was conducted by collecting data from experimental studies and calculating and comparing effect sizes across them. In particular, this meta-analysis focused on the findings of published papers and dissertations databases over the past decade, geared to studies conducted in Iranian universities.

The present study employed meta-analytic techniques, commonly used in the field of medical sciences, to integrate the findings of studies conducted in Iran over the period 2006-2016. More specifically, the purpose of the study was to summarize the results of previous studies into a single estimate to provide a quantitative synthesis of the research literature on EFL oral accuracy/fluency, hoping to open new directions for researchers, practitioners and educators to capitalize on the recent findings regarding the differential effects of various techniques and practices on EFL students' speaking performance.

The rationale for choosing the meta-analytical method of research was the lack of scientifically-based systematic reviews in which evidence on the research question has been systematically identified, reviewed and summarized according to determined criteria to draw inclusive conclusions about the contributions of different practices to EFL learners' oral performance based on the synthesized findings of previous studies. It was hoped that with a close quantitative examination, future research for implementing different practices for promoting oral accuracy and fluency would be more precisely directed.

Several advantages can result from the synthesis of studies on oral accuracy/fluency. First, since all of the studies drew data from university-based classes, the review can offer valuable insight into the effectiveness of proposed interventions in academic contexts. Furthermore, it aids to determine the significance of academic achievement as an outcome of the introduced treatments. Last but not least, it can identify qualitative information with respect to such study features as university type, participants' gender and grade, study design, quality of the experiments, as well as the duration, frequency and timing of the treatment. The specific research questions addressed in this study were as follows:

1. To what extent are study outcomes (accuracy and fluency) affected by the methodology of research?

2. To what extent are study outcomes (accuracy and fluency) the by-products of contextual factors?

3. To what extent do previous primary studies comply with quality standards?

The study attempted to combine both quantitative and qualitative methods in order to provide a more encompassing view of all records available for the time period under study. The quantitative method provided explanations to question one and two, and both the quantitative and the qualitative module sought answers to question three.

Method

The study adopted the meta-analysis as the statistical technique for the procedure. Meta-analytic procedures are statistical techniques used to review and synthesize independent studies in a systematic way within a specific area of research. More specifically, the meta-analysis uses the data set as the unit of analysis and permits tests of hypothesis in terms of associations obtained in the data sets (Masgoret & Gardner, 2004). The outcome of each study experiment is realized as an effect size, being it the difference between the mean for the control group and the mean for the treatment group, divided by the overall standard deviation (Cavanaugh, Gillan, Kromrey, Hess & Blomeyer, 2004).

Inspired by procedures set forth by Hedges and Olkin (1985) and Lipsey and Wilson (2001), the methodology of the present meta-analytical study followed four phases:

- literature search and identification of relevant studies,
- selection of eligible studies and determination for inclusion,
- coding of study reports,
- effect size calculation and data analyses.

In the following, each phase is described in detail.

Phase I: Literature Search and Identification of Relevant Studies

The search for relevant studies was as in-depth as possible. Adopting In'nami and Koizumi's (2010) recommended approach in finding qualified databases used for meta-analyses in applied linguistics, the current literature search was conducted using five databases: Social Science Citation Index (SSCI), Linguistic and Language Behavior Abstracts (LLBA), ProQuest Digital

Dissertation Full Text (PQDT), Educational Resources Information Center (ERIC) and PsycINFO. It was believed that these databases which are frequently used by meta-analysts provided a proper coverage of representative journals.

In addition to the above-mentioned databases and due to the probable lack of the cataloguing of Iranian journals in international databases, the following methods were also taken into account to broaden the literature search:

- Searching web sites known to contain research related to foreign language teaching such as the Ministry of Science, Research, Technology, Ministry of Education, Jahad Daneshgahi Scientific Information Database (SID), Iranian Research Institute for Information Science and Technology (IRANDOC), CIVILCA, and Noor Magazines Database (noormags),
- Searching scholarly journals that may not be indexed (e.g. elmi-tarviji journals),
- Employing general search engines (e.g., Google Scholar, bing, etc.) in keyword searches for other manuscripts that either had not been catalogued in the databases or were currently under review,
- Directly inquiring with a number of researchers known to be actively studying related fields.
- Manually conducting investigation into the references of articles and abstracts, and
- Maximize the scope of studies for consideration by employing search strategies that included a variety of combinations of key terms in such a way that search words included different forms of terms as *oral, speaking, proficiency, production, accuracy, fluency,* as well as *second language acquisition (SLA), second language learning, second language teaching, foreign language learning, foreign language teaching, ESL* and *EFL* (i.e. oral proficiency; speaking accuracy, etc.). It deserves notice that in case, based on the abstract/description of the retrieved article, relevance to the present study could not be determined, it was saved for possible inclusion. The resulting collection included 74 articles.

Phase II: Selection of Eligible Studies and Determination for Inclusion

In order to select eligible studies as well as synthesize the literature for the quantitative phase of the meta-analysis, screening criteria were narrowly specified as follows:

1. The study was published between 2006 and 2016, because it was believed that seminal studies prior to 2006 have already been indexed during this period and their results discussed.
2. The study was reported in English.
3. The publication type of the study was a journal article or a doctoral/master's dissertation.
4. The research focused on Iranian EFL university students.
5. The study had an experimental or quasi-experimental design. Only studies that examined the effect of independent variables on learners' accurate/fluent oral production through an experimental or quasi-experimental design could provide the required data for the analysis. That is to say since qualitative studies, narrative reports, and ex-post facto and other similar designs offer information not acquired in this research, they were excluded from the analysis.
6. The study included one experimental group and one control group or compared the outcome measures of the study before and after the treatment. The effect sizes in the present study were calculated based on the differences between the control group and the experimental group for the studies with two groups and between the pretest and the post-test for the studies with only one group.
7. The dependent variables (i.e., accuracy and fluency) were based on well-established instruments that provided quantifiable data.

After the studies were collected and read, eligibility for inclusion was determined based on the criteria. Any inconsistencies between the researchers were resolved after discussions, so that in total, 14 studies which met all inclusion criteria were selected for the next phase. The rest of the studies were on target respecting the topic, yet either were qualitative, literature reviews or commentary papers that focused on conceptual frameworks such as definitions, approaches, theories, etc., experimental but insufficient in reporting quantitative data (to enable effect size extraction), equivocal in terms of the scope of the

participants of the study, or whose participants were out of focus (e.g., they were carried out in language schools or high schools, among non-Iranian EFL learners or students of others majors than English.) Figure 1 shows the results of the literature search geared to inclusion criteria, and Figure 2 illustrates the studies classified by their measured outcomes.

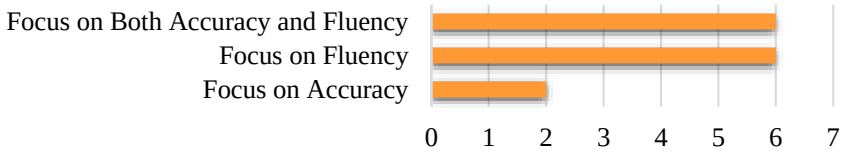

Figure 1. Collected study reports classified by type

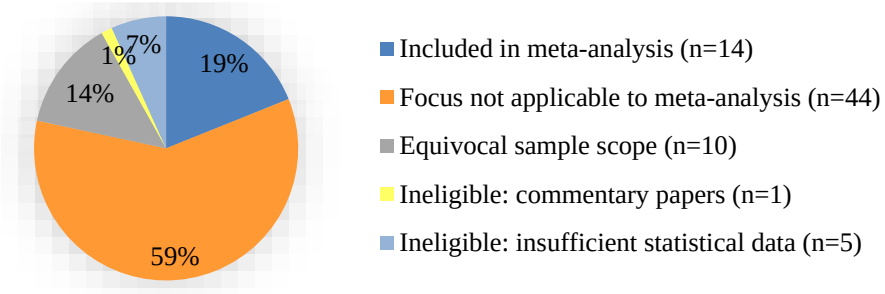

Figure 2. Studies included in the meta-analysis classified by measured outcomes

Phase III: Coding of Study Reports

As a result of the initial search for relevant literature, 74 articles were found. In the second stage, these articles were filtered through the seven screening criteria described above, leaving 14 studies as the body of the meta-analysis for coding. According to Lipsey and Wilson (2001), the coding of study features helps to unravel different factors associated with variations in the

phenomenon from features related to method. In so doing, the three researchers independently coded each study. Afterwards, the coding was discussed among the researchers on a study-by-study basis, and discrepancies, that occurred infrequently, were resolved. Table 1 contains a full description of the variables and the levels included in the coding manual.

Table 1
The Coding Manual

Study descriptors	Feature No.	Feature title	Definition	Codes
Identification	1	Study ID	Assigning an identification number to each study	ID
	2	Author	Last name of the first author	"name"
	3	Publication year	The publication year	"year"
	4	Participants' L2 proficiency level	The participants' second language proficiency level, e.g., low, mid, and high levels	low=1, mid=2, high=3, not reported=0
	5	Participants' grade	Participants' university level	freshman=1, sophomore=2, junior=3, senior=4, higher levels= 5, not reported=0
	6	University type	The participants' university type, e.g., state, azad, non-governmental, payam noor	state=1, azad=2, non-governmental=3, mixed=4, not reported=0
	7	First language (L1)	The participants' mother tongue	persian or not reported=1, azeri=2, kurdish=3, mixed=4
	8	Participants' gender	The participants' gender	female=1, mixed=2, not reported=0
Methodology	9	Sample size	Total sample size of this study	N
	10	Research design	Design of study, e.g. true-experimental, time-series	The time-series=1, factorial=2, posttest-only experimental=3, pretest-treatment-posttest experimental=4,
	11	Randomization	Random sampling	Yes=1, no=2, not reported=0
	12	Instrumentation	The data collected for the analysis in the study	Narration=1, monologic tasks=2, interview=3, dialogic tasks=4
	13	Reliability type	The type of the reliability having been reported	Inter-rater=1, intra-rater=2, test reliability=3, not reported=0
	14	Reported reliability	The different reliability types of the study	N or not reported=0
Intervention	15	Target language features	The language aspect that was examined. e.g., lexical development, grammatical competence	overall proficiency=1, syntactic accuracy=2
	16	Duration	The length of study in weeks	N or not reported=0
	17	Frequency	The number of sessions per week	N or not reported=0
	18	number of treatment sessions	The total number of treatment sessions	N or not reported=0
	19	Session length	The duration of undergone treatment in minutes	N or not reported=0
	20	Class/home activity	The activities in the study	Watching movies=1, listening and retelling=2, topic discussion=3, recording oral productions=4, picture-cued story telling=5, memorizing formulaic expressions=6, not applicable=0
	21	Independent variable (treatment)	The treatment given to the participants in the study	Task planning time=1, task repetition=2, task type=3,

				podcast reconstruction=4, online chat=5, dialog journals=6, speaking strategic planning=7, interaction=8, rote learning=9, metacognitive awareness=10, feedback=11, photomontage=12
	22	Dependent variable	Fluency, accuracy, complexity, other	Fluency=1, accuracy=2, complexity=3, both fluency and accuracy=4, all=5
	23	Measure of fluency	The method of the analysis of fluency	average number of words, T-units and syllables based on Gilabert(2004)=6, calculating syllables, pauses, repetitions and substitutions based on Farrokhi&Mahmoudi(2012)=5, IELTS speaking band score=4, Interview scoring profile, based on Khabiri(2003)=3, Counting repetitions, false starts, etc. based on Skehan&Foster(1999)=2, subjective measurement=1, not reported=0
	24	Significance of development of fluency	Whether the treatment resulted in significant development in fluency	Yes=1, no=2, not reported=0
Outcome measures				error-free verb forms and error-free T-units based on Gilabert(2004)=2, Counting ratio of error-free clauses (errors per t-unit) based on Bygate(2001)=1, not
	25	Measure of accuracy	The method of the analysis of accuracy	reported=0
	26	Significance of development of accuracy	Whether the treatment resulted in significant development in accuracy	Yes=1, no=2, not reported=0
Quality assessment	27	Quality assessment	a quality score based on Effective Public Health Practice Project , 1998	Strong=1, moderate=2, weak=3

The coding manual encompasses five categories of study descriptors, namely identification, methodology, intervention, outcome measures and quality assessment. Among the variables included in the studies, those of intervention (e.g., target language features, class/home activities given and the treatments) and outcome measures (e.g., dependent variables, i.e., whether the focus of a particular study was on accuracy, fluency or both, the measure of accuracy/fluency and the significance of study results) were of particular interest.

As for the "intervention" study descriptor, the focus of the studies in terms of target language feature was mainly on the overall proficiency of the language learners. All in all, studies investigated the impacts of eleven different independent variables, wherein six different activity types were implemented.

The nature of the research in four studies required that no class/home activities were given as instructional materials.

In terms of outcome measures, "accuracy" and "fluency" were operationally defined in various ways across studies. In particular, for fluency, most studies employed Skehan and Foster's (1999) proposed scheme (n=5), and the remaining studies either implemented other rubrics or did not report any measurement criteria (subjective measurements which in turn resulted in overall holistic quality ratings were also grouped as "not reported"). As for "accuracy", most studies employed calculating the ratio of error-free clauses, that is, errors per t-unit, based on Bygate's (2001) scale (n=6), whereas one study did not report any rubric. Table 2 presents the detailed extracted variables and the associated data.

Table 2
Summary of the Coding of the Primary Studies

Feature No.	**Feature title**														
1	Study ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
2	First Author	Rou hi	Birja ndi	Abdi	Hass askh ah	Safar i Vesa l	Asaa dinej ad	Bara dara n	Akef	Mor adi	Farr okhi	Ghe naati	Seifo ori	Maft oon	Aha ngari
3	Publication year	2006	2008	2012	2015	2015	2015	2009	2010	2014	2014	2015	2016	2009	2011
4	Participants' L2 proficiency level	0	2	2	1	2, 3	1	2	1	2	1	2	2	2	2
5	Participants' grade	0	2	0	1	5	0	3	1	1	1	4	2	1	0
6	University type	4	2	2	1	2	2	2	2	1	3	2	2	2	2
7	First language (L1)	2	2	4	1	1	1	1	3	1	4	1	2	1	4
8	Participants' gender	0	2	2	2	1	0	2	0	2	2	2	2	2	2
9	Sample size	37	120	40	33	40	60	52	60	20	45	120	114	109	40
10	Research design	1	2	3	1	1	4	4	4	4	2	4	2	3	3
11	Randomization	0	1	1	0	0	2	0	1	1	0	1	1	1	1
12	Instrumentation	1	2	2	4	2, 3	3	3	3	1	1	1	1	2	2
13	Reliability type	0	1	0	1	0	1	3	1	3, 1	3	0	1	0	0
14	Reported reliability	0	0.99	0	0	0	0.79, 0.9	0.81	0.74	0.84, 0.9	0.77	0	0.87, 0.93	0	0
15	Target language features	1	1	1	1	1	1	1	1	1	1	1	1	2	2
16	Duration	0	1	1	0	2	0	0	16	0	16	0	0	12	1

17	Frequency number of treatment sessions	0	1	1	2	1	0	0	0	0	2	0	0	1	1
18	Session length	0	1	1	0	1	10	20	0	10	20	0	15	10	1
19	Class/home activity	0	0	0	90	25	30	90	0	10	90	0	0	0	0
20	Independent variable (treatment)	1	0	0	3	0	2	3	4	5	6	3	5	3	0
21	Dependent variable	1	2, 3	1	12	3	4	5	6	7	8, 9	4	10	11	1
22	Measure of fluency	5	5	4	5	5	4	1	1	1	1	1	1	2	2, 3
23	Significance of development of fluency	2	2	2	6	2	0	4	3	4	5	1	2	*	*
24	Measure of accuracy	1	1	1	1	2	1	1	1	1	1	1	1	*	*
25	Significance of development of accuracy	1	1	1	2	1	0	*	*	*	*	*	*	1	1
26	Quality Assessment	1	1	2	1	1	1	*	*	*	*	*	*	1	2
27		3	1	3	3	1	1	1	2	2	2	1	3	3	3

Finally, inspired by Shadish and Haddock (1994, cited in Goldberg, Russell & Cook, 2003, p. 9), a variable representing "quality assessment" was computed from the subset of the coded variables. In this regard, for each study, quality assessment was based on Quality Assessment Tool for Quantitative Studies developed by the Effective Public Health Practice Project (EPHPP), an 8-item quantitative scoring method to be applied to assess the quality of Quantitative studies on public health interventions (see the Appendix). The scale was created by building on tools, evidence and expert consensus providing a standardized means to evaluate study quality and develop recommendations for study findings. The final results of the tool, whose reliability and validity were assessed and reported as high by Thomas, Ciliska, Dobbins and Micucci (2004), lead to a rating in:

- selection bias
- study design
- confounders
- blinding
- data collection methods

- withdrawals and dropouts
- intervention integrity
- analysis

Phase IV: Effect Size Calculation and Data Analyses

The meta-analytic part of the data analysis requires the calculation or extraction of effect sizes. To calculate the effect sizes using Cohen's (1977), the post-test data of each experimental group and each control/comparison group were contrasted. It was assumed in case there was more than one post-test in a study, the first, i.e. immediate, post-test was considered for further analyses to control for the effect of retention as an extraneous variable. Moreover, in studies with only one group of subjects, the effect size was computed by the comparison of group conditions before and after being exposed to treatments. This enabled results from multiple-group-design studies to be analyzed with one-group-only-design data. Also, for studies considering both accuracy and fluency as outcome measures, two independent effect sizes were calculated. Since it has been documented that the computed effect size tends to be biased when based on small sample sizes, Hedge's correction was applied to Cohen's *d*. The output index is known as the unbiased standardized mean difference effect size or the weighted mean effect size (Hedges, Shymansky, & Woodworth, 1989; Lipsey & Wilson, 2001; Yun, 2011).

Based on Cohen's explanation of effect size, the effect size around 0.8 or above is considered a large effect, around 0.5 a medium effect, and less than 0.2 a small effect. To determine the significance of the mean effect sizes, 95% confidence intervals were calculated for each weighted mean effect size. The confidence intervals which do not contain the value zero are interpreted as statistically significant. The narrower the confidence interval, the more reliable the obtained effect. Therefore, a positive effect size, with a 95 percent confidence interval not containing zero, is an indication that the implemented treatment was significantly effective on the development of accuracy and/or fluency.

Results

In this section, a summary of the findings is presented. The analyses focused on three outcome variables commonly reported by studies that investigate the effects of different types of treatments on students' oral accuracy and/fluency. In the following, the findings pertinent to each of the variables are presented separately.

Characteristics of the Studies

Based on the identified 27 study features, 14 articles were coded (detailed results are presented in Table 2). Being published between 2006 and 2016, the 14 studies were all journal articles. Sample sizes ranged from 20 to 120, all with adult learners including both males and females (890 in total). Eight studies were conducted with students with Persian as their mother tongue or which did not report the L1 of the participants, and the rest were conducted with students with other local languages (namely Azeri and Kurdish) or a combination of L1s. In two studies, the participants were chosen from the State- run universities, in 10 studies the samples were taken from the Islamic Azad University, in one study sampling was done from a non-governmental university and finally in the last one the sample was chosen from a combination of different universities.

Overall, two target language features were examined: general oral proficiency in 12 studies and syntactic accuracy in two studies. As mentioned earlier, six studies examined both oral accuracy and oral fluency, six studies probed into only oral fluency, and only two examined oral accuracy. Of the fourteen studies, three studies included more than one comparison (namely, studies with IDs 2, 5 and 10). A summary of the study characteristics of the primary studies are presented in Table 3.

Table 3
Summary of the Primary Studies

Study ID	N	L1	Outcome measures	L2 proficiency	Treatment	Treatment duration	Treatment activity
1	37	Azeri	Accuracy/Fluency	Not reported	task planning time	0	Watching movies
2	120	Azeri	Accuracy/Fluency	Mid	task repetition & task type	1	not applicable
3	40	Persian & Azeri	Accuracy/Fluency	Mid	task planning time	1	not applicable
4	33	Persian	Accuracy/Fluency	Low	photomontage	0	topic discussion
5	40	Persian	Accuracy/Fluency	Mid & High	task type	2	not applicable
6	60	Persian	Accuracy/Fluency	Low	podcast reconstruction	0	listening and retelling
7	52	Persian	Fluency	Mid	online chat	0	topic discussion
8	60	Kurdish	Fluency	Low	dialog journals	16	recording oral productions
9	20	Persian	Fluency	Mid	speaking strategic planning	0	picture-cued story telling
10	45	Mixed	Fluency	Low	interaction & rote learning	16	memorizing formulaic expressions
11	120	Persian	Fluency	Mid	podcast reconstruction	0	topic discussion
12	114	Azeri	Fluency	Mid	metacognitive awareness	0	picture-cued story telling
13	109	Persian	Accuracy	Mid	Feedback	12	topic discussion
14	40	Persian & Azeri	Accuracy	Mid	task planning time	1	not applicable

Among the studies, four studies involved low-level learners; eight studies, mid-level learners; and one study, mixed-level learners. The Mixed-level was a

group including both mid- and high-level learners. One study did not report learners' target language proficiency level (Rouhi, 2006). On this occasion, following Norris and Ortega (2000), the amount of L2 instruction that the learners received prior to the investigation as a proxy for L2 proficiency was taken into account. However, since the study did not report the grade level of the learners either, the required data were missing and the associated item was labeled as *Not reported*.

In terms of the interventions, the studies implemented a diverse spectrum of treatments which is an indication of the wide range of efforts and ELT practices having being experimented: task planning time, task repetition, task type, podcast reconstruction, online chat, dialog journals, speaking strategic planning, interaction, rote learning, metacognitive awareness, feedback, and the implementation of photomontage. Of the studies, three administered the instructional activities in one week, one in 2 weeks, one in 12 weeks and two in 16 weeks, while seven studies failed to report the duration of the experiment. Two studies conducted the experiments that included 20 sessions of treatment, one study 15 sessions, three studies 10 sessions, four studies only one session, and four studies failed to report the overall treatment sessions.

In order to gain a better scope of the features of class/home activities used in the treatments, what learners were asked to do in the different studies was also extracted. Four studies asked learners to have discussions on various topics, while two wanted that learners tell stories based on picture-cued tasks. Watching movies, listening and retelling, recording oral productions and memorizing formulaic expressions, were other class/home activities each of which was administered in one study. The nature of the treatments in four studies did not require that learners do any type of activity.

The Overall Impressibility of Oral Accuracy and Fluency

To answer the first research question about the overall impressibility of oral accuracy and fluency as a result of students' being exposed to study treatments, the unbiased standardized mean difference effect sizes (i.e. weighted mean effect sizes) were calculated based on the sample size of each study, as shown in Table 4 and Table 5 respectively. In total, there were 32 comparison groups

allowing the calculation of effect sizes across the primary studies, 13 of which yielding descriptions of accuracy and 19 describing fluency.

Table 4
Weighted Mean Effect Sizes on Accuracy

Study ID	**Label (multiple comparisons)**	**Mean (experimental/ comparison)**	**SD (experimental/ comparison)**	**unbiased standardized mean difference effect size (g)**	**95% confidence interval (lower/upper)**
1	-	0.46/0.28	22.98/16.39	0.009	-0.6443/0.6623
2	A	0.6129/0.5223	0.48360/0.35116	0.2144	-0.0394/0.4681
	B	0.5890/0.3393	0.3441/0.2685	0.8091	0.1644/1.4537
	C	0.5890/0.6388	0.3441/0.3646	-0.1405	-0.761/0.4801
	D	0.3393/0.6388	0.2685/0.3646	-0.9354	-1.2019/-0.6689
3	-	0.6320/0.6715	0.23681/0.31379	-0.1421	-0.7627/0.4785
4	-	1.7834/1.3095	0.19646/0.26104	2.0514	1.4553/2.6474
5	A	0.31/0.15	0.084/0.072	2.0452	1.5044/2.5861
	B	0.31/0.29	0.084/0.089	0.2311	-0.2086/0.6708
	C	0.15/0.29	0.072/0.089	-1.7295	-2.2432/-1.2158
6	-	22.32/17.96	3.134/3.169	1.3834	0.8201/1.9468
13	-	0.1473/0.411	0.12405/0.18253	-1.6869	-2.1241/-1.2497
14	-	0.63/0.67	0.23/0.31	-0.1465	-0.7672/0.4741

Table 5
Weighted Mean Effect Sizes on Fluency

Study ID	**Label (multiple comparisons)**	**Mean (experimental/ comparison)**	**SD (experimental/ comparison)**	**unbiased standardized mean difference effect size (g)**	**95% confidence interval (lower/upper)**
1	-	0.85/2.56	1.79/4.52	-0.4974	-1.1608/0.1659
2	A	0.5868/0.4277	0.64384/0.44268	0.288	0.0336/0.5423
	B	0.4208/0.2913	0.4939/0.2893	0.32	0.0653/0.5746
	C	0.4208/0.5710	0.4939/0.4797	-0.3085	-0.563/-0.054
	D	0.2913/0.5710	0.2893/0.4797	-0.7061	-0.9669/-0.4453
3	-	0.3770/0.6955	0.45702/0.28339	-0.8376	-1.484/-0.1912
4	-	3313.3939/1112.2121	1819.79355/449.60363	1.6607	1.1011/2.2202
5	A	0.05/0.04	0.062/0.048	0.1804	-0.2588/0.6195
	B	0.05/0.08	0.062/0.076	-0.4326	-0.8759/0.0108
	C	0.04/0.08	0.048/0.076	-0.6293	-1.0783/-0.1803
6	-	22.32/17.96	3.134/3.169	1.3834	0.8201/1.9468
7	-	74.81/70.00	7.547/9.274	0.5647	0.0292/1.1002
8	-	3.025/2.375	0.696/0.739	0.9055	0.3742/1.4369
9	-	6.7/5.1	1.05/1.37	1.3109	0.3448/2.277

10	A	-82.18750/-80.53333	86.85637/102.22375	-0.0174	-0.7331/0.6983
	B	-82.18750/-91.25000	86.85637/130.54322	0.0817	-0.6342/0.7977
	C	-80.53333/-91.25000	102.22375/130.54322	0.0914	-0.6246/0.8075
11	-	14.1333/13.0667	1.72191/1.73564	0.617	0.2507/0.9832
12	-	0.36/0.60	0.25/0.66	-0.4842	-0.8567/-0.1116

Since some studies either compared more than one experimental group to a control/comparison group or were based on the factorial design so that they examined more than one independent variable, they generated more than a small effect size because of the calculations resulting from the different comparisons between the different groups. On such occasions, the different paired comparisons are labeled as a, b, c or d in Table 6 and Table 7.

Table 6
Subgroups Weighted Mean Effect Sizes on Oral Accuracy

Feature	**Variable**	**Number of comparisons**	**Weighted mean effect size (g)**	**95% CI (lower-upper)**
participants' L2 proficiency level	low	2	1.522	0.8201/2.6474
	mid	7	-0.2897	-2.1241/1.4537
	mixed	3	0.1823	-2.2432/2.5861
	NR	1	0.009	-0.6443/0.6623
participants' L1	Persian (or not reported)	6	0.38245	-2.2432/2.6474
	Azeri	5	-0.00868	-1.2019/1.4537
	Mixed	2	-0.1443	-0.7672/0.4785
participants' gender	female	3	0.182267	-2.2432/2.5861
	mixed	8	0.309483	-2.1241/2.6474
	NR	2	0.6962	-0.6443/1.9468
instrumentation	narration	1	0.009	-0.6443/0.6623
	monologic tasks	8	-0.14811	-2.2432/2.5861
	interview	3	0.125475	-2.2432/2.5861
	dialogic tasks	1	2.0514	1.4553/2.6474
study duration	Mid (between one month and one semester)	1	-1.6869	-2.1241/-1.2497
	Short (less than a month)	9	0.022867	-2.2432/2.5861
	NR	3	1.147933	-0.6443/2.6474
All studies		13	0.299562	-2.2432/2.6474

Note. NR=Not Reported, NA=Not Applicable

Table 7
Subgroups Weighted Mean Effect Sizes on Oral Fluency

Feature	**Variable**	**Number of comparisons**	**Weighted mean effect size**	**95% CI (lower-upper)**
participants' L2 proficiency level	low	6	0.4375	-0.7331/2.2202
	mid	9	0.084911	-1.484/2.277
	mixed	3	-0.29383	-1.0783/0.6195
	NR	1	-0.4974	-1.1608/0.1659
participants' L1	Persian (or not reported)	8	0.45455	-1.0783/2.277
	Azeri	6	-0.23137	-0.7061/0.5746
	Kurdish	1	0.9055	0.3742/1.4369
	Mixed	4	-0.17048	-1.484/0.8075
participants' gender	female	3	0.182267	-2.2432/2.5861
	mixed	13	0.069417	-1.484/2.277
	NR	3	0.597167	-1.1608/1.9468
instrumentation	narration	7	0.157429	-1.1608/2.277
	monologic tasks	8	-0.26571	-1.484/0.6195
	interview	3	0.9512	0.0292/1.9468
	dialogic tasks	1	1.6607	1.1011/2.2202
study duration	long (one semester or longer)	4	0.2653	-0.7331/1.4369
	short (less than a month)	8	-0.26571	-0.8376/0.6195
	NR	7	0.650729	-1.1608/2.277
All studies		19	-0.00761	-2.2432/2.5861

Note. NR=Not Reported, NA=Not Applicable

As for oral accuracy, the unbiased effect sizes range from -1.7295 to 2.0514, with seven positive (four large, no medium, three small) and six negative effect sizes; this demonstrates that the overall effect sizes obtained in these studies tend to be small (Table 4). The weighted mean effect size across the 8 studies having examined oral accuracy was m=0.1509 which is a small mean effect size; moreover, the overall confidence interval of -2.2432 to 2.6474 indicates that the result is not statistically significant since the confidence interval includes zero (Norris & Ortega, 2009). The analyses show that, as for oral

accuracy, the experimented ELT practices are just as effective as regular classroom instruction.

However, the range of the obtained effect sizes with a standard deviation of 1.2, indicates that some ELT practices tend to be more beneficial in comparison to regular classroom instruction, while others were not or in some cases even worse. Moreover, the analysis of the 95 percent confidence intervals reveals the fact that three out of the seven confidence intervals pertinent to the seven obtained positive effect sizes encompass zero. This shows that, overall, among the 13 comparisons on the effectiveness of the experimented treatments, only four yielded to significant outcomes (studies with IDs 2b, 4, 5a and 6). The study with ID 6, as will be discussed in the next section, was later excluded from the analysis, since it considered oral proficiency as a whole construct and did not report an independent score as for oral accuracy. This indicates that almost in 77% of the experimented treatments, students performed as well as students in regular programs with no significant improvement in oral accuracy.

The unbiased effect sizes of the 19 oral fluency measures range from -0.8376 to 1.3834, with eleven positive (four large, three medium, twelve small) and eight negative effect sizes; this demonstrates that, like the effect sizes obtained for oral accuracy, the overall effect sizes obtained from these studies tend to be small (see Table 5). The weighted mean effect size across the 12 studies having examined oral fluency was m=0.1837 (small); moreover, the overall confidence interval of -1.484 to 2.277 indicates that the result is not statistically significant since the confidence interval includes zero. Therefore, the analyses show that the experimented ELT practices are just as effective on oral fluency as regular classroom instruction.

The range of the extracted effect sizes with a standard deviation of 0.74, shows that some ELT practices appear to be more effective in contrast to regular classroom instruction. In line with this, the analysis of the 95 percent confidence intervals reveals the fact that three out of the eleven confidence intervals pertinent to the eleven obtained positive effect sizes encompass zero. This indicated that, overall, among the 19 comparison groups on the effectiveness of the experimented treatments, eight studies yielded to significant outcomes (Studies with IDs 2a, 2b, 4, 6, 7, 8, 9 and 11) among

which study with ID 6 was later excluded from the analysis, since it considered oral proficiency as a whole construct and did not report an independent score as for oral fluency. This indicates that almost in 63% of the experimented treatments, students performed as well as students in regular programs with no significant improvement in oral fluency.

Following the analyses of the effect sizes, the coded study features were tested to determine sources of significant effect size variations. Of the 27 study features, four (namely participants' gender, randomization, reported reliability and length of the sessions) could not be contrasted since not all the studies included the required data for coding the features. This means that some features in some studies were not compared due to the lack of data. Thus, analyzing the variance was meaningless for such features because of the missing data, resulting in cases in which a value of "Not reported" was coded.

Moreover, some variables were not included in the meta-analysis in the first place due to the absence of data in some studies, thus incomparable: the frequency of students' attendance in classes, the level of the preparation of the teachers, the amount of teachers' experiences in language education and control for the effects of probable parallel training.

Contextual Factors Influencing Oral Accuracy and Fluency

To answer the second research question, analyses were conducted with the coded study features being as independent variables. In line with Lin, Huang and Liou (2013), Plonsky (2011), Cavanaugh et al. (2004), Goldberg et al. (2003) and Norris and Ortega (2000), the data associated with the following variables were extracted and synthesized as they were hypothesized to have been effective on the overall oral proficiency of L2 learners in terms of accuracy and fluency: participants' L2 proficiency level, participants' L1, participants' gender, instrumentations, study duration. and quality assessment. The unbiased standardized mean difference effect sizes and 95% confidence intervals (CI) for each study and its subgroups which represented the aforementioned contextual factors were calculated (see Table 6 and Table 7 for oral accuracy and fluency respectively). The findings aid in figuring out which contextual factors might have been influential in learners' L2 oral accuracy and fluency.

To begin with, the weighted mean effect size (g) of each L2 proficiency level associated with the oral accuracy of the comparison group was calculated. The results revealed the fact that the studies were highly influential in promoting students' oral accuracy at low-levels (g=1.522) since The result were statistically significant as the 95% CI of the two effect sizes did not include zero. In mixed-levels, the effect size (g=0.1823) was small and not significant, as the 95% CI included zero, while mid-levels did not experience an overall improvement in oral accuracy since weighted mean effect size was negative.

To account for the variation in the effects of the participants' mother tongue on the students' oral accuracy, the weighted mean effect sizes for the 13 subgroups were calculated. The results show that the students with Persian as their mother tongue experienced a rather small positive effect size (g=0.3824). This result was not found to be statistically significant at 95% confidence interval. The findings also showed that the results in the experiments with groups including participants with Azeri as their L1 yielded to neutral effects while mixed groups in terms of L1 experienced a negative effect size at 95% confidence interval. It deserves notice that since Persian is the official language in Iran, the groups which did not report the students' mother tongue were categorized as Persian by default.

To address the degree of the improvement of oral accuracy with the two genders, the weighted mean effect sizes were calculated. The treatments with females were found to experience a small positive effect size, though the results with none of the subgroups were statistically significant. A simple look at the studies shows that females and mixed groups were the target of researchers' experiments while single-sex male participants have not been under investigation.

In order to observe the impacts of the different types of instrumentation on oral accuracy, the weighted mean effect sizes of the studies geared to each and every instrument were calculated. It was found that interviews and dialogic tasks in other forms were the most effective ones (g=0.9512 and g=1.6607 respectively) and the result was statistically significant based on the 95% confidence intervals. As for narration, the analysis showed a small weighted mean effect size (g=0.157429) with a statistically insignificant outcome, while

other monologic tasks (a combination of personal, narrative and decision-making tasks) were found to have negative effects on the students' performance. It should be noted here that there was only one effect size contributing to the weighted mean effect size for the dialogic task subgroup; Therefore, this particular result should be interpreted with caution.

To investigate the extent to which the duration of treatments influenced the development of the students' oral accuracy, the effect sizes were computed across the three pre-defined categories of treatment periods: short, mid, long. It was found that the studies had nearly a neutral effect when the treatment period was short (g=0.022867) and a negative effect when the treatment duration was defined as mid (g=-1.6869); none of these findings was statistically significant at 95% confidence interval, though. Also, since no study investigated the impact of the treatments longitudinally, no mean effect size was extracted. Among the 13 comparison pairs, nine did not report the duration of treatments, thus the related effect sizes obtained from the studies were labelled as NA (i.e., not applicable), while the mean effect size (g=1.147933) was not statistically significant.

As for fluency, the analysis of the weighted mean effect sizes showed a medium positive improvement with low-level participants (g=0.4375, 95% CI) while mid-levels and mixed-levels experienced an insignificant neutral and negative development respectively based on obtained scores on their oral performance.

To investigate the extent to which the variation in the effects of the participants' mother tongue influenced the development of the students' oral fluency, the effect sizes were calculated across the 19 subgroups. The results show that the students with Persian as their mother tongue experienced a rather medium positive effect size (g=0.45455), the results of which was found to be not statistically significant (95% CI). This result was not found to be statistically significant at 95% confidence interval. Similar to accuracy, the mother tongue of the students in studies on fluency which did not report students' L1 was labeled as Persian. The findings also showed that results in experiments with groups including participants with Azeri as well as mixed-L1 groups yielded negative effects. Only the Kurdish subgroup experienced a high

effect size (g=0.9055, 95% CI), a result which needs further investigation since it is based on only one single study.

The calculated weighted mean effect sizes also show that the degree of the improvement of oral accuracy with females was positive (g=0.182267) while the treatment on mixed-gender groups were ineffective. Both results were found to be statistically insignificant at 95% confidence interval.

In terms of the effects of instruments, interviews with a high weighted mean effect size (g=0.9512) at 95% confidence interval were influential in improving the students' oral fluency, while other instrument types either were ineffective (monologic tasks) or had a very small share of impact on oral fluency (narration).

To address the degree of the effectiveness of the duration of the study, the weighted mean effect sizes revealed the fact that long-term treatments had a small positive and short-term treatments a negative impact on the students' oral fluency, neither results were statistically significant at 95% confidence interval.

Quality Assessment of the Included Papers

While meta-analyses provide the most highly rated recommendations for evidence-based treatment (Kung et al., 2010), interpretations on the part of the findings of the primary studies should be treated with caution. A number of instruments have been created and developed in order to assess the quality of meta-analyses as well as primary research. In this regard, an 8-item assessment rubric was developed by reviewing available instruments in literature and creating a list of components.

A rater manual, whose purpose was to describe items in the tool and to help raters to score study quality, accompanied the tool. It deserves consideration that not all the included items applied to every primary study. Thus, in case there was any ambiguity in a study in terms of the report of a feature which made scoring a challenge, or where the presence or absence of a study feature could not be located, the code "not reported", "not applicable" or "can't tell" was employed.

After researchers individually rated each primary study, discrepancies were discussed and resolved, and the final decision was made based upon the rating codes: strong, moderate and weak, where "strong" was given to papers with no

weak rating on the individual items of the tool, "moderate" to papers with only one weak rating and "weak" to papers with two or more weak ratings on the individual items of the tool. The results showed that, among the included papers, six studies were scored strong, three moderate and five weak. A comparison of the quality rank of the included studies with their associated overall effect sizes at 95% interval indicates that while studies 2a, 4, 5a and 6 showed a high positive impact of the implemented treatments on oral accuracy, only study 4 (i.e., Hassaskhah & Rahimizadeh Asli, 2015) was assessed as a high-quality experiment. As for oral fluency, among the eight comparisons with high positive effect sizes at 95% confidence interval (studies 2a, 2b, 4, 6, 7, 8, 9, 11), study 4 (Hassaskhah & Rahimizadeh Asli, 2015) showed to highly comply with quality standards and studies 8 and 9 were assessed to encompass a medium share of quality. Details regarding the quality scoring and assessment are found in the Appendix.

Discussion

This study employed meta-analytic procedures to synthesize findings across multiple studies to investigate the effects of potential factors on students' oral accuracy and fluency. A large number of studies which were primarily identified for inclusion in the analysis had to be eliminated either because they were qualitative, they focused on subjects other than the desirable group of participants (namely university EFL learners), or they failed to report statistics needed for effect size calculation. All in all, the analyses indicate that the overall effect of the different instruction types employed in previous studies has only had a small eeffect on the students' oral performance. In other words, the overall effect of the employed treatments in previous research has led to improvements in students' performance just similar to the outcomes of regular classroom instruction. This impact is found in the both phases of the study, namely on oral accuracy as well as fluency.

However, among the independent variables having been investigated, three different types of treatment were identified to be significantly effective in promoting learners' oral accuracy: the employment of narrative tasks, photomontage and interviews. In contrast to accuracy, a larger number of treatments were found to be influential in developing learners' oral fluency. An

analysis of the obtained effect sizes in the extracted subgroups indicates that employing photomontage, keeping dialogue journals and strategic planning are highly effective, while online chatting, task repetition and the application of narrative tasks in class instruction rank next in enhancing learners' oral fluency.

In interpreting the results of previous studies, some points deserve notice. In the study conducted by Asaadinezhad and Gorjian (2015), the researchers took a holistic approach toward having an operational definition of the construct of speaking proficiency in such a way that the data related to accuracy and fluency were not included in the paper. Therefore, discrete scores for oral accuracy and fluency could not be extracted and the outcomes of the study had to be excluded from the meta-analysis. Moreover, as to the absence of a control group in Moradi and Talebi's (2014) study, that is, the one-group pretest-posttest design, the related scores of the pre- and the post-test of the experimented group were compared and the effect sizes were extracted.

A closer look at the synthesized data reveals the fact that almost every primary study investigated the effects of one specific independent variable on speaking accuracy and fluency. This ends in interpretations of some divergent results which are based on experiments with multiple types of treatments, thus low reliable justifications. Therefore, the results inevitably show variations in the degree of improvement that students have experienced, as well as a need for more convergent information if firm conclusions are to be drawn. Therefore, outcomes of such investigations are advised to be treated with caution until future researches are done.

Moreover, a number of research studies employed the proposed interventions in one single session (namely Birjandi & Ahangari, 2008; Abdi, Eslami & Zahedi, 2012; Safari Vesal, Safari Vesal & Tavakoli, 2015; Ahangari & Abdi, 2011). The question here is whether the application of only one session of a specific treatment would result in valid interpretations, or the conclusions which are based on single-shot observations could as well be by-products of other confounding factors such as test-retest effect, practice effect, etc. Thus, the problem whether single-shot interventions could be as effective a treatment to the oral performance of ELT undergraduates as regular classroom instruction calls for further research.

The complex nature of SLA and the many contextual variables involved in the employment of various interventions might be, in most cases, the main reason resulting in the small overall effect of the treatments. Despite the fact that our analyses of the contextual variables show that variations in L2 proficiency, participants' L1, participants' gender, instrumentation and study duration are likely to contribute to changes in learners' oral performance, due to the small sample sizes of some subgroups of factors, it is difficult to conclude which (and to what extent) contextual variations play major roles in promoting oral performance. Also, as it is clear, since none of the 14 studies included in the meta-analysis involved advanced learners (i.e. senior undergraduates), primary investigations that could target at advanced learners are recommended.

The present meta-analysis has also unveiled other limitations on the part of previous studies. First, because of the very limited number of published and unpublished master's and doctoral dissertations on the development of oral performance in general, and the promotion of oral accuracy and fluency as probably the most favorable consequences of speaking classes in particular, a certain inevitable level of publication bias existed in this study. Thus, more comprehensive studies, particularly empirical researches in the form of dissertations, should be conducted to examine the impressibility of students' oral performance. Second, the absence of meta-analytical studies in the field calls for analyses and syntheses of findings extracted from early researches.

More detailed descriptions of the employed treatments including the rationale in designing tasks or the implementation procedures would also aid to open new directions in identifying more precisely the variables that have a definitive or major influence on oral performance in the Iranian EFL context. In addition, as mentioned in earlier sections, some variables were not included in the meta-analysis due to the absence of the data in some studies. Among the many intervening and extraneous variables, further research could also probe into the potential effects of the frequency of students' attendance in classes, teachers' experiences in teaching, the level of the preparation of teachers (e.g., the number and duration of in-service programs), parallel training courses and so forth.

Another thought-provoking problem in previous studies is the absence of the examination of learning retention. Our reviews show that no study has

involved a delayed post-test; thus, the longitudinal effects of the experimented treatments on oral performance could not be verified in this meta-analysis. Future researches are encouraged to incorporate designs which could assess beyond an immediate post-treatment observation to obtain a long-term picture of the effectiveness of the implemented instruction. Besides, since nearly all synthesized researches in the present study took a holistic approach toward assessing oral proficiency in terms of target language structure, conducting analytical studies that focus on different language forms is also recommended.

Last but not least is the very few number of studies having investigated the effects of independent variables on oral performance that reported a simultaneous improvement in oral accuracy as well as fluency. Birjandi and Ahangari (2008) justified employing narrative tasks (in the form of story-telling tasks) in speaking classes. Their findings indicated that narrative tasks would lead to a simultaneous improvement in oral accuracy and fluency. In another study, Hassakhah and Rahimizadeh Asli (2015) introduced photomontage as a speaking task to facilitate talking in EFL classes. Addressing Skehan's (2009) explanation of the tensions which exist between form (accuracy) and fluency, they suggested photomontage tasks as a resolution to the problem of facilitating oral accuracy and fluency at the same time. None of the studies, however, reported the number of treatment sessions as well as the length of the study. In addition, the assessment of learning retention, that is, delayed post-test, was not documented in either of the studies. Besides, both studies focused on students' holistic oral proficiency and neither targeted at any specific form.

The present study is significant in that it adds to the body of literature on the effectiveness of different types of treatment in instructed second language learning. The fact that, to the best of our knowledge, no meta-analytical studies have ever been conducted to depict the overall status quo of oral performance among ELT undergraduates in the Iranian universities' English departments augments the significance of the present study. All in all, while the problems of the impressibility of oral performance among university ELT undergraduates and of the factors influencing the effectiveness of employed classroom treatments were addressed, more research is called upon to further investigations.

References

Note: Studies included in the meta-analysis are marked with an asterisk (*).

* Abdi, M., Eslami, H., & Zahedi, Y. (2012). The impact of pre-task planning on the fluency and accuracy of Iranian EFL learners' oral performance. *Procedia - Social and Behavioral Sciences, 69*, 2281 – 2288.

* Ahangari, S., & Abdi, M. (2011). The effect of pre-task planning on the accuracy and complexity of Iranian EFL learners' oral performance. *Procedia - Social and Behavioral Sciences, 29*, 1950 – 1959.

* Akef, K., & Nossratpour, S. (2010). The impact of keeping oral dialogue journals on EFL learners' oral fluency. *Journal of English Language Studies, 1*(2), 127-142.

Ansarian, A. A., & Chehrazad, M. H. (2015). Differential effects of focused and unfocused recasts on the EFL learners' oral accuracy. *Colombia Applied Linguistics Journal, 17*(1), 86-97.

* Asaadinezhad, N., & Gorjian , B. (2015). The effect of reconstruction podcast on pre-intermediate EFL learners' speaking proficiency. *International Journal of Language Learning and Applied Linguistics World, 8*(3), 132-145.

Askari, K., & Langroudi, J. (2014). The effectiveness of Ur model in developing Iranian EFL learners' fluency and accuracy in speaking. *Applied Linguistic and Language Research, 1*(1), 75-86.

* Baradaran, A., & Khalili, A. (2009). The impact of online chatting on EFL learners' oral fluency. *Journal of English Language Studies, 1*(1), 63-77.

Binder, C. (1988). Precision teaching: Measuring and attaining exemplary academic achievement. *Youth Policy, 10*(7), 12-15.

Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst, 19*(2), 163-197.

Binder, C., Haughton, E., & Bateman, B. (2002). *Fluency: Achieving true mastery in the learning process.* Professional Papers in Special Education, Charlottesville, VA: University of Virginia. Retrieved May, 2016 from http://special.edschool.virginia.edu/resources/papers.html/Binder-et-al_Fluency.pdf

* Birjandi, P., & Ahangari, S. (2008). Effects of task repetition on the fluency, complexity and accuracy of Iranian EFL learners' oral discourse. *The Asian EFL Journal*, *10*(3), 28-52.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 23-48). Harlow: Longman.

Cavanaugh, C., Gillan, K. J., Kromrey, J., Hess, M., & Blomeyer, R. (2004). *The effects of distance education on K–12 student outcomes: A meta-analysis.* Naperville, IL: Learning Point Associates.

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535-544.

Chomsky, N. (1965) *Aspects of the theory of syntax*. MIT Press: Cambridge.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (1st ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, *11*, 367-383.

Ebsworth, M. E. (1998). Accuracy & fluency: Which comes first in ESL instruction? *ESL Magazine*, *1*(2), 24-26.

Effective Public Health Practice Project. (1998). *Quality assessment tool for quantitative studies.* Retrieved May, 2016 from http://www.ephpp.ca/index.html

Ellis, R. (1994). *The Study of second language acquisition*. Oxford: Oxford University Press.

* Farrokhi, F., & Mahmoudi, A. (2014). A socio-cognitive approach to developing oral fluency and naturalness in Iranian EFL learners. *International Journal of Applied Linguistics & English Literature*, *3*(2), 1-15.

Ferris, D. (2004). The "grammar correction" debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime?). *Journal of Second Language Writing*, *13*(1), 49-62.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*, 299-323.

* Ghenaati, M. J., & Madani, D. (2015). The effect of exposure to TV and radio news on the improvement of Iranian EFL learners' speaking fluency. *Research Journal of English Language and Literature*, *3*(4), 398-411.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A metaanalysis of studies from 1992 to 2002. *The Journal of Technology, Learning, and Assessment, 2*(1). Retrieved May, 2016, from http://www.bc.edu/research/intasc/jtla/journal/pdf/v2n1_jtla.pdf

* Hassaskhah, J., & Rahimizadeh Asli, S. (2015). Photomontage: A new task to change speaking into talking classrooms. *Cogent Education, 2*, 1-11. Retrieved May, 2016 from https://www.cogentoa.com/article/10.1080/2331186X.2015.1125333.pdf

Hazrativand, P. (2012). The Effect of Typographical Input Enhancement on Iranian EFL Learners'Accuracy in Oral Production of Narratives. *International Journal of Applied Linguistics and English Literature, 1*(4), 76-85

Hedges, L.V., & Olkin, I. (1985) Statistical Methods for Meta-Analysis. Orlando: Academic Press.

Hedges, L. V., Shymansky, J., & Woodworth, G. (1989). *Practical guide to modern methods of meta-analysis.* Arlington, VA: National Science Teachers Association Press.

Horowitz, D. (1986). Process not product: Less than meets the eye. *TESOL Quarterly, 20*(1), 141144

In'nami, Y., & Koizumi, R. (2010). Can structural equation models in second language testing and learning research be successfully replicated? *International Journal of Testing, 10*, 262-273.

Iwashita, N., Elder. C., & McNamara, T. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(3), 401– 436.

Johnson, K., & Johnson, H. (1999). *Encyclopedic dictionary of applied linguistics.* Oxford: Blackwell Publishing.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*, 145–164.

Kumaravadivelu, B. (2006). *Understanding language teaching: From method to postmethod*. NY: Routledge.

Kung, J., Chiappelli, F., Cajulis, O. O., Avezova, R., Kossan, G., & Chew, L. (2010). From systematic reviews to clinical recommendations for evidence-based health care: validation of revised assessment of multiple systematic reviews (R-AMSTAR) for grading of clinical relevance. *The Open Dentistry Journal, 4*, 84–91.

Lambert, C.P., & Engler, S. (2007). Information distribution and goal orientation in second language task design. In M. P. G. Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 25-43). Clevedon: Multilingual Matters.

Larsen-Freeman, D. (2006). Functional grammar: On the value and limitations of dependability, inference, and generalizability. In M. Chalhoub-Deville, C. Chapelle, & P. Duff (Eds), *Inference and generalizability in applied linguistics.* Amsterdam: Benjamins.

Lin, W., Huang, H., & Liou, H. (2013). The effects of text-based SCMC on SLA: A Meta-Analysis. *Language Learning & Technology*, *17*(2), 123–142.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

* Maftoon, P., & Kolahi, S. (2009). The impact of recasts on the syntactic accuracy of Iranian EFL university students' oral discourse. *The Journal of Applied Linguistics, (*2)2, 160-178.

Masgoret, A. M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning, 53,* 123-163.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20,* 83–108.

* Moradi, Z., & Talebi, S. H. (2014). The effect of pre-speaking strategies instruction in strategic planning on Iranian EFL students' awareness as well as students' fluency and lexical resources in speaking. *Procedia - Social and Behavioral Sciences, 98,* 1224 – 1231.

Nation, I. S. P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking.* NY: Routledge.

Nayar, P. B. (1997). ESL/EFL dichotomy today: Language politics or pragmatics? *TESOL Quarterly, 31*(1), 9–37.

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

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30,* 555-578.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21,* 109- 148.

Oxford, R. L. & Burry-Stock, J. A. (1995). Assessing the use of language learning strategies worldwide with the ESL/EFL version of the Strategy Inventory for Language Learning (SILL). *System, 23*(1), 1-23.

Pienemann, M., & KeBler, J. (2011). *Studying processability theory: An introductory textbook.* US: John Benjamins Publishing.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning, 61*(4), 993-1038.

Pollock C. W. (1997). *Communicate what you mean: A concise advanced grammar* (2nd ed.). NY: Prentice Hall.

Rafie, Z. F., Rahmany, R., & Sadeqi, B. (2015). The differential effects of three types of task planning on the accuracy of L2 oral production, *Journal of Language Teaching and Research, 6*(6), 1297-1304.

Richards, J. C., & Schmidt, R. (2010). *Longman dictionary of language teaching and applied linguistics* (4th ed.). London: Pearson Longman.

Rivers, W. M. (1981). *Teaching foreign language skills* (2nd ed.). Chicago: University of Chicago Press.

Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning, 45,* 283-331.

Robinson, P., & Ellis, N. C. (2008). Conclusions: Cognitive linguistics, second language acquisition and L2 instruction - Issues for research. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 489–545). NY: Routledge.

Rouhi, A. (2006). Striking an effective balance between accuracy and fluency in task-based teaching (Unpublished doctoral dissertation). Tehran University, Tehran.

* Rouhi, A., & Marefat. H. (2006). Planning time effect on fluency, complexity and accuracy of L2 output. *Pazhuhesh-e Zabanha-ye Khareji, 27,* 123-141.

* Safari Vesal, N., Safari Vesal, M., & Tavakoli, M. (2015). The effect of task type on complexity, accuracy, and fluency of Iranian EFL candidates' oral production: IELTS interview test in focus. *Proceedings of The 2nd National Applied Research Conference on English Language Studies, Tehran, Iran.* Retrieved May, 2016 from http://www.civilica.com/Paper-ELSCONF02-ELSCONF02_095.html

* Seifoori, Z. (2016). Metacognitive awareness and the fluency of task-based oral output across planning conditions: The case of Iranian TEFL students. *Iranian Journal of Language Teaching Research, 4*(1), 11-26

Semke, H. D. (1984). Effects of the red pen. *Foreign Language Annuals, 17,* 195-202.

Shiriyan, Z., & Nejadansari, D. (2014). The effect of literature-response activities on the complexity, accuracy, and fluency of Iranian EFL learners' L2 oral productions. *Journal of Applied Linguistics and Language Research, 1*(2), 12-26

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*, 185–211

Skehan P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(1), 93-120.

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers, *ELT Journal, 65*, 71-79.

Thomas, B. H., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing, 1*(3), 176-184.

Widdowson, H. G. (1997). EIL, ESL, EFL: Global issues and local interests. *World Englished, 16*(1), 135–146.

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: a meta-analysis. *Computer Assisted Language Learning, 24*(1), 39-58.

Appendix

Quality assessment tool for quantitative studies
COMPONENT RATINGS

A) SELECTION BIAS
(Q1) Are the individuals selected to participate in the study likely to be representative of the target population?
1 Very likely
2 Somewhat likely
3 Not likely
4 Can't tell
(Q2) What percentage of selected individuals agreed to participate?
1 80 - 100% agreement

2 60 - 79% agreement
3 less than 60% agreement
4 Not applicable
5 Can't tell
B) STUDY DESIGN
Indicate the study design
1 Randomized controlled trial
2 Controlled clinical trial
3 Cohort analytic (two group pre + post)
4 Case-control
5 Cohort (one group pre + post (before and after))
6 Interrupted time series
7 Other specify _____
8 Can't tell
Was the study described as randomized? If NO, go to Component C.
No Yes
If Yes, was the method of randomization described? (See dictionary)
No Yes
If Yes, was the method appropriate? (See dictionary)
No Yes
C) CONFOUNDERS
(Q1) Were there important differences between groups prior to the intervention?
1 Yes
2 No
3 Can't tell
The following are examples of confounders:
1 Race 2 Sex
3 Marital status/family
4 Age
5 SES (income or class)
6 Education
7 Health status
8 Pre-intervention score on outcome measure
(Q2) If yes, indicate the percentage of relevant confounders that were controlled (either in the design (e.g. stratification, matching) or analysis)?
1 80 - 100% (most)
2 60 - 79% (some)

3 Less than 60% (few or none)
4 Can't Tell
D) BLINDING
(Q1) Was (were) the outcome assessor(s) aware of the intervention or exposure status of participants?
1 Yes
2 No
3 Can't tell
(Q2) Were the study participants aware of the research question?
1 Yes
2 No
3 Can't tell
E) DATA COLLECTION METHODS
(Q1) Were data collection tools shown to be valid?
1 Yes
2 No
3 Can't tell
(Q2) Were data collection tools shown to be reliable?
1 Yes
2 No
3 Can't tell
F) WITHDRAWALS AND DROP-OUTS
(Q1) Were withdrawals and drop-outs reported in terms of numbers and/or reasons per group?
1 Yes
2 No
3 Can't tell
4 Not Applicable (i.e. one time surveys or interviews)
(Q2) Indicate the percentage of participants completing the study. (If the percentage differs by groups, record the lowest).
1 80 -100%
2 60 - 79%
3 less than 60%
4 Can't tell
5 Not Applicable (i.e. Retrospective case-control)
G) INTERVENTION INTEGRITY
(Q1) What percentage of participants received the allocated intervention or exposure of interest?

1 80 -100%
2 60 - 79%
3 less than 60%
4 Can't tell
(Q2) Was the consistency of the intervention measured?
1 Yes
2 No
3 Can't tell
(Q3) Is it likely that subjects received an unintended intervention (contamination or co-intervention) that may influence the results?
4 Yes
5 No
6 Can't tell
H) Analyses
(Q1) Indicate the unit of allocation (circle one)
community organization/institution practice/office individual
(Q2) Indicate the unit of analysis (circle one)
community organization/institution practice/office individual
(Q3) Are the statistical methods appropriate for the study design?
1 Yes
2 No
3 Can't tell
(Q4) Is the analysis performed by intervention allocation status (i.e. intention to treat) rather than the actual intervention received?
1 Yes
2 No
3 Can't tell
GLOBAL RATING FOR THIS PAPER (circle one):
1 STRONG (no WEAK ratings)
2 MODERATE (one WEAK rating)
3 WEAK (two or more WEAK ratings)
With both reviewers discussing the ratings:
Is there a discrepancy between the two reviewers with respect to the component (A-F) ratings?
No Yes
If yes, indicate the reason for the discrepancy
1 Oversight
2 Differences in interpretation of criteria
3 Differences in interpretation of study

Final decision of both reviewers (circle one):	1	STRONG
	2	MODERATE
	3	WEAK

Biodata

Arman Toni is a PhD candidate of applied linguistics and currently a lecturer at MSRT where he teaches SLA theories and psycholinguistics. His research interests include cognitive psychology in language acquisition, instructed second language learning and form-focused instruction.

Jaleh Hassaskhah is an associate professor of TEFL at the University of Guilan. Her areas of research include exploring the processes of teaching and learning and examining the environmental factors influencing these processes. She has published articles and books on the related issues both nationally and internationally.

Parviz Birjandi is a professor of research methods and statistics and the principle of the Department of English at IAU, Science and Research Branch, Tehran. He is well-known for his publication of articles in applied linguistics as well as the authorship of English language textbooks for Iranian high schools.