2023

A systematic review of proficiency assessment methods in bilingualism research

Daniel J. Olson
Purdue University, danielolson@purdue.edu

Recommended Citation

Olson, D. J. (2023). A systematic review of proficiency assessment methods in bilingualism research. *International Journal of Bilingualism,* 1–36. https://doi.org/10.1177/13670069231153720

A systematic review of proficiency assessment methods in bilingualism research "

Daniel J. Olson "

Purdue University
640 Oval Dr., West Lafayette, IN, USA, 47907
danielolson@purdue.edu

Abstract: "

Aims and Objectives/Purpose/Research Questions "

Proficiency assessment is a key methodological consideration in the field of bilingualism, and previous reviews have highlighted significant variability in both the use and type of assessment methods. Yet, previous reviews of proficiency assessment methods in bilingualism have failed to consider key study characteristics (e.g., methodology and subfield) that may impact the choice of proficiency assessment method. This paper provides an updated systematic review of proficiency assessment methods in the field of bilingualism, analyzing trends within different methodological approaches and linguistic subfields.

Design/Methodology/Approach

A systematic review was conducted, examining recent research articles in the field of bilingualism, broadly defined. A total of 17 journals (of 100) and 140 empirical research articles (of 478) with bilingual participants fit the relevant inclusionary criteria.

Data and Analysis

Studies were coded for several characteristics, including methodology (e.g., quantitative vs. qualitative), linguistic subfield (e.g., psycholinguistics), and the method of proficiency assessment (e.g., standardized testing, self-reporting).

Findings/Conclusions

Analyses revealed a number of different methods of proficiency assessment currently used in bilingualism research. However, different trends were found by methodology type and linguistic subfield. Broadly, the results revealed greater use of proficiency assessments in quantitative research than qualitative research. Moreover, while there was significant variability in all of the subfields examined, several within-subfield tends were identified.

Originality

This study provides an update to previous findings, establishing current proficiency assessment practices in bilingualism research. In addition, acknowledging the unique needs of different types of research, this study is the first to examine trends within different methodological approaches (i.e., quantitative vs. qualitative) and subfields of bilingualism (e.g., psycholinguistics, sociolinguistics).

Significance/Implications

The notable variability in proficiency assessment methods within different subfields suggests a greater need for subfield-specific norms to facilitate comparative analysis. Several key

considerations are given for the selection of proficiency assessment methods in bilingualism research.

Keywords: Language proficiency, Proficiency assessment, Language dominance, Systematic review, Second language acquisition

Introduction

Among the most long-standing methodological issues in the study of bilingualism has been who should be classified as bilingual and how to assess their abilities in their two languages. While the public conceptualization of a bilingual is likely akin to a *balanced bilingual*, who has equal "mastery" of each of their two languages (Wei, 2000, p. 6), the field of bilingualism has generally taken a broader approach. Montrul (2016), for example, defines a bilingual as someone who has knowledge of two languages, "although not necessarily to the same degree" (Montrul, 2016, p. 15). Similarly, Grosjean (2008) considers bilinguals to be those who use two or more languages or varieties in their everyday lives, "although rarely equally fluent in all language skills in all their languages" (p. 243). These definitions, adopted in the current study, are broad enough to encompass bilinguals who actively employ their languages in multiple contexts and those who are in the process of acquiring their second language (L2).[1] As such, most definitions of bilingualism rely explicitly or implicitly on the notion of proficiency, making proficiency a key methodological concept in bilingualism research. Yet, previous research has shown that many studies in the field fail to assess proficiency or fail to assess proficiency in a consistent manner across studies, limiting the ability to effectively compare outcomes across larger bodies of research (Gaillard & Tremblay, 2016; Grosjean, 2008; Tremblay, 2011). This systematic review provides an examination of recent proficiency assessment methods in bilingualism research and specifically accounts for different study characteristics (i.e., methodology and subfield), a factor not previously examined in large-scale reviews.

Defining Proficiency

While defining proficiency is seemingly intuitive, broadly characterized as a speaker's general abilities in a given language, researchers have highlighted the fact that the theoretical underpinnings of proficiency remain "underdeveloped" and "undertheorized" (Alderson, 2006, p. 12). Thomas (1994) provides a broad definition of proficiency as referring to a person's overall competence in a language. Early approaches to proficiency focused principally on evaluating a speaker's linguistic knowledge in a given language. This linguistic knowledge was evidenced in decontextualized language forms and translation exercises (for a review see Barnwell, 1996). Building on this line, other early models of proficiency proposed a two-dimensional conceptualization, with linguistic knowledge (e.g., lexical, morphological, syntactic) and the four language skills (i.e., listening, reading, speaking, and writing) as key components of proficiency (e.g., Carroll, 1972). Yet, in his review of this line of work, Hulstijn (2015) noted that these early approaches to proficiency failed to account for the fact that language is used in real-life communicative contexts.

Responding to this decontextualized approach, subsequent models of language proficiency began to rely on notions of communicative competence (e.g., Hymes, 1972), specifically addressing the degree to which a speaker is able to successfully engage in communicative situations in context. A communicative competence approach to proficiency highlights the multiple competencies required for successful interaction, including grammatical, sociocultural and structural competence (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980). The dual roles of both linguistic knowledge and communicative skills are evidenced in Hulstijn (2011), who provides a more comprehensive definition of proficiency as "the extent to which an individual possesses the linguistic cognition necessary to function in a given communicative situation" (p. 242). The core components of language proficiency include the "phonetic-phonological,

morphonological, morphosyntactic, and lexical domains" (p. 242), which are complimented by the (peripheral) strategies or skills such as interactional abilities, strategic competencies, and knowledge of different types of oral and written discourses. These core components are required for almost all communicative interaction, while the peripheral components are recruited selectively depending on the demands of the task.

Within this communicative framework, two complementary theoretical approaches should be noted. First, a number of researchers conceptualize of an overarching category of proficiency, entailing both the knowledge and skills components (e.g., Hyltenstam, 2016). This unified proficiency category suggests that *both* linguistic knowledge and communicative skills are needed to communicate effectively. In this framework, a speaker's proficiency in a given language is operationalized as a continuous variable and ranges from low proficiency to high/advanced proficiency, with nativeness or native-likeness as an endpoint or benchmark. In contrast, Hulstijn (2015) distinguishes between proficiencies in Basic Language Cognition (BLC) and Higher Language Cognition (HLC). BLC refers to the linguistic knowledge (i.e., phonetics, phonology, morphology, syntax) and automaticity with which this knowledge is used by all native speakers of a given language. HLC refers to the domain where differences are found in native speakers, such as low-frequency lexical and morphosyntactic structures. Referring to his core-periphery distinction, the BLC falls within the core components of language proficiency, while the HLC "pertains to both the core and periphery" (p. Hulstijn, 2015, p. 52). Within this framework, proficiency is not conceptualized as a single continuous variable, but as a "dichotomy" (Hulstijn, 2015, p. 22). For non-native speakers, this dichotomy breaks proficiency into the subcomponents, although these subcomponents may be evaluated as dual continuous variables. While an in-depth treatment of the theoretical nature of proficiency is beyond the scope of the current paper, most models acknowledge proficiency as some combination of knowledge and skills that facilitate successful comprehension and production of the target language (Gaillard & Tremblay, 2016; Ingram, 1985).[2]

Measuring Proficiency

In spite of the theoretical complexity of proficiency, researchers have long sought to provide a measure of proficiency for bilinguals, including second language learners (for a review see Ingram, 1985). Yet, given its inherent complexity, measuring proficiency is not particularly straight-forward. Responding to this complexity, Gaillard and Tremblay (2016) propose several key components of a well-designed proficiency assessment. Proficiency assessments should be valid and reliable, correlating well with other measures of proficiency and producing consistent results for similar populations. Assessments should also be global, assessing multiple linguistic domains beyond those that are directly tested in a given study. Considering more practical concerns, Gaillard and Tremblay (2016) suggest that researchers should choose a proficiency assessment sufficiently granular to identify different degrees of proficiency relevant for a given study, but succinct enough to be completed within existing time constraints. Moreover, the goals of a proficiency assessment may differ between theoretical and/or empirical research and educational contexts (Hulstijn, 2011). Taken as a whole, the appropriate proficiency assessment for a given research paradigm may depend on a number of factors.

Previous Methods of Proficiency Assessment

To date, there have been a wide variety of differing proficiency assessments used in bilingualism research. While a comprehensive list is likely impossible, some of the most common include standardized testing, self-rating, area specific proficiency tests, multi-component tests, institutional frameworks, oral proficiency interviews, and curricular standards.[3] These proficiency assessments are briefly detailed below, with some initial discussion of advantages and limitations.

Among the most common methods, a number of language-specific standardized tests have been developed for the assessment of proficiency, such as the Test of English as a Foreign Language (TOEFL, Educational Testing Service, 2020a) and the *Diploma de Español como Lengua Extranjera* (DELE, Universidad de Salamanca, n.d.). This category includes tests that largely rely on written modalities and addressing multiple linguistic subcomponents. Evidence is available for some standardized testing suggesting that they correlate with other indicators of proficiency (e.g., for TOEFL see Educational Testing Service, 2020b). As standardized tests may be time-consuming, others have proposed shorter or modified versions, such as the Modified DELE (Montrul & Slabakova, 2003). Among the advantages of standardized tests are their comprehensiveness, their correlation with other indicators of proficiency, and the fact that they are widely available. Yet, standardized tests are limited both theoretically, by their decontextualized nature (Thomas, 1994), and practically by the fact that they are often time-consuming (Tremblay, 2011) and not comparable across languages (Hulstijn, 2012).

Another method of proficiency assessment is direct self-rating (or self-assessment), which commonly uses quantitative questionnaires. Domain-general self-rating requires participants to rate their proficiency in a specific language, usually via Likert-scale rating (Marian et al., 2007). Domain-specific self-rating requires participants to rate their proficiency across a number of linguistic subskills (e.g., speaking, listening, reading, writing) (e.g., Bahrick et al., 1994). Relatedly, indirect self-assessments ask participants to rate their abilities in a variety of communicative tasks or situations (e.g., I can talk about what I do on the weekends) (Tigchelaar et al., 2017) and measure proficiency through a composite score of all communicative tasks (for a review see Ma & Winke, 2019). Quantitative self-ratings, which have been shown to correlate with performance on other proficiency test (e.g., Jia et al., 2002) are often easy to administer, comparable across multiple languages, and able to incorporate a variety of skills, and non-linguistic components (e.g., attitudes). Yet, self-assessments are inherently subjective measures, rather than objective measures, and may not offer a sufficient level of granularity for some studies.

Area specific proficiency tests have been used to examine proficiency in a single linguistic component or skill, but taken as representative of proficiency as a whole. For example, receptive vocabulary tests (e.g., Peabody Picture Vocabulary Test- Revised; Dunn & Dunn, 2007), available in a number of languages, have been used to assess participant proficiency. Syntactic and/or morphosyntactic complexity as a measure of proficiency can be assessed through analysis of written texts (e.g., Housen & Kuiken, 2009). Phonetic and phonological proficiency can be assessed via accent ratings (e.g., Flege et al., 2002) or acoustic measurements. Fluency has been assessed via sentence duration measures (e.g., Flege et al., 2002) or mean length utterance measures (e.g., Rice et al., 2010). Closely related, some authors have developed study-specific cloze tests to measure proficiency. Broadly, a cloze test consists of a text with a number of

lexical items deleted (Anderson, 1971). Participants are asked to provide the missing word, either spontaneously or from a list of options. Performance on some cloze tasks have been shown to correlate with performance on other proficiency assessments (e.g., Bachman, 1985). Several advantages of area specific proficiency tests include their availability in a variety of languages, their ease in administering and scoring, and that they may be tailored to a specific study need. However, these skill-specific proficiency measures generally assess one component while being taken to represent the broader concept of proficiency.

Beyond single-component tests, other approaches to proficiency assessment attempt to measure multiple components in a single evaluation. For example, Gaillard and Tremblay (2016) employed an elicited imitation task that entails assessment across multiple linguistic levels (e.g., vocabulary, syntax, phonetics). These scores may then be combined or averaged to provide an overall proficiency score. The elicited imitation task reflects L2 processing efficiency, which is taken as indicative of proficiency (Van Moere, 2012). A recent review provided evidence for a correlation between performance on an elicited imitation task and other measures of proficiency, with greater correlations found for longer imitation task lengths (Kostromitina & Plonsky, 2022). According to Gaillard and Tremblay (2016), assessing multiple linguistic components in a single assessment serves to add an additional layer of validity to the measure of proficiency.

Several methods provide a more holistic view of proficiency. For example, institutional frameworks, such as the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001), provide a holistic proficiency assessment for a number of languages. A speaker's CEFR proficiency level can be determined by several means, including via standardized testing and self-assessment through a language portfolio, allowing for cross-linguistic comparison. Also providing a more holistic view of proficiency, in which multiple linguistic components can be assessed in conjunction with communicative behaviors or skills, oral proficiency interviews (OPIs) have been widely used in second language acquisition. For example, the ACTFL OPI is a standardized oral production test that assesses language use "in real-world situations in a spontaneous and non-rehearsed context" (American Council of Teachers of Foreign Languages, 2012). Again, a number of studies have shown that the ACTFL OPI proficiency ratings are largely reliable (Surface et al., 2008) and correlate with other measures of proficiency (Tschirner et al., 2012). While the ACTFL OPI focuses on oral communication skills, separate tests are available for reading, writing, and listening, potentially providing learners with multiple proficiency ratings across the different skills. While both institutional frameworks and oral proficiency interviews are comprehensive in nature, address multiple skills, and are comparable across languages, they have been criticized as creating discrete labels for a continuous variable and the lack of consistency between different labeling systems (Menke & Malovrh, 2021). Both institutional frameworks and oral proficiency interviews are labor-intensive and may require specialized training on the part of the administrator.

Although not an independent test, curricular standards (i.e., classroom level or years of instruction) have also been found to be a common way to assess proficiency (e.g., Tremblay, 2011). In two separate systematic reviews of proficiency assessment in second language acquisition, institutional status was found to be the most common method of assessing or approximating proficiency (Thomas, 1994; Tremblay, 2011). Although this method provides a

quick and easy proficiency assessment, the different resulting groups or levels have been shown to be heterogeneous in their behavioral outcomes (Tremblay, 2011). Moreover, as different institutions have different curricular standards and reporting of such levels (e.g., intermediate-level) is often "accompanied by no criteria for interpretation" (Norris & Ortega, 2000, p. 455), comparison between studies and groups using curricular standards to assess proficiency may be limited.

Critiques of Proficiency Measurements

Menke and Malovrh (2021) review of a number of the key critiques of current approaches to proficiency measurement. From a theoretical perspective, among the most prominent critiques is the underdeveloped theoretical conceptualizations of proficiency that underlie proficiency assessments. As proficiency assessment has developed in tandem or ahead of theoretical conceptualizations of proficiency, some have argued that the conceptualization of proficiency is determined by existing proficiency assessment methods, rather than the other way around (e.g., Lantolf & Frawley, 1988), leading to questions about construct validity. Similarly, Menke and Malovrh (2021) highlight the fact that little empirical research has linked proficiency levels and assessments directly with well-documented developmental sequencies (Hyltenstam, 2016), although such research is now beginning to emerge (for a review see Menke & Malovrh, 2021). On a more practical level, while theoretical conceptualizations of proficiency highlight the role of the communicative context, several researchers note that proficiency assessments are largely absent of real-world discursive and sociolinguistic situations (McAloon, 2015). Previous research has shown that the speech produced during proficiency assessments is fundamentally different than speech produced in real-world contexts. Van Lier (1989), for example, notes several key differences between speech produced in OPIs, arguably one of the most naturalistic approaches, and in real conversations. These differences include the asymmetrical roles of the interviewer and interviewee, participants' varying levels of comfort with "showing off" their linguistic skills (p. 502), and sociopragmatic failures on the part of the interviewer impacting assessment of the interviewee (for a discussion of written assessments see Weigle & Friginal, 2015). The decontextualized nature of proficiency assessments, reminiscent of the early linguistic knowledge tests, brings in to question the ecological validity of proficiency testing. Moreover, proficiency assessments that provide discrete labels for different proficiency levels (e.g., ACTFL, CERF) have been criticized as arbitrarily dividing a continuous variable into discrete categories (Hyltenstam, 2016), which may not align across different proficiency assessments (Menke & Malovrh, 2021).

While much of the theoretical discussion takes place within the subset of literature on second language acquisition, in which theoretical conceptualizations and assessments of proficiency focus on a developing L2 (in contrast to a stable L1), additional considerations should be taken when applying the concept of proficiency to the broader category of bilinguals. First, as noted by Grosjean (2008), the notion that a bilingual is two monolinguals in one speaker is problematic. A bilingual's knowledge or skills may be uneven across different linguistic domains and skills. For example, bilingual heritage speakers may appear (near) native-like with respect to phonetics/phonology, but less so in the morphosyntactic domain (Polinsky & Kagan, 2007). Similarly, some communicative skills (e.g., reading and writing) may be underdeveloped relative to other skills (e.g., speaking and listening), responding to a bilingual's language experience (Grosjean, 2008). As most conceptualizations of proficiency entail multiple domains of linguistic

knowledge, it is unclear how such uneven development should be accounted for. Even within a single linguistic domain (e.g., lexical domain), a bilingual's knowledge may be context dependent. The complementary principle suggests that a bilingual's linguistic knowledge in each of their languages may be expected to differ. Bilinguals may use one language in work or technical settings and another language in social settings. Thus, bilinguals may develop unique, contextually-specific linguistic knowledge (e.g., vocabulary) in each of their languages. For example, Gasser (2000, as cited in Grosjean, 2008), found that a bilingual's vocabulary knowledge was topic-dependent (e.g., home topics = Language A vs. work topics = Language B).

A second consideration that is particularly relevant for assessing proficiency in bilingual speakers is the notion of the monolingual native speaker as the proficiency standard norm (see Menke & Malovrh, 2021). Again, the use of a monolingual benchmark suggests that bilinguals are (or may develop to be) two monolinguals in a single speaker. This conceptualization of proficiency relative to monolingual norms ignores the well-documented evidence that a bilingual's two languages inherently interact in complex ways.

Methodological Issues in Proficiency Assessment in Bilingualism Research

Within research on bilingualism, proficiency and proficiency assessments are employed with several different functions (see Luk & Bialystok, 2013), depending on the goals of the research. First, proficiency may be used as an inclusionary criterion for a given study, such that participants in the study must reach a certain level of proficiency in both of their languages to be eligible for the study. Within this function, studies may administer a proficiency assessment to ensure that a participant qualifies under the adopted definition of bilingualism (e.g., Chung, 2018). Second, proficiency may be used to establish between group comparisons within a single study. Among the most common approach is the establishment of a certain proficiency threshold to separate participants into monolingual and bilingual groups (e.g., Grey et al., 2018) or between novice and advanced language learners (e.g., Köylü, 2018). Third, proficiency may be used as a correlate of specific linguistic behavior or cognitive task. In this case, proficiency often serves as a variable of interest, and a bilingual's performance on a linguistic or cognitive task is analyzed with respect to their level of proficiency (e.g., Thomas-Sunesson et al., 2018). Finally, among the most important functions of proficiency for advancing collective knowledge in the field, proficiency may be used to facilitate cross-study comparison. This cross-study comparative function may be seen as overlaying the three other sub-functions, but necessarily relies on a degree of comparability of the proficiency measures used across the various studies.

Considering the role of proficiency assessment in cross-study comparison, Tremblay (2011) notes that "Although it may not be reasonable to expect all researchers to use the exact same proficiency assessment method, comparisons between studies… would be more amenable if there were at least some consensus on acceptable procedures that researchers could use to estimate proficiency and more consistency in the characterization of these procedures" (p. 343). This failure to produce consistent characterizations of participant proficiency may lead to an inability to reconcile seemingly divergent results. For example, Grosjean (2008) notes that specific methodological issues in the study of bilingualism, and most notably in how we describe a bilingual's abilities (e.g., language proficiency, language history, language use) in their two languages, has resulted in "conflicting results" (p. 241). He further details a series of inconsistent

results involving bilinguals in the fields of psycholinguistics and neurolinguistics that could be potentially clarified by attention to such methodological issues. These inconsistent results represent some fundamental issues in the field, including selective vs. non-selective language processing, presence vs. absence of a language switching cost, and language localization, among others (Grosjean, 2008). This sentiment has been echoed by a variety of researchers who all note that a lack of effective comparative measures serves as an obstacle to knowledge building (see Gaillard & Tremblay, 2016; Grosjean, 2008; Marian et al., 2007; Norris & Ortega, 2000; Tremblay, 2011).

Given the important role that proficiency plays in the field, and notably in the comparability of results across multiple studies, proficiency assessment remains a key methodological consideration. Grosjean (2008) highlights two main methodological issues related to proficiency assessment in the field of bilingualism: the failure to measure proficiency and the failure to measure proficiency in a consistent manner across studies. First, many studies fail to control for the variable of proficiency. Considering the presence (or absence) of proficiency assessment in prior research, the results of several prior systematic reviews of proficiency should be considered. Hulstijn's (2012) review of studies from a single journal in the field of bilingualism found that 20% of studies did not assess subjects' proficiency in any way. This result is paralleled in Tremblay's (2011) review of proficiency in second language acquisition (approx 9%). When proficiency was assessed, only a portion of studies used an objective measure (Hulstijn, 2012: 55%; Tremblay, 2011: 36%). Second, when proficiency is assessed, it is not operationalized in a consistent manner across studies, limiting the ability to compare outcomes (Gaillard & Tremblay, 2016; Tremblay, 2011). Previous systematic reviews have found a wide variety of different types of proficiency assessments (Hulstijn, 2012; Thomas, 1994; Tremblay, 2011). These methodological shortcomings of proficiency assessment have been noted in both the broader field of bilingualism (e.g., Grosjean, 2008; Marian et al., 2007) and in studies focused on the subset of bilinguals undergoing L2 acquisition (Tremblay, 2011) and have been documented through several large-scale systematic reviews (Hulstijn, 2012; Thomas, 1994; Tremblay, 2011). The overarching theme that emerges from these previous reviews is the wide variability in proficiency assessment methods used in the study of bilinguals and the limiting effect that such variability may have on the generalizability of results.

Finally, several authors have noted the important role that study characteristics (e.g., methodology and subfield) might play in selecting a proficiency assessment. Thomas (1994, p. 308–309), for example, states that "proficiency doesn't have a uniform function" in research (see also Tremblay, 2011). While previous reviews have highlighted the existing variability in proficiency assessment methods, the role of study characteristics, such as methodology or subfield, has not previously been accounted for when examining proficiency assessment methods in the field of bilingualism.

Research Questions

Given the role that proficiency assessment plays in creating a cohesive body of knowledge (e.g., Grosjean, 2008) and the perceived lack of consistency in methods of proficiency assessment in the field (e.g., Marian et al., 2007; Tremblay, 2011), the current project adds to ongoing methodological conversations in several important ways. First, this study examines proficiency assessment in the broader field of bilingualism, following current definitions of bilingualism

which include bilinguals across a variety of stages (i.e., learners and stable bilinguals) (Montrul, 2016). Second, this study provides an update to previous findings (Hulstijn, 2012; Thomas, 1994; Tremblay, 2011), offering an overview of recent proficiency assessment practices. And finally, acknowledging the unique needs of different types of research, this study is the first to examine trends within different methodological approaches (i.e., quantitative vs. qualitative) and subfields of bilingualism (e.g., psycholinguistics, sociolinguistics). Three research questions are addressed:

(RQ1) How is proficiency assessed in research in the field of bilingualism?
(RQ2) Does the type of proficiency assessment vary by study methodology?
(RQ3) Does the type of proficiency assessment vary by subfield?

Methodology

To address the above research questions, a systematic review of the literature was conducted. A systematic review is characterized by the use of systematic and explicit methods to identify, select, and critically appraise relevant research, which allows for the analysis and synthesis of results or trends in a given field (Higgins & Green, 2008). There are several advantages to a systematic review, including the ability to leverage a robust data set to draw broad, unbiased conclusions from the cumulative evidence on a given topic. The current systematic review differs from a traditional (narrative) systematic review in that it does not present an exhaustive analysis of all studies in which proficiency was assessed, but rather provides an analysis of a large, methodically-collected, representative, and recent sample of studies in the field of bilingualism.

Journal Eligibility Criteria

To conduct the review, it was necessary to systematically compile a large database of studies in bilingualism. The selection of the initial database differs somewhat from previous systematic reviews of proficiency assessment (Hulstijn, 2012; Thomas, 1994; Tremblay, 2011). In an effort to provide a more comprehensive picture of current proficiency assessment methods, the choice was made to include a greater number of journals over a shorter time period. The overall size of the current database is comparable to previous systematic reviews (Hulstijn, 2012; Thomas, 1994; Tremblay, 2011).[4]

The initial inclusionary criteria used to select the journals was based on notions of representativeness, recency, and scope. To provide a representative sample, while reasonably restricting the number of studies that could fall under the broad category of bilingualism, the top 100 (of 753) journals from the Scimago Journal Rankings (SJR) in the category of *Language and Linguistics* (Scimago, n.d.) were examined.[5] In order to focus on recent methods of proficiency assessment, only research articles published during 2018 were examined. At the onset of the project, 2018 was the latest, full year of publications available. Finally, considering scope, as the goal was to assess the common methods in the field of bilingualism, only journals that made specific reference to bilingualism in their publicly-available aims and scopes were included. To that end, the publisher descriptions for each of the initially-considered 100 journals were examined for the presence of several pre-determined keywords. Keywords that triggered the inclusion of a given journal included: *bilingual, bilingualism, multilingual, multilingualism, trilingual, trilingualism, code-switching, language switching, translanguaging,* and *interlingual.* A total of 17 journals were identified that used one (or more) of these keywords (Table 1).

[Table 1]

Article Eligibility Criteria

The 17 journals identified contained a total of 478 research articles. Only research-oriented studies were included in the database; editorials, book reviews, and commentaries were not considered. It was then necessary to set specific inclusionary criteria for each individual study. Only studies considering human participants that were broadly described as *bilinguals, multilinguals,* or *learners* were included. The decision was made to include all types of bilinguals, including learners, drawing on a broad definition of bilingualism (for discussion see Grosjean, 2008; Wei, 2000). Particular attention was paid to descriptions in the *Participants* or *Methods* subsections of each article. A significant number of variations of these terms were found, and the decision was made to err on the side of inclusion in any ambiguous descriptions. Each study was considered for inclusion by a single, trained research assistant. Of an initial set of 478 articles, 140 were included in the final data set (see Table 1), comparable to previous systematic reviews (Hulstijn, 2012; Thomas, 1994; Tremblay, 2011). Figure 1 illustrates the identification and screening process via a modified PRISMA flow-chart.

[Figure 1]

Given the wide variety of study designs and participant descriptions, the reliability of the application of the inclusionary criteria was assessed to ensure consistency. A randomly selected subset of approximately 13% (61 articles) of articles from the initial database was blindly recoded for inclusion by a second coder. Results showed broad agreement on the application of the inclusionary criteria, with 87% of the studies coded identically by the two raters. A chi-squared test showed that the two ratings were strongly correlated, $\chi^2(1) = 31.126$, $p < .001$, Cramer's V = 0.714.

Article Coding Procedure

Having identified the articles for inclusion, each article was coded for study characteristics, languages, and proficiency assessment. Within study characteristics, each article was coded for the type of study (i.e., quantitative ($n = 86$), qualitative ($n = 34$), mixed ($n = 20$)) and subfield of study (Table 2). Subfield coding was done qualitatively, using a cyclical approach from open coding to axial coding to final category assignment (e.g., Corbin & Strauss, 1990). Open coding refers to the initial process of identifying subfields for each article in the data set. When possible, open coding relied on the authors' own terminology and keywords to determine the subfield. Axial coding consisted of grouping subcategories into broader emergent themes and testing them against the data. During this stage, the decision was made to eliminate the categories of *second language acquisition* and *bilingualism*, as the majority of articles fit into one or both of these categories, effectively limiting their utility. Finally, final category assignment refers to the process of assigning each study a subfield tag or tags based on the final set of themes identified through the axial coding process. Each article was tagged for up to two subfields. For example, Mostafizar Rahman (2018), which addressed tense/lax vowel merger in the L2 English of L1 Bangala speakers and examined differences in gender, was coded as having a quantitative approach within the subfields of phonetics/phonology and sociolinguistics. A total of 225 subfield tags were assigned.

[Table 2]

Languages were coded using the terms in a given study, respecting the authors' choice of terminology. For example, some languages or varieties were described by different authors using different terms (e.g., Mandarin vs. Chinese), and some studies focused on specific language varieties (e.g., Namibian English). These terms were included as unique entries in the total count of languages. While the initial coding attempted to classify languages as L1 or L2, given the variety of study designs and participant profiles, this was not possible. As such, articles were coded for the languages used by participants, regardless of order of acquisition, resulting in a total of 380 language tags. Several studies (n = 7) were identified as including participants from 10 or more different languages or language pairings (e.g., Dewaele, 2018) and were coded separately. Three studies received tags of "unknown", as the languages of the participants were not specified. A total of 65 unique languages were identified in the data set (Table 3), excluding studies with participants from 10 or more different language backgrounds.

[Table 3]

Finally, and representing the main focus of the current systematic review, each study was coded for the presence (or absence) of a proficiency assessment and the type of proficiency assessment used. As some studies assessed proficiency in multiple ways, each study was permitted an unlimited number of proficiency assessment type tags. While a number of categories were established prior to the initial coding procedure, drawing on the previous literature (e.g., Thomas, 1994; Tremblay, 2011), coding was cyclical and several categories were added during the initial coding stage. As noted by Tremblay (2011) in her review, a precise classification of the proficiency assessment methods employed is often challenging, as many studies fail to report sufficient details. To provide a measure of reliability, each article was blindly coded by two trained coders. Any disagreements in the coding were subsequently discussed to reach agreement.

Results

Overall Use of Proficiency Assessment

Of the 140 articles included in this systematic review, proficiency was assessed in 118 (approximately 84%). The cyclical coding system identified nine separate methods for measuring proficiency in the data set: standardized testing, quantitative self-rating, other types of questionnaires, institutional frameworks, other proficiency tests, curricular standards, qualitative self-ratings, holistic assessments, and area specific proficiency. Of the studies that included proficiency assessment, thirty-eight (32%) were identified as employing more than one measure of proficiency. Table 4 illustrates overall frequency of the different proficiency assessment methods. Each of these assessment method categories is detailed below, with examples from the data set. It is worth noting that, in general, most studies did not explicitly provide rationale for their choice of proficiency assessment method.

[Table 4]

The most common method identified was standardized testing. Standardized tests were classified as large-scale tests of proficiency that are widely available, as well as their derivatives (e.g., the

Modified DELE). Standardized tests were largely found for single languages, particularly English. Examples of standardized tests in the data set include the TOEFL (e.g., Köylü, 2018), the modified DELE (e.g., Jegerski, 2018), the TOEIC (e.g., Nakayama et al., 2018), among others.

Two different questionnaire types were found in the data: quantitative self-ratings and other questionnaires. *Quantitative self-ratings* were defined as those methods in which participants were asked to provide a numerical evaluation, usually via a Likert-scale response, of their own proficiency. These quantitative self-ratings were provided either for a participant's proficiency in a language (e.g., Hopf et al., 2018) or their proficiency across a number of linguistic subskills (e.g., López & Vaid, 2018). Although many studies used Likert-type self-ratings, there was considerable variability in the self-rating structure. Unlike standardized testing, which generally was used to assess proficiency in one language, self-ratings were sometimes used to evaluate proficiency in two or more languages (e.g., Ellis et al., 2018). The category of *other questionnaires* was used to designate those questionnaires that addressed proficiency, albeit not through self-ratings or did not detail the mechanisms sufficiently to qualify as quantitative self-rating category. This designation was common in studies with child participants, as parents or caregivers responded to a questionnaire about the child's language use patterns or competencies. In some cases, formal questionnaires were used to directly assess children's proficiency (e.g., Child Multilingualism Questionnaire (CLALVCLA, 2012)), while in other cases, language exposure measures were used as a proxy for proficiency (e.g., Language Exposure Assessment Tool; DeAnda et al., 2016). This category was also common for teacher assessments of learners' abilities (e.g., Curdt-Christiansen & La Morgia, 2018). Broadly, the "other questionnaire" category resulted in quantitative results, although some used open-ended, qualitative questions (e.g., Paradowski & Bator, 2018).

The designation of *other proficiency test* was used to encompass any test of linguistic competence, used to determine proficiency, not included in the category of standardized test. Proficiency assessment methods in this category were most often author-developed methods of assessing proficiency. Notably, multiple studies used variations of the cloze-task as a measure of proficiency (e.g., Gánem-Gutiérrez & Gilmore, 2018).

Institutional frameworks consisted of large-scale descriptive tools to classify proficiency. Again, these frameworks serve as a reference for a participant's level of proficiency, but do not necessarily provide details about how to classify participants. The CEFR (Council of Europe, 2001) was the most common institutional framework in the data set, but evaluation methods were varied and not always specified. For example, reference levels were determined via participant self-report (e.g., Dahm & De Angelis, 2018), external assessors (e.g., Hijazo-Gascón, 2018), or curricular equivalents (e.g., Christiner et al., 2018).

Curricular standards, also called institutional status (e.g., Thomas, 1994), describes cases in which proficiency is discussed with respect to some curricular benchmark in an educational setting. Specifically, curricular standards refer to the (expected) proficiency levels on the basis of the courses in which a student was enrolled or the type of assignments given in an educational setting. Most commonly, curricular standards included the number of years or types of language courses a participant had taken (e.g., Suzuki & Sunada, 2018) or the results of a recent language

exam (Zhou & Zhou, 2018). Exemplifying the notion of expected curricular equivalence, Christiner et al., (2018) write, "after four years of learning, pupils should have reached the A1 level as described by the CEFR" (p. 459).

Both qualitative self-ratings and holistic assessment were largely based on qualitative reporting, but differed in who conducted the assessment. Q*ualitative self-ratings* relied on participant self-reporting, but did not use, or did not report, numerical outcomes. Qualitative self-reporting, more common in qualitative research, often consisted of a single mention of participant-asserted proficiency. For example, in addition to other measures of proficiency, Wilczewski et al., (2018) state that participants "declared that they had either good or proficient knowledge of English" (p. 592). *Holistic assessment* of participant proficiency was based on qualitative observations of a third party, usually the researcher or a classroom instructor (e.g., Seals, 2018). For example, Seals (2018) states that one participant "displayed proficiency in all three languages" (p. 333). In these cases, the assessment criteria was not always transparent. A second form of holistic assessment can be found in native speaker quantitative judgments of participant productions (e.g., Saito et al., 2018), which rely on native speakers' intuitions.

Area specific proficiency, among the least represented in the data set, referred to those studies that assessed proficiency in one particular linguistic domain. For example, Luk and Shirai (2018) used mean length utterance (MLU) to establish language proficiency (and dominance).

Among studies that did not use a proficiency assessment, there were several approximations to proficiency that should be noted. In several studies, authors described the general proficiency characteristics of the larger population, but not for their study participants (e.g., Karimzad & Sibgatullina, 2018). This was notably the case in which data was collected from participants in an anonymous manner. In other cases, studies reported exposure, rather than proficiency, particularly in the case of infants and children (e.g., Fort et al., 2018). As such, certain design factors of the studies themselves limited authors' abilities to use proficiency assessment methods.

Proficiency Assessment Method by Study Type

Considering the study type (quantitative, qualitative, and mixed methods), a clear trend emerges (Figure 2). Quantitative-oriented studies provided one or more assessments of proficiency in 82 of the 86 studies, approximately 95%. In contrast, qualitative studies, which were notably less represented in the data set, included proficiency assessments in 19 of 34 cases, approximately 55%. The rate of proficiency assessment in mixed method studies fell in in between those found in the quantitative and qualitative studies (17 of 20, 85%).

[Figure 2]

Considering the proficiency assessment method used in the different types of studies, again a clear difference is found between the different study types. In the quantitative studies that assessed proficiency, standardized testing (31.7%), quantitative self-reporting (30.4%), other and proficiency tests (18.3%), were the most common methods employed. In contrast, in qualitative studies, holistic assessments (31.6%) were the most common forms of proficiency assessment, followed by curricular standards and other proficiency questionnaires (26.3% each). Finally,

mixed methods studies relied most often on standardized testing (41.2%), other questionnaires (29.4%), and other proficiency tests (23.5%). The contrast is most notable between the quantitative and qualitative studies, with the two most common forms of proficiency assessment for the qualitative studies among the least represented in the quantitative studies (holistic assessment: 7.3%; curricular standards: 8.5%).

Proficiency Assessment Method by Subfield

First, considering the role of subfield, trends were analyzed with respect to the presence or absence of a proficiency assessment method (Figure 3). Across all subfields, the presence of a proficiency assessment method varied widely. Notably, several subfields were consistent in their use of a proficiency assessment method: psycholinguistics, cognitive linguistics, neurolinguistics (100%), vocabulary and lexical development (100%), semantics (100%), and syntax and morphology (95.7%). In contrast, several other subfields showed somewhat less reliance on proficiency assessment methods: language pedagogy (79.4%), pragmatics (77.8%), language attitudes, ideologies, and identity (68.8%), and sociolinguistics (64.3%). Although interpretation of these results should consider the representation of a given subfield in the data set, notable variation can be seen even among the most well-represented subfields.

[Figure 3]

Second, results were analyzed to compare the different proficiency assessment methods across the range of subfields. Figure 4 illustrates the use of each of the nine proficiency assessment methods across the nine most common subfields in the data set. The percentage illustrated in Figure 4 represents the frequency of each proficiency assessment method relative to the total number of identified proficiency assessment methods. Again, a single study could receive up to two different subfield designations and multiple proficiency assessment method tags. An initial examination of Figure 4 highlights the wide variation in proficiency assessment methods found within different subfields. In the nine most common subfields, all showed at least six different methods of proficiency assessment, and four showed eight or nine different methods of proficiency assessment. This finding highlights the overall diversity of proficiency assessment methods in bilingualism research.

Finally, it is worth considering some of the broad trends that emerge from the data, focusing on the most well-represented subfields. With respect to the subfield of psycholinguistics, cognitive linguistics, and neurolinguistics, there was a clear preference for proficiency assessment methods that provided a quantitative outcome. Of the 36 studies in this subfield that employed proficiency assessment, approximately 41.6% used some form of self-rating questionnaire and 41.6% used a standardized test. Also notable, the methods of curricular standards (5.6%), holistic assessment (11.1%), and institutional frameworks (13.9%) were less utilized in this subfield. In contrast, among studies employing proficiency assessment in the subfield of language pedagogy, quantitative self-rating was rarely used (3.7%), while standardized testing (37.0%) and institutional frameworks (33.3%) were the most common. While standardized testing was the most common proficiency assessment method in the data set, it was noticeably absent from two other well-represented subfields: language attitudes, ideologies, and identities (13.6%) and sociolinguistics (0.0%). These subfields instead relied on other types of questionnaires (attitudes and identities: 40.1%; sociolinguistics: 38.9%) and curricular standards (attitudes and identities:

27.3%; sociolinguistics: 27.8%). Again, as each study could receive up to two subfield tags, there was likely considerable overlap between these two subfields.

[Figure 4]

Discussion

The current study addresses the assessment of proficiency, a key methodological issue in the study of bilingualism. The discussion follows directly from the three initial research questions and provides some tentative recommendations for the field.

Directly related to Research Question 1, the analysis revealed that approximately 84% of studies employed one or more methods of proficiency assessment. Although the scope of this review is somewhat different than previous reviews (e.g., for SLA see Tremblay, 2011; for bilingualism see Hulstijn, 2012), this result is in line those studies. Considering the overall types of proficiency assessments found here (Table 4), there was broad overlap with the previous reviews (Hulstijn, 2012; Thomas, 1994; Tremblay, 2011). In the current study, standardized testing represented the most common approach to proficiency assessment. Curricular standards (e.g., number of years of study), the most common approach seen in past reviews (Thomas, 1994; Tremblay, 2011), were also well-represented in the data set. Notable innovation was found in the use of self-ratings, which were poorly represented in previous reviews (e.g., Tremblay, 2011: 7% of studies), but more prominent in the current data set (quantitative self-rating = 29%; qualitative self-rating = 15%). Given the long-standing call for including proficiency assessment in bilingualism research (e.g., Grosjean, 2008), it is noteworthy that there was no overall increase in the presence of proficiency assessments relative to past reviews of the literature. Moreover, the current study, as with previous reviews, highlighted a wide range of proficiency assessments found in the field.

Turning to the role of study type (RQ2) and subfield (RQ3) on proficiency assessment, both factors previously unaccounted for in large-scale reviews, the results showed clear trends for both the presence and type of proficiency assessment by study characteristics. For study type, for example, there was a clear trend for proficiency assessment methods to be used in quantitative research (95%) and less so in qualitative research (55%), and different types of assessment were found for each study type. Similarly, relevant differences emerged by subfield, with several fields showing near universal use of proficiency assessment (e.g., psycholinguistics) and others demonstrating a less consistent pattern (e.g., sociolinguistics). Moreover, the type of assessment was also subfield dependent, although significant variability remained. These trends highlight the importance of considering a study's characteristics in the selection of a proficiency assessment method.

While the data presented in this study have been parsed in two different ways (i.e., by study type and by subfield), in part to provide multiple study characteristics for future research to consider, it should be noted that these characteristics are not independent. For example, while some subfields showed broad representation of multiple study types (e.g., pragmatics: quantitative = 33%, qualitative = 22%, mixed = 44%), others skewed strongly towards one study type (e.g., psycholinguistics: quantitative = 97%, qualitative = 3%, mixed = 0%). For details on the distribution of study type by subfield see Appendix A. As such, although results for study type

and subfield may be driven in part by the interaction between these characteristics, researchers may choose to consider both in tandem when choosing proficiency assessment methods.

Evaluation and Selection of Proficiency Assessment Methods

The choice of whether to assess proficiency, and which type of proficiency assessment to choose, is a crucial decision in bilingualism research. The decision to include a measure of proficiency is related to a variety of factors, including the aims of the project, methods employed, and/or population (Hulstijn, 2012). If researchers determine that a proficiency assessment method is necessary, the choice of which proficiency assessment method to use should be carefully considered and thoroughly justified. In short, researchers should engage meaningfully with key decisions surrounding proficiency assessment and provide their criteria to readers. Potential aspects to consider include:

(1) Is proficiency assessment needed in this study to effectively answer the research questions? $
(2) How is proficiency theoretically conceptualized for the current study?
(3) Is proficiency in a single language, proficiency in both languages, or the relative measure of language dominance, most appropriate for the current study?
(4) What is the function of the proficiency assessment in the current study?
(5) Is a single method of proficiency sufficient or would this study benefit from multiple measures of proficiency?
(6) Are there subfield specific norms that may facilitate within subfield comparisons?
(7) Does a given measure allow for easy comparison across multiple subfields?
(8) Is there evidence for the validity and reliability of a given proficiency measure?
(9) What are the consequences, both positive and negative, of a given proficiency measure?

Considering these questions, a few possible standards are suggested here, with the acknowledgement that proficiency assessment should depend on a particular study's goals and that selection of a particular assessment is subject to a variety of linguistic, financial, and temporal constraints. Starting from the theoretical basis, authors would benefit from carefully considering the theoretical construct of proficiency (e.g., Hulstijn, 2015; Hyltenstam, 2016) and particularly avoiding the circular practice whereby the construct of proficiency is defined by the measures of proficiency (e.g., Lantolf & Frawley, 1988). Rather, selection of the measure of proficiency should be related to the *a priori* conceptualization of the construct of proficiency, and the nature of the links between concept and operationalization should be made explicit. While proficiency assessment methods that directly assess both linguistic knowledge and skills (e.g., Carroll, 1972) within an interactional context (e.g., Hymes, 1972), such as OPIs, may be impractical for a variety of reasons, authors should express the link between proficiency and their methodology for assessing proficiency. For example, if proficiency measured using standardized testing, researchers may acknowledge the implicit (assumed) link between this knowledge-oriented methodology and the broader skills and communicative function of proficiency (e.g., Hulstijn, 2011). Considering the function of the proficiency measure, Luk and Bialystok (2013) note that in bilingualism research, proficiency assessment may serve as an inclusionary criteria, to establish between groups comparisons (e.g., monolingual vs. bilingual), or as a correlate of specific linguistic tasks or behaviors. The function of the proficiency measure in a given study may guide the selection of the assessment, with more granular quantitative measures used when proficiency functions as a correlate of specific linguistic or cognitive tasks.

Moreover, given the overarching goal of creating research that is cross-comparable (Tremblay, 2011), the selection of a proficiency assessment measure should take into account overarching trends within the field, or as shown in the current review, trends within a given subfield or methodological approach.

In determining whether to use a proficiency assessment, it is recommended that all research in the field use some form of proficiency assessment, unless a concrete reason may be otherwise articulated. Inclusion of some form of proficiency assessment, even if not particularly relevant for the study's research questions, will allow future researchers to effectively reference study results. If proficiency is not measured, justification may be provided (e.g., this study employs previously collected data from anonymous sources, preventing any proficiency assessment). Given the broad definition of bilingualism (e.g., Montrul, 2016), proficiency measures should likely be collected in both languages when possible. Collecting proficiency assessment in both languages is a reasonable way to avoid the pitfalls of 'assuming' native competence in one of the participant's languages. This is particularly relevant given prior suggestions that proficiency may naturally vary in a bilingual's L1 (Treffers-Daller, 2011), bilingual speakers may undergo L1 attrition (Schmid, 2013), and the substantial evidence that learning an L2 impacts a speaker's L1, even at the beginning stages of L2 acquisition (de Leeuw & Celata, 2019). Moreover, collecting measures in both languages permits a degree of post-hoc approximation of language dominance, if dominance is not specifically addressed.

In determining which specific proficiency assessment to select, researchers may consider whether one or more measures is warranted, and if a given measure allows for easy comparison with previous literature, follows subfield and methodological trends, and has been evaluated with respect to validity and reliability. The use of multiple proficiency measures would also permit both following subfield specific norms, as found in the current results, and wider comparability across subfields. It is relevant to note that both standardized testing and quantitative self-ratings were among the most common forms of proficiency assessment across the data set in the current study, providing an initial point of comparison with previous literature. It is noteworthy that these two most common approaches represent objective and subjective measures of proficiency. Where possible, the use of multiple measures of proficiency should be entertained. Particularly recommended is combination of objective (e.g., standardized tests) and subjective (e.g., quantitative self-ratings) measures, which will both facilitate comparison with other research, as well as provide a degree of triangulation for proficiency measures. If multiple assessments are not practical or possible, researchers may look to the most common methods within a given methodological approach or subfield as a starting point. In comparing the most common approaches, relative to many standardized tests (e.g., TOEFL), subjective self-rating measures of proficiency have comparatively low barriers for usage. Broadly, they require little specialized training for administration and scoring, can be completed in a relatively short amount of time, and are available or easily translated into multiple languages (although see Delgado et al., 1999; Tomoschuk et al., 2019). In addition, given the fact that several self-rated questionnaires that have been evaluated for both validity and reliability are freely available, when feasible we recommend using established measures rather than self-rating questionnaires that are designed *ad hoc* for a single study. The particular self-rating questionnaire to be selected may be driven by subfield trends. As using multiple measures may be impractical in some circumstance, given the initial evidence in favor of reliability and validity, and the low barriers for usage, established,

quantitative self-rating questionnaires may represent a practical option for many studies in the field, particularly when only a single proficiency measure is employed.

Further considering the cross-comparative function of proficiency assessments, while curricular standards were well-represented in the data set, and among the most common in other systematic reviews (e.g., Tremblay, 2011; Thomas, 1994), as noted in Norris and Ortega (2000), they were largely presented without additional contextualizing information. That is, labels presented via curricular standards, such as years of study or a specific course enrollment (e.g., third-year French course), are "inexact" (Norris & Ortega, 2000, p. 455), resulting in heterogeneous outcomes (e.g., Tremblay, 2011) and representing an impediment to careful cross-comparison. Although such labels may suffice for a given study, it is difficult to predict how future research may benefit from comparison. For example, data presented in a study in which proficiency serves as a descriptor of the participants may be easily reinterpreted and combined with other studies to allow proficiency to serve as a variable of interest (e.g., independent variable). As such, it is strongly recommended that additional proficiency measures be used in conjunction with curricular standards where possible. Similarly, qualitative self-ratings and holistic assessments may be enhanced by contextualization and/or triangulation with other proficiency assessment methods.

Related to the validity of the proficiency assessment, following recommendations by Gaillard and Tremblay (2016), authors should choose proficiency assessments that have been evaluated for reliability and validity, and correlated with other measures of proficiency. Explicit discussion of the reliability and validity of a given measure may both justify the selection of a given proficiency assessment and inform future research on the topic.

And finally, each method of proficiency assessment has clear strengths and weaknesses that authors should weigh in relation to their own research aims. As an example, it is worth considering the two most commonly found methods of proficiency assessment in the current review: standardized testing and quantitative self-ratings. From a theoretical perspective, there remains debate about whether many standardized tests account for the "skills' component of proficiency (Thomas, 1994), instead focusing more heavily on decontextualized knowledge. Related to the function, it is clear that standardized testing may suffice for multiple different functions, although some tests may not be sufficiently granular for certain behavior–proficiency correlations. Many standardized tests have been well-evaluated for validity and reliability, and correlated with other measures of proficiency (e.g., Educational Testing Service, 2020b). Moreover, researchers may choose standardized tests for their comprehensive nature and to facilitate comparison with previous literature on the same language. Yet, as a result of the comprehensive nature, standardized tests are often time-consuming (Tremblay, 2011). Moreover, the language-specific nature of most standardized tests limits both cross-linguistic comparison and measures of dominance (Hulstijn, 2012), which require proficiency measures in each of a bilingual's two languages. It should be noted that the predominance of standardized testing in the current review is likely related to the languages under study. English and Spanish were by far the most common languages examined in the data set (Table 3), two languages for which large-scale standardized tests and their derivatives are widely available (e.g., TOEFL, DELE, etc.). Minority and/or indigenous languages without such tests are at a distinct disadvantage and at risk of being left out of comparative work. In contrast, quantitative self-ratings, may be designed to

encompass both knowledge and skills components of proficiency, and may suffice for a variety of different functions. With respect to validity and reliability, there is some evidence that quantitative self-ratings may correlate with other behavioral measures (e.g., Marian et al., 2007). Unlike standardized testing, quantitative self-rating questionnaires may be translated into a variety of languages, facilitating cross-study comparison (although for different ratings in different languages see Delgado et al., 1999; Tomoschuk et al., 2019). Yet, quantitative self-ratings may not offer a sufficiently granular measure, a focus on specific linguistic subcomponents, or an adequately objective perspective for some areas of study.

While there is still unlikely to be a sole, "best" measure of proficiency, choices surrounding proficiency assessments should be justified in study methodologies. In the current data set, most studies did not provide any specific rationale for their choices regarding proficiency assessment. In the future, authors should explicitly provide some rationale for the selection of proficiency assessment methods. These justifications may rely on previous points of comparison in the relevant literature, trends for a given subfield or methodology, or relevant constraints on researchers' or participants' time and resources. Adding a brief discussion of the proficiency assessment choices may also serve to further develop conversation surrounding subfield-specific norms and best practices.

Ongoing Calls for Consensus

Previous research has highlighted the "natural diversity" in proficiency assessment methods that arises from different study goals and characteristics (Thomas, 1994, p. 309). As noted by Tremblay (2011), while not all studies have the same needs related to proficiency assessment, some degree of consensus by subfield would facilitate cross-study comparison and building of scientific knowledge. The results of the current study serve both as a starting point for identifying some such methodological and subfield-specific consensuses and highlight the need for greater methodological consolidation. Specifically, the current results highlight some broad trends in the types of proficiency assessment methods by both methodological characteristics (Figure 3) and subfield (Figure 4). The existing trends illustrated here may serve as an additional consideration in the selection and justification of proficiency assessment methods, along with study goals, methodological characteristics, and/or population. Yet, at the same time, the results highlight the ongoing variability in proficiency assessment methods, with every subfield analyzed evidencing at least six different proficiency assessment methods. In short, while there appear to be some emerging subfield-specific norms, variability in proficiency assessment remains an obstacle to comprehensive, comparative work in the field. The findings here build on the previous calls by for more uniformity in proficiency assessment methods across the field (e.g., Tremblay, 2011) by specifically emphasizing the need for methodology-specific and subfield-specific conventions.

Acknowledgements

I am grateful to Ellen Deemer and Jacob King for their efforts in data coding. All errors remain my own.

Declaration of Conflicting Interests

No potential conflict of interest was reported by the author.

[1] While Montrul (2016) and others take a broad view of who is bilingual, distinctions in types or profiles of bilinguals may remain relevant. For a discussion of different bilingual profiles and terminology, see Wei (2000).

[2] An important distinction should be made between *proficiency* and *language dominance*, two related but easily "confusable" constructs (Birdsong, 2006, p. 47). Language dominance generally refers to a bilingual's relative abilities in each of their two languages and encompasses proficiency and other factors (e.g., language use, language history) (see Montrul, 2016). The current study examines proficiency assessment but acknowledges that these two constructs are variably employed in bilingualism research (Gertken et al., 2014). A detailed discussion of dominance is beyond the scope of the current paper.

[3] For example studies from the current data set that employ these different types of proficiency assessment, see the results section.

[4] Tremblay (2011) examined three journals from 2000–2008 (144 studies); Thomas (1994) examined four journals from 1988–1992 (157 studies); Hulstijn (2012) examined a single journal from 1998–2011 (140 studies).

[5] While SJRs provided a useful tool to collect relevant sources, they should not necessarily be taken as a direct assessment of the quality of any journal or individual study.

References

Alderson, C. (2006). *Diagnosing foreign language proficiency: The interface between learning and assessment.* Continuum Publishing

American Council of Teachers of Foreign Languages. (2012). ACTFL Proficiency Guidelines 2012. https://www.actfl.org/resources/actfl-proficiency-guidelines-2012

Anderson, J. (1971). Selecting a suitable 'reader': Procedures for teachers to assess language difficulty. *RELC Journal*, *2*(2), 35–42.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 535–556.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bahrick, H. P., Hall, L. K., Goggin, J. P., Bahrick, L. E., & Berger, S. A. (1994). Fifty years of language maintenance and language dominance in bilingual Hispanic immigrants. *Journal of Experimental Psychology: General*, *123*(3), 264–283. https://doi.org/10.1037/0096-3445.123.3.264

Barnwell, D. P. (1996). *A history of foreign language testing in the United States: From its beginnings to the present.* Bilingual Press.

Birdsong, D. (2006). Dominance, proficiency, and second language grammatical processing. *Applied Psycholinguistics*, *27*(1), 46–49. https://doi.org/10.1017/S0142716406060048

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, *1*(1), 1–47.

Carroll, J. (1972). Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allend & R. N Campbell (Eds.), *Teaching English as a second language* (pp. 313–320). McGraw-Hill.

Christiner, M., Rüdegger, S., & Reiterer, S. M. (2018). Sing Chinese and tap Tagalog? Predicting individual differences in musical and phonetic aptitude using language families differing by sound-typology. *International Journal of Multilingualism, 15*(4), 455–471. https://doi.org/10.1080/14790718.2018.1424171

Chung, E. S. (2018). Second and heritage language acquisition of Korean case drop. *Bilingualism*, *21*(1), 63–79.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, *13*(1), 3–21.

Cornell Language Acquisition Lab and Virtual Center for Language Acquisition (CLALVCLA). 2012. Virtual Linguistic Lab Child Multilingualism Questionnaire. Virtual Center for Language Acquisition, Cornell University. http://www.cornell.edu/vcla

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Curdt-Christiansen, X. L., & La Morgia, F. (2018). Managing heritage language development: Opportunities and challenges for Chinese, Italian and Pakistani Urdu-speaking families in the UK. *Multilingua, 37*(2), 177–200. https://doi.org/10.1515/multi-2017-0019

Dahm, R., & De Angelis, G. (2018). The role of mother tongue literacy in language learning and mathematical learning: is there a multilingual benefit for both?. *International Journal of Multilingualism, 15*(2), 194–213. https://doi.org/10.1080/14790718.2017.1359275

DeAnda, S., Bosch, L., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2016). The language exposure assessment tool: Quantifying language exposure in infants and children. *Journal of Speech, Language, and Hearing Research*, *59*(6), 1346–1356.

de Leeuw, E., & Celata, C. (2019). Plasticity of native phonetic and phonological domains in the context of bilingualism. *Journal of Phonetics*, *75*, 88–93. https://doi.org/10.1016/j.wocn.2019.05.003

Delgado, P., Guerrero, G., Goggin, J. P., & Ellis, B. B. (1999). Self-assessment of linguistic skills by bilingual Hispanics. *Hispanic Journal of Behavioral Sciences*, *21*(1), 31–46. https://doi.org/10.1177/0739986399211003

Dewaele, J. M. (2018). "Cunt": On the perception and handling of verbal dynamite by L1 and LX users of English. *Multilingua, 37*(1), 53–81. https://doi.org/10.1515/multi-2017-0013

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test*. Pearson Assessments.

Educational Testing Service. (2020a). *Test of English as a Foreign Language.* www.ets.org/toefl

Educational Testing Service. (2020b). Validity Evidence Supporting the Interpretation and Use of TOEFL iBT® Scores. *TOEFL® Research Insight Series, Volume 4*. https://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf

Ellis, C., Thierry, G., Vaughan-Evans, A., & Jones, M. W. (2018). Languages flex cultural thinking. *Bilingualism: Language and Cognition, 21*(2), 219–227. https://doi.org/10.1017/S1366728917000190

Flege, J. E., MacKay, I. R., & Piske, T. (2002). Assessing bilingual dominance. *Applied Psycholinguistics*, *23*(4), 567–598. https://doi.org/001:1O.1017.S0142716402004046

Fort, M., Ayneto-Gimeno, A., Escrichs, A., & Sebastian-Galles, N. (2018). Impact of bilingualism on infants' ability to learn from talking and nontalking faces. *Language Learning*, *68*, 31–57. https://doi.org/10.1111/lang.12273

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, *66*(2), 419–447. https://doi.org/10.1111/lang.12157

Gánem-Gutiérrez, G. A., & Gilmore, A. (2018). Tracking the real-time evolution of a writing event: Second language writers at different proficiency levels. *Language Learning, 68*(2), 469–506. https://doi.org/10.1111/lang.12280

Gertken, L. M., Amengual, M., & Birdsong, D. (2014). Assessing language dominance with the bilingual language profile. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 208–225). Multilingual Matters.

Grey, S., Sanz, C., Morgan-Short, K., & Ullman, M. T. (2018). Bilingual and monolingual adults learning an additional language: ERPs reveal differences in syntactic processing. *Bilingualism: Language and Cognition*, *21*(5), 970–994. https://doi.org/10.1017/S1366728917000426

Grosjean, F. (2008). *Studying bilinguals*. Oxford University Press.

Higgins, J. P. T., & Green, S. E. (2008). *Cochrane handbook for systematic reviews of interventions.* Wiley.

Hijazo-Gascón, A. (2018). Acquisition of motion events in L2 Spanish by German, French and Italian speakers. *The Language Learning Journal, 46*(3), 241–262. https://doi.org/10.1080/09571736.2015.1046085

Hopf, S. C., McLeod, S., & McDonagh, S. H. (2018). Linguistic multi-competence of Fiji school students and their conversational partners. *International Journal of Multilingualism, 15*(1), 72–91. https://doi.org/10.1080/14790718.2016.1241256

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229–249. https://doi.org/10.1080/15434303.2011.565844

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, *15*(2), 422–433. https://doi.org/10.1017/S1366728911000678

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins.

Hyltenstam, K. (2016). Introduction: Perspectives on advanced second language proficiency. In K. Hyltenstam (ed.), *Advanced proficiency and exceptional ability in second languages* (pp. 1–14). Walter de Gruyter.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *On communicative competence* (pp. 269–293). Penguin.

Ingram, D. E. (1985). Assessing proficiency: An overview on some aspects of testing. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 215–277). Multilingual Matters.

Jegerski, J. (2018). Sentence processing in Spanish as a heritage language: Self-paced reading study of relative clause attachment. *Language Learning, 68*(3), 598–634.

Jia, G., Aaronson, D., & Wu, Y. (2002). Long-term language attainment of bilingual immigrants: Predictive variables and language group differences. *Applied Psycholinguistics*, *23*(4), 599–621. https://doi.org/10.1017/S0142716402004058

Karimzad, F., & Sibgatullina, G. (2018). Replacing "them" with "us": Language ideologies and practices of "purification" on Facebook. *International Multilingual Research Journal, 12*(2), 124-139. https://doi.org/10.1080/19313152.2017.1401449

Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, *44*(3), 886-911. https://doi.org/10.1017/S0272263121000395

Köylü, Y. (2018). Comprehension of conversational implicatures in L2 English. *Intercultural Pragmatics*, *15*(3), 373–408. https://doi.org/10.1515/ip-2018-0011

Lantolf, J. P., & Frawley, W. (1988). Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10*(2), 181–195. https://www.jstor.org/stable/44488172

López, B. G., & Vaid, J. (2018). Fácil or A piece of cake: Does variability in bilingual language brokering experience affect idiom comprehension?. *Bilingualism, 21*(2), 340–354. https://doi.org/10.1017/S1366728917000086

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, *25*(5), 605–621. https://doi.org/10.1080/20445911.2013.795574

Luk, Z. P. S., & Shirai, Y. (2018). The development of aspectual marking in Cantonese-English bilingual children. *International Review of Applied Linguistics in Language Teaching, 56*(2), 137–179. https://doi.org/10.1515/iral-2014-0018

Ma, W., & Winke, P. (2019). Self-assessment: How reliable is it in assessing oral proficiency over time?. *Foreign Language Annals*, *52*(1), 66–86. https://doi.org/10.1111/flan.12379

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50,* 940–967. https://doi.org/10.1044/1092-4388(2007/067)

McAloon, P. (2015). From proficiency to expertise: Using HR evaluation methods to assess advanced foreign language and culture ability. In T. Brown & J. Brown (Eds.), *To advanced proficiency and beyond: Theory and methods for developing superior second language ability* (pp. 153–170). Georgetown University Press.

Menke, M. R., & Malovrh, P. A. (2021). The (limited) contributions of proficiency assessments in defining advancedness. In M. R. Menke & P. A. Malovrh (ed.), *Advancedness in second language Spanish: Definitions, challenges, and possibilities* (pp. 1–16). John Benjamins Publishing.

Montrul, S. (2016). Dominance and proficiency in early and late bilingualism. In C. Silva-Corvalán & J. Treffers-Daller (Eds.), *Language dominance in bilinguals: Issues of measurement and operationalization* (pp. 15–35). Wiley-Blackwell. https://doi.org/10.1017/CBO9781107375345

Montrul, S., & Slabakova, R. (2003). Competence similarities between native and near-native speakers: An investigation of the preterite-imperfect contrast in Spanish. *Studies in Second Language Acquisition*, *25*(3), 351–398. https://doi.org/10.1017/S0272263103000159

Mostafizar Rahman, A. R. M. (2018). Tense-lax merger: Bangla as a first language speakers' pronunciation of English monophthongs. *Asian Englishes, 20*(3), 220–241. https://doi.org/10.1080/13488678.2017.1327834

Nakayama, M., Lupker, S. J., & Itaguchi, Y. (2018). An examination of L2-L1 noncognate translation priming in the lexical decision task: insights from distributional and frequency-based analyses. *Bilingualism*, *21*(2), 265–277. https://doi.org/10.1017/S1366728917000013

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417-528. https://doi.org/10.1111/0023-8333.00136

Paradowski, M. B., & Bator, A. (2018). Perceived effectiveness of language acquisition in the process of multilingual upbringing by parents of different nationalities. *International Journal of Bilingual Education and Bilingualism, 21*(6), 647–665. https://doi.org/10.1080/13670050.2016.1203858

Polinsky, M., & Kagan, O. (2007). Heritage languages: In the 'wild'and in the classroom. *Language and Linguistics Compass*, *1*(5), 368–395. https://doi.org/10.1111/j.1749-818X.2007.00022.x

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments, *Journal of Speech, Language, and Hearing Research, 3*(2), 333–349. https://doi.org/10.1044/1092-4388(2009/08-0183)

Saito, K., Dewaele, J. M., Abe, M., & In'nami, Y. (2018). Motivation, emotion, learning experience, and second language comprehensibility development in classroom settings: A cross-sectional and longitudinal study. *Language Learning*, *68*(3), 709–743. https://doi.org/10.1111/lang.12297

Schmid, M. S. (2013). First language attrition. *Linguistic Approaches to Bilingualism*, *3*(1), 94-115. https://doi.org/10.1075/lab.3.1.05sch

Scimago. (n.d.). *Scimago Journal and Country Rank: Language and Linguistics.* Retrieved September 20, 2019, from https://www.scimagojr.com/help.php.

Seals, C. A. (2018). Positive and negative identity practices in heritage language education. *International Journal of Multilingualism, 15*(4), 329–348. https://doi.org/10.1080/14790718.2017.1306065

Surface, E. A., Poncheri, R. M., & Bhavsar, K. S. (2008). Two studies investigating the reliability and validity of the English ACTFL OPIc with Korean test takers. *The ACTFL OPIc® Validation Project Technical Report*.

Suzuki, Y., & Sunada, M. (2018). Automatization in second language sentence processing: Relationship between elicited imitation and maze tasks. *Bilingualism, 21(*1), 32–46. https://doi.org/10.1017/S1366728916000857

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–307.

Thomas-Sunesson, D., Hakuta, K., & Bialystok, E. (2018). Degree of bilingualism modifies executive control in Hispanic children in the USA. *International Journal of Bilingual Education and Bilingualism, 21*(2), 197–206. http://doi.org/10.1080/13670050.2016.1148114

Tigchelaar, M., Bowles, R. P., Winke, P., & Gass, S. (2017). Assessing the validity of ACTFL Can-Do statements for spoken proficiency: A Rasch analysis. *Foreign Language Annals*, *50*(3), 584–600. https://doi.org/10.1111/flan.12286

Tomoschuk, B., Ferreira, V. S., & Gollan, T. H. (2019). When a seven is not a seven: Self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition, 22*(3), 516–536. https://doi.org/10.1017/S1366728918000421

Treffers-Daller, J. (2011). Operationalizing and measuring language dominance. *International Journal of Bilingualism*, *15*(2), 147–163. https://doi.org/10.1177/1367006910381186

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition*, *33*, 339–372. https://doi.org/10.1017/S0272263111000015

Tschirner, E., Bärenfänger, O., & Wanner, I. (2012). Assessing evidence of validity of assigning CEFR ratings to the ACTFL oral proficiency interview (OPI) and the oral proficiency interview by computer (OPIc). http://doi.org/10.13140/RG.2.2.30276.68482

Universidad de Salamanca. (n.d.). *Dioloma de Español como Lengua Extranjera.* www.dele.org

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), 489–508. https://doi.org/10.2307/3586922

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, 29, 325– 344. http://doi.org/10.1177/0265532211424478

Wei, L. (2000). Dimensions of bilingualism. In L. Wei (Ed.), *The bilingualism reader* (pp. 3–25). Routledge.

Weigle, S. C., & Friginal, E. (2015). Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: Effects of language background, topic, and L2 proficiency. *Journal of English for Academic Purposes*, *18*, 25–39. https://doi.org/10.1016/j.jeap.2015.03.006

Wilczewski, M., Søderberg, A. M., & Gut, A. (2018). Intercultural communication within a Chinese subsidiary of a Western MNC: Expatriate perspectives on language and communication issues. *Multilingua, 37*(6), 587–611. https://doi.org/10.1515/multi-2017-0095

Zhou, W., & Zhou, M. (2018). Role of self-identity and self-determination in English learning among high school students. *Journal of Language, Identity & Education, 17*(3), 168–181. https://doi.org/10.1080/15348458.2018.1433537

Appendix A

[Table 5]