# A Meta-Analysis of the Effect of Ten Language Immersion Programs on Academic Achievement

**Sean R. Hill**
*Central Michigan University*

**Abstract:** Dual language immersion programs are becoming increasingly popular across the United States. Many sources have reported positive academic achievement for immersion students relative to their peers, especially after a few years in the program. However, not all authors have reported gains, and many of those gains are made with uneven groups. Further, many studies fail to report the effect sizes for practical significance. This meta-analysis examines ten studies that met the inclusion criteria to determine whether an immersion or bilingual effect exists in the data. Although there is an overall positive effect for immersion programs, it is small and the potential for unreported studies would likely cancel it out. Lastly, the meta-analytic results are still based heavily on quasi-experimental designs and non-randomized populations, indicating that the immersion effect on academic achievement should be considered cautiously.

**Key words:** dual language immersion, bilingual advantage, immersion effect, academic achievement, performance in content areas

Since the 1960s and the creation of the first French language immersion schools in Canada, proponents of dual language or bilingual education programs have often argued that students in these programs show an initial lag in performance relative to students in traditional programs. However, shortly after the introduction of instruction in English, immersion students perform on par and often surpass their peers. Indeed, this has led to what has been termed the immersion effect or the bilingual advantage, and enrollment in immersion education is cited frequently by parents as a way to increase the cultural capital of their children—to give them an advantage over their peers as they continue their education and enter the world of work (Williams, 2017).

Language immersion programs have been increasing in popularity throughout the United States, although the exact number of programs that are offered is difficult to determine. According to the Center for Applied Linguistics (2011), there were three immersion programs across the United States in 1971. By the end of the 1980s, there were still fewer than 100 programs nationally. By 2011, there were 448 schools that housed immersion programs. Some states have directed sustained efforts at creating dual language immersion (DLI) programs and provided top-down leadership and funding. For example, Utah passed an International Education Initiative bill in 2008 and began DLI programs in 15 elementary schools. As of 2018, there are currently 195 DLI schools in Utah alone (Lee, 2018). Similarly, the passage of Proposition 58 in California in 2016, which effectively repealed Proposition 227's English-only focus in 1998, allowed for the expansion of DLI programs for English language learners (ELLs) and more parental voice in the creation of DLI programming (Felton, 1999;

Hopkinson, 2017). This is expected to dramatically increase DLI programs and enrollment in those programs. As Williams (2017) has noted, when immersion programs are created, schools often reach maximum enrollment overnight. Despite parent and educator expectations of immersion students eventually surpassing peers in traditional classrooms, it is important to examine whether there is an immersion effect that indicates this outcome.

## Literature Review

Despite the rapid growth in the DLI program model, there are disparate results reported in the literature, as well as theoretically. Historically, bilingualism was seen as a detriment to total English fluency. MacNamara (1966) described the balance hypothesis, which supposed one total sum for language ability. Under this model, increased achievement in a second language came at a cost to the first. This was coupled with the development of reported confusion in the language development of bilingual children who often code-switched: they would begin speaking in one language and switch to a different language. Rather than viewing code-switching as a more precise or efficient way of expressing a point, researchers often proposed that this demonstrated the inability to master one language at the expense of a second. However, the early studies suffered from serious methodological concerns. For example, there was no control for socioeconomic status and recent immigrant groups were often used for the examples of diminished capacity and proficiency with English (Cummins, 1977; Diaz, 1983; Kirkici, 2004). Moreover, there was no accounting for the proficiency levels of the subjects, who sometimes performed better or worse than their monolingual peers. Lastly, linguistic elitism was present. Spanish was often referred to as a weak language, whereas English was considered a strong language. Indeed, much of the early research and the theories surrounding the development of bilingual students was rooted in research that focused on working class immigrant groups for whom English was a second language.

Because there had been little focus on the development of bilingualism and its academic effect on students whose first language was English, the results of the first French immersion schools for anglophone students surprised many when these students reached parity with anglophone students in traditional English-only curricula. Indeed, Cummins (1979) proposed the Linguistic Interdependence Hypothesis, elucidating two forms of bilingualism—additive and subtractive—in order to explain the paradoxical effects in the literature. Additive bilingualism is a condition where the inclusion and growth of an additional language increases the performance in a students' first language or in other academic subjects. On the other hand, negative bilingualism refers to the phenomenon where instruction in a second language reduces the proficiency of a student's first language and academic achievement in other areas.

Cummins (1979), who also examined the interaction of effects when a student was from a language majority or language minority group, proposed an underlying cognitive store of information regarding linguistic awareness and processing. Labeled the two-iceberg model, each language would have distinct surface features; however, the deep structures of the language had many characteristics that were shared. This is termed the common underlying proficiency. Cummins (1979) posited that as students became more proficient in a second language, they would pass a developmental threshold where effects would become additive rather than subtractive. Therefore, the theory that drives the beliefs about DLI programs considers paradoxical findings from the literature.

There have been few meta-analyses over the body of immersion research and most literature reviews in published works refer back to the same set of studies. One of the primary resources that is cited by many researchers is Swain and Lapkin's (1982) *Evaluating bilingual education: A Canadian case study*. They reviewed ten years of data from different boards of education in Canada and listed whether traditional or immersion students had a statistically significant advantage in their achievement in different subjects across many grade levels. Solely statistically significant scores were reported; no effect sizes were reported at this time. Likewise, they also reported whether or not there was a significant difference between the cohorts' scores. Although this work was important in synthesizing the effects of multiple immersion programs from the 1970s, this foundation does not report much statistically relevant information. Summarily, this case study effectively served as a meta-analysis of the country's French immersion programs, and has lent support to the idea that academic performance is typically lower in immersion students relative to their peers in the early grades but that those students reach parity or often surpass traditional students by the later grades. Essentially, the meta-analysis of these programs was an exercise in vote counting, representing a tally of statistically significant differences without an accounting of the size of those differences, only the direction.

Another oft-cited resource is Willig's (1985) "A meta-analysis of selected studies in the effectiveness of bilingual education," although its focus is not specifically on language immersion programs for language majority students. Rather, this analysis examines bilingual education and immersion programs in the context of students whose first language is not English. Although Willig contended that these program models support the academic achievement of students in bilingual education, her meta-analysis did not weight studies by the inverse of the variance (Borenstein, Hedges, Higgins, & Rothstein, 2009). Even though she included all effect sizes, a few studies accounted for the majority of her interpretations of the results. Further, her findings were rebuked by Baker (1987) at the U.S. Department of Education. Studies that cite Willig's (1985) meta-analysis supporting bilingual education fail to cite its criticisms.

The only meta-analysis of immersion programs for language majority students was conducted over 24 studies involving English immersion programs in Hong Kong. Lo and Lo (2013) reported that their results were inconclusive, although the overall mean effect size was -.28 with a mean *SE* of .01. Measures of variability were high, $Q(21)=1240$, $p < .01$, $T=.32$, $I^2=98.31\%$ (p. 57). Overall, immersion students were more proficient in their second language, English, relative to other students, but in general suffered in other academic areas, including their first language (p. 47). The only content area that appeared unaffected in the immersion setting was achievement in mathematics. In another study, although not a meta-analysis, the claim of an immersion effect and its role in potentially increasing student achievement in other content areas was recently reported with the first and second cohorts of immersion students in Utah. Watzinger-Tharp, Swenson, and Mayne (2016) found that "neither target language nor program significantly predicted student growth in math" (p. 11), indicating that immersion education was ineffective in increasing math achievement. On the other hand, students in the program did not experience a significant decline in mathematics achievement relative to their peers in traditional classrooms.

## Multiple Terms Create Havoc in the Literature

It is important to note that in the following discussion of terms, English is considered the main language and the United States is the country of reference. In other countries, the first language of the country would be considered the first language and language majority students would speak that language. However, there are additional concerns in that regard because English can also carry a prestige factor relative to other languages, which could differentially affect outcomes.

To begin, there are a plethora of terms used to describe this field, which the U.S. Department of Education's Office of English Language Acquisition (2015) has identified as an impediment to effective summary. These include first language (L1), second language (L2), and target language (TL) and may also describe English language learners (ELLs) or students in English as a second language (ESL) programs. This relates to the language status as well, where students that already speak English are considered language majority students, whereas ELL or ESL students are also classified as language minority students. This distinction is important for its potential to convey important information about the socioeconomic characteristics or ethnic heritage of the students. Much of the research is classified under bilingual education, which also overlaps with bilingual-bicultural education. Programs of this type have often been utilized with immigrant communities or indigenous groups. These models are also sometimes termed dual language immersion models, or more specifically two-way immersion or potentially developmental or transitional bilingual

programs. Programs that serve ESL or ELL students receive title funding under the Elementary and Secondary Education Act, currently authorized as the Every Student Succeeds Act. Even California's Proposition 58 is an example of the confusion in the terms: it specifically aims to increase ELL's access to two-way immersion programs and bilingual education. This is understandable because Proposition 227 sought to remove access to these programs for ELLs and immigrant groups (Felton, 1999; Hopkinson, 2017). However, it has been noted that many programs that begin as two-way immersion programs often become one-way immersion programs as these coveted opportunities change the economic demographics of school districts (Williams, 2017).

The aims of two-way immersion programs and the students they target are very different from those of students in one-way or foreign language immersion programs. One-way immersion programs typically consist of language majority students acquiring a second language. Two-way programs traditionally require one third to one half of the students in a cohort to be from a language minority group, in order for the students to have sufficient interaction with and models of authentic language use by other students (Tedick & Wesely, 2015). English immersion programs, on the other hand, are programs where ELL students only receive instruction in English and are sometimes labeled submersion programs because students either sink or swim in English (Hall, Smith, & Wicaksono, 2011). In this context, the students' ability in their first language is superfluous, and the goal is to reach English proficiency as quickly as possible, often at the expense of success in other core academic areas. Kennedy and Medina (2017) suggested that multiple studies have corroborated positive effects for ELLs in the dual language model.

Moreover, even one-way immersion programs have many different models. Programs can be either early or late. Early programs typically begin in either kindergarten or first grade, whereas late programs often begin in late elementary or early middle school grades. They can also range from total programs to partial programs (Tedick & Wesely, 2015). In total immersion programs, students often experience one to three years of instruction completely in the TL before instruction in English begins. The early French immersion programs often utilized this model. Full immersion, on the other hand, often utilizes a 90:10 or 80:20 split in TL to L1 usage, with the proportion of the L1 usage increasing each year. Partial or dual immersion programs often employ a 50:50 split in language usage. The partial or dual programs often began in response to parental concerns that their children were lagging relative to children in traditional classrooms, and would not be ready for national or state assessments in English language arts. Bournot-Trites and Reeder (2001) examined total versus partial immersion programs and students' math achievement with language proficiency as the manipulated variable. They reported support for Cummins's (1979) Linguistic Interdependence Hypothesis.

Lastly, the utilization of these terms does not include all the ways that immersion program models are described in the literature. Despite the confusion in terms listed above, this literature can also be found under searches for content-based language instruction or medium of instruction (Tedick & Wesely, 2015). Medium of instruction can therefore also be categorized as "L2 as the medium of instruction" or medium listed by language (Lo & Lo, 2014). Finally, immersion programs may also be listed under terms that are more generalized, such as examples in research related to second language acquisition. To summarize, it is difficult to conclusively synthesize the literature because the terminology used to classify it has not been fully standardized in the field.

## Method

### Literature Search and Inclusion Criteria

The literature search is presented following the structure outlined by Moher, Liberati, Tetzlaff, and Altman, (2009), which describes the original search returns for database 1 followed by closer readings to determine if the inclusion criteria are met for database 2. The ultimate inclusion criteria are included for database 3 and require close reading to ensure that articles have the necessary data reported in order to calculate effect sizes.

The search features included the different keywords listed above in the following databases: ERIC, PsycARTICLES, ProQuest Dissertations, and PsycINFO. For inclusion, articles had to be written in English and examine students in elementary and secondary education. College students or immersion experiences through study abroad were not included within the scope of the literature search. Also, additional articles were examined from the reference lists of articles returned. Searches of prior meta-analyses and narrative reviews returned few meta-analyses, which have been described above. A keyword search with one Boolean operator in Google Scholar was also conducted simultaneously for "language immersion, bilingual education, content-based language instruction, language-mediated learning, content and language integrated learning, language majority student" AND "academic achievement," excluding patents and citations, yielded 89 results in 0.04 seconds. Database 1 consisted of articles that appeared to relate to the meta-analysis topic.

Database 2 entailed reading the abstract for a study. Database 2 included articles that contained original data not found in other published and unpublished studies that compared students whose first language was English in immersion programs to the performance of students in traditional classrooms. Additionally, hand searches of journals that specialize in foreign language or bilingual research were conducted. All six years of the *Journal of Immersion and Content-Based Language Instruction* were reviewed, as well as the 21 volumes of the

*International Journal of Bilingual Education and Bilingualism*. *Foreign Language Annals* was searched from 2003 to 2018. One of the manual searches was also conducted of the American Council of Immersion Education's (ACIE) Newsletter publications from the Center for Advanced Research on Language Acquisition at the University of Minnesota. This resource was available from 1997 to 2011 and included 189 potential articles. Finally, landmark studies and other meta-analyses, such as those presented above, were also perused. However, as stated previously, many of the studies examined the effectiveness of bilingual education, which is not necessarily the same as immersion education in this context.

For inclusion in Database 3, studies also had to include either an effect size (Cohen's *d*), or provide enough information from mean, standard deviation, population size, or confidence intervals for the calculation of Cohen's *d*. Examples of studies that met the inclusion criteria for database 2 but not 3 because they did not report means, standard deviations, or *n* include Arthur (2004), Essama (2007), Haj-Broussard (2005), Watzinger-Tharp, Swenson, and Mayne (2016), among others. Arthur (2004) reported on the growth of the program from a single strand located within one elementary building to its expansion building-wide with a transition into middle school. Although this article is important in discussing the complexities of operating an immersion program amid competing concerns, the quantitative justification that was used for expansion solely compared the percentage of students that passed the standardized statewide ELA and mathematics assessments.

Although Arthur (2004) reported a clear advantage for students in the immersion program relative to the non-immersion students over two different testing periods by sixth grade, he also discussed overcrowding in non-immersion classrooms relative to small class sizes in the immersion classes as cohorts advanced through grade levels. Unfortunately, there was not enough data provided to calculate effect sizes. Essama (2007) compared students in a French immersion program to national averages and the 50$^{th}$ and 75$^{th}$ percentiles. When appropriate descriptive statistics were reported, they examined the performance of students in the immersion program by ethnicity. There was not a comparison between immersion and non-immersion students. Similarly, Haj-Broussard (2005) performed an ANCOVA to compare the ELA and mathematics performance of French immersion and regular education students, particularly focusing on African American performance. However, she did not report enough descriptive statistics to calculate Cohen's *d*. Another study, Jones (2005), only reported data for statistically significant results. Grade levels without statistically significant differences in test scores between immersion and non-immersion students were not reported, only *p* values. Even the recent article on the progress of students in Utah's dual immersion programs, which failed to show that immersion programming or language of instruction corresponded with student

growth in mathematics, did not provide descriptive statistics that could be used to calculate Cohen's *d* (Watzinger-Tharp, Swenson, & Mayne, 2016, p. 11).

### Study Coding

Ten studies remained which met the inclusion criteria outlined above. For each article, the names of the authors and dates of publication were recorded, as well as the state and/or country where the program was established. The publication status coding was established to distinguish journal articles, book chapters, and online newsletters. Dissertations and theses were classified as unpublished, and it was verified that these did not overlap with other published works. The type of program was then recorded as either total, full, or partial, with the approximate level of language time partition. Full programs included either a 90/10 or 80/20 split, and these programs were classified as full programs even if the immersion programs moved to a 50/50 model after two years. Similarly, if the programs began as a 50/50 model and transitioned to a full program, they were still classified as a partial or dual program.

The second primary characteristic recorded was whether the program was one-way or two-way. One-way foreign language programs are set up primarily to teach students a second language, whereas two-way programs provide instruction in two languages to an approximately even numbers of students that do and do not speak the language of instruction in order to improve language proficiency in both the L1 and L2. Next, the language of the immersion program reported was recorded. Only two articles recorded more than one language; the remaining articles had one language of instruction. Next, studies were coded for the grades that the program covered, as well as the approximate number of years of their participation in immersion education. Late immersion programs begin in either late elementary or middle school, and it is clearly important to distinguish between students in fifth grade that have had five or six years of immersion education and students with one year of immersion education. One study by Strickland and Hickey (2016) reported the student age; in this study immersion students were compared against all other Irish students in a national dataset.

Lastly, articles were coded with the necessary information to determine effect sizes. Each calculable effect size for the grade level was listed. Because each article generally provided multiple measures, and as assumptions of independent samples are contradicted with the inclusion of each effect size in the meta-analysis, effect sizes in Hedge's *g* are calculations from the multiple measures by instruction content area and for each study as a whole. Those are calculated using a fixed effects model and the variability and standard errors are reported for each study's overall effect size. Table 1 describes the characteristics of the 10 studies in Database 3.

Table 1. *Characteristics of Studies in Database 3*

| Author, Year | State, Country | Publication Status | Type of | 1 or 2 Way | Language | Grade (Years in imm.) | ES in Cohen's d or Hedge's g for mean ES |
|---|---|---|---|---|---|---|---|
| Artzer, 1990 | OH, USA | Diss. | Partial 50/50 | 1 | Spanish or French | 7th-8th | Total Mean ES=.193 (Var=0.013, SE=.115) Mean ELA ES=.205 Math=.521 Reading=.232 7th Read=.359 8th Read=.232 7th Lang=.252 8th Lang=-.069 Mean Math ES=.169 (Var=.040, SE=.200) 7th Math=-.083 8th Math=.521 |
| Fortune & Song, 2016 | USA | Journal | Total | 1 | Chinese | 3(3-4) – 5(5-6) | Total Mean ES=.174 (Var=0.003, SE=.050) Mean Math ES=.439 (Var=.004, SE=.061) 5th Math=.175 3rd Math=.523 Fixed ELA Mean ES=-.056 (.005, SE=.071) 5th ELA=.005 3rd ELA=-.086 |
| Greene VonCannon, 2015 | NC, USA | Diss. | Full | 1 | Spanish | 1st (2) | Total Mean ES=.045 (Var=.007, SE=.083) Text Reading Comp=-.383 Nonsense Words CLS=-.019 Nonsense Word/ WW Read=.271 Oral Comp=-.059 Written comp=.647 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Marian, Shook & Schroeder, 2013 | IL, USA | Journal | Partial K-2 to Full 3 -5 | 2, only Lang Maj used | Spanish | 3rd (4) – 5th (6) | Total Mean ES=.614 (.007, SE=.084)<br>Mean Read ES=.513 (Var=.014, SE=.118)<br>Mean Math ES=.716 (Var=.014, SE=.119)<br>3rd Read=.634<br>4th Read=.442<br>5th Read=.360<br>3rd Math=.712<br>4th Math=.722<br>5th Math=.721 |
| Mukai, Downs & Sato, 2005 | AK, USA | Journal | Partial | 1 | Japanese | 4th (4-5) and 7th (max 7) | Total Mean ES=.551 (Var=.004, SE=.066)<br>Read=.402<br>Lang=.475<br>Math=1.115<br>Science=.571<br>Soc. Stud=.222 |
| Strickland & Hickey, 2016 | Ireland | Journal | Not known | 1 / 2 | Gaelic | 9-year-olds | Total Mean ES=.070 (Var=0.002, SE=.041)<br>Math=.060<br>ELA=.079 |
| Jacobsen, 2013 | NC, USA | Diss. | Full, 65:35 | 1 | Chinese | 3rd-5th (4-6) | Total Mean ES=.545 (Var=.008, SE=.087)<br>Mean Read ES=.400<br>5th Read=1.049<br>4th Read=.492<br>3rd Read=.424<br>Mean Math ES=.551<br>5th Math=.949<br>4th Math=.509<br>3rd Math=.626<br>5th Science=.946 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Hill & Otani, 2014 | MI, USA | Thesis | 50/50 | 1 | Chinese | K-3rd (3/4) | Total Mean ES=.170 (Var=.007, SE=.080)<br>3rd Math=.723<br>3rd Math State=.500<br>2nd Math=.120<br>1st Math=.089<br>K Math=-.283<br><br>3rd Read=-.412<br>3rd ELA State=.252<br>2nd Read=.224<br>1st Read=.007<br>K Read=.159 |
| Thomas, Collier & Abbott, 1993 | VA, USA | Journal | 50/50 | Most 1, some 2 | Spanish (4), Japanese (3), French (1) | 1-3(1-2) | Total Mean ES=.050 (Var=0.001, SE=.033)<br>Mean Read ES=.216 (.005, SE=.068)<br>Jap 2nd/3rd Read=.111<br>All, 1st/2nd Read=.242<br>Mean Math ES=-.002 (.002, SE=.038)<br>Jap 2nd/3rd Math=.166<br>All, 1st/2nd Math=0.00<br>Jap 2nd Math=.428<br>All, 1st Math=-.091 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Padilla, Fan, Xu & Silva, 2013 | CA, USA | Journal | 50/50 | 2 | Chinese | 2(2/3)-5(5/6) | Total Mean ES=0.049 (Var=.003, SE=.055)<br>Mean Math ES=.302 (.007, SE=.081)<br>5th Math=.853<br>4th Math=.512<br>3rd Math=.147<br>2nd Math=.232<br>Mean ELA/Wr ES=-.167 (.006, SE=.075)<br>5th ELA=.180<br>4th ELA=.311<br>3rd ELA=.088<br>2nd ELA=-.828<br>4th Writ12=.110<br>4th Writ13=.334 |

***Note. Hedge's g provides the correction for the number of participants in the studies. Studies with smaller sample sizes are more like to return more extreme effect sizes so Hedge's g takes the sample sizes into account.***

Notes on Two Studies

Jacobson (2013) utilized three comparison schools: a local neighborhood school with similar rates of free and reduced lunch (NSS), a magnet school with similar rates of free and reduced lunch that performed higher academically (MSS), and another neighborhood school with lower rates of free and reduced lunch (NSL) with high academic achievement (p. 89). For this analysis, the magnet school (MSS) was chosen under the assumption that parents would have made a choice to send their children to this program and that socioeconomic status effects would be mitigated. Lastly, the combined effect size listed for Jacobson (2013) for math and reading is based off the total math and reading scores listed in the dissertation, whereas the total fixed mean effect size is calculated using a fixed effect model of the combined math, combined reading, and fifth grade science scores. Each specific grade level effect size in math and reading is presented for grade level comparisons.

In order to include Padilla, Fan, Xu, and Silva (2013), it is important to note the procedure that I followed. It was possible to determine the effect size from a calculation of frequency scores under each score on a scale of 0 to 8 or 1 to

4, depending on the year. However, only proficiency ratings were provided for the math and ELA scores, as well as for the proportion of students and the total number of immersion or traditional that fell under five proficiency categories ranging from far below proficient to advanced. There was no scaled score or standard deviation provided for these students. One must therefore be careful when considering the effect sizes, as they are based on the mean scores and standard deviations of the proficiency categories, which are ordinal data. Commonly, state education agencies provide scales that list the advanced category with the highest number. I followed this ordinal coding and determined that far below proficient=0, below proficient=1, basic proficiency=2, proficient=3, and advanced=4. In this way, I was able to include effect sizes that I assume to be close to what scaled score means and standard deviations would have provided.

## Results

This meta-analysis followed Borenstein, Hedges, Higgins, and Rothstein's (2009) method for performing meta-analyses and incorporated Hedge's correction for small sample size. A discussion of different methods for performing a meta-analysis are described in the limitations and directions for future research section. For some parts of the analysis, the 58 coded effect sizes are discussed and treated as independent samples. The diagnostic tests, on the other hand, look solely at the ten studies and the calculated composite effect size and variation from each one.

### Main Effect

Despite the paradoxical claims presented in the theoretical frameworks and the range of effects presented in Table 1, the meta-analysis of the effects of immersion education on students whose first language is not the L2 indicates an overall very small positive effect. However, it is important to note that the confidence intervals indicate that the effect hovers around null results. Table 2 presents the results with all 58 coded effect sizes in the model. However, the $z$ value is to be discarded because the $k$ is based off those 58 effect sizes, which were derived from ten studies, indicating that the assumption of independent samples is broken. The fixed effects results are presented as a comparison to the random effects model. The fixed effects model suggests that there are no other contributing factors to the achievement scores of students in these programs other than a main program effect. The random effects model indicates a small mean effect size and the $I^2$ suggests much underlying variability that is not explained by the main effect of the program. Therefore, it is likely that moderators play a key factor determining the overall effectiveness of immersion education. Calculations were completed in R.

Table 2. *Summary Report for Meta-Analysis of Academic Achievement for Immersion Students: All Effect Sizes as Independent Samples*

| Statistic | Fixed Effects Results | Random Effects Results |
|---|---|---|
| K | 58 | 58 |
| Mean ES | 0.17 | 0.26 |
| z(p-value) | 9.64(.0000) | 5.91(.0000) |
| SE | 0.02 | 0.04 |
| 95% CI | 0.14, 0.21 | 0.17, 0.34 |
| Q(p-value) | -- | 284.53 (.0000) |
| Tau-squared | -- | .0745 |
| I-squared | -- | 79.97% |
| 95% Credibility Interval | -- | -0.29, 0.81 |

In order to maintain the assumption of independent samples, mean effect sizes and variability were calculated for each study independently, despite the differences in grade levels or content areas. As such, the *k* is greatly reduced from 58 in the previous table. Using the mean effect size and variability of each study, the overall main effect of immersion education was still small in the fixed effects model, although the confidence intervals no longer included negative effect sizes. Assuming that there are other factors that determine immersion student performance relative to those from traditional classrooms, a random effects model was performed. The meta-analytic results indicate a statistically significant, small positive effect for immersion education. The confidence intervals do not include zero, lending further support for a main effect due to immersion. However, the large $Q(9)=108.46$, $p < .000$, $T^2=.0395$, and $I^2=91.7\%$, strongly suggest that there are moderating variables affecting student achievement between immersion and traditional students.

Moderator Analyses

Due to the high levels of variability surrounding the composite effect sizes, it is highly probable that other factors play a vital role in determining how immersion students' performance relates to that of students in traditional classrooms. Therefore, moderator analyses were performed. Table 4 displays the English language arts mean effect sizes performed as a separate meta-analysis. The fixed effects result is for the purpose of comparison with the random effects model. This analysis was calculated with R. Although it is an ecological fallacy to directly compare the effect sizes with the *z* score—these are separate meta-analyses which do not include a pooled tau squared, and the samples are not independent—it

appears that the subject content of the ELA meta-analysis has a lower effect size relative to the mean when all effect sizes were included in the analysis. Relative to the model with one overall effect size per study, this pattern also appears. It is plausible that immersion education affects student achievement to less of a degree in the content of ELA relative to mathematics.

Table 3. *Summary Report for Meta-Analysis of Academic Achievement for Immersion Students: Composite Effect Size for Each Study*

| Statistic | Fixed Effects Results | Random Effects Results |
|---|---|---|
| K | 10 | 10 |
| Mean ES | 0.16 | 0.24 |
| z(p-value) | 9.03(.0000) | 3.60(.0003) |
| SE | 0.02 | 0.07 |
| 95% CI | 0.13, 0.20 | 0.11, 0.37 |
| Q(p-value) | -- | 108.46 (.0000) |
| Tau-squared | -- | .0395 |
| I-squared | -- | 91.70% |
| 95% Credibility Interval | -- | -0.24, 0.72 |

***Note. Random Effects model is calculated with the method of moments.***

Table 4. *Separate English Language Arts Meta-Analysis as Moderator Report for Academic Achievement for Immersion Students: All Effect Sizes as Independent Samples*

| Statistic | Fixed Effects Results | Random Effects Results |
|---|---|---|
| K | 32 | 32 |
| Mean ES | 0.13 | 0.16 |
| z(p-value) | 4.88(.0000) | 2.77(.0056) |
| SE | 0.03 | 0.06 |
| 95% CI | 0.08, 0.18 | 0.05, 0.27 |
| Q(p-value) | -- | 124.23 (.0000) |
| Tau-squared | -- | .0671 |
| I-squared | -- | 75.05% |
| 95% Credibility Interval | -- | -0.38, 0.70 |

***Note. Random Effects model is calculated using the method of moments.***

Table 5 displays the mathematics mean effect sizes performed as a separate meta-analysis. The fixed effects result is for the purpose of comparison with the random effects model. This analysis was calculated with R. The same problems of direct comparison are present with the mathematics meta-analysis, but it appears that immersion education may have a greater impact on mathematics achievement. It is important to note, however, the large $I^2$ present in both separate analyses. Other factors potentially affect achievement to a greater extent.

Table 5. *Separate Mathematics Meta-Analysis as Moderator Report for Academic Achievement for Immersion Students: All Effect Sizes as Independent Samples*

| **Statistic** | **Fixed Effects Results** | **Random Effects Results** |
|---|---|---|
| K | 26 | 26 |
| Mean ES | 0.20 | 0.36 |
| z(p-value) | 8.15(.0000) | 5.33(.0000) |
| SE | 0.02 | 0.07 |
| 95% CI | 0.15, 0.24 | 0.23, 0.49 |
| Q(p-value) | -- | 151.24 (.0000) |
| Tau-squared | -- | .0825 |
| I-squared | -- | 83.47% |
| 95% Credibility Interval | -- | -0.25, 0.97 |

***Note. Random Effects model is calculated using the method of moments.***

Further moderating effects analyses would consider separate pooled and non-pooled tau-squared between the groups to better estimate the true score. In addition, a meta-regression would be useful to check for a continuous moderator based on the grade level and years of instruction in immersion programming. Other categorical moderators of interest include language of instruction, one-way versus two-way programming, and partial or dual language immersion versus total and full immersion programs. Similarly, it would be interesting to test whether there is a difference in effect sizes and variability between the published and unpublished studies.

## Diagnostic Analyses

Meta-analyses suffer from the potential of publication bias. Studies that report significant results are more likely to be published relative to studies which effect little practical significance. It is therefore difficult to determine from the

studies that were coded whether immersion education has a measurable effect on academic achievement. With so few published and unpublished results in the literature, it is probable that other studies have not been published or completed.

This is an especially important consideration with programming that takes place in a dynamic environment like that of public education. Due to the increased accountability standards faced by school districts, ineffective programs that either produce null or negative effects on student achievement are unlikely to continue. Even when programs are successful, peripheral problems related to the exclusive nature of immersion programs with advancing grade levels and student attrition may negatively impact even a successful program's longevity. One of the author's own manuscripts was in part rejected for publication because an immersion program was terminated despite the academic performance of participating students. The article reviewers specifically stated that the program's termination did not fit with the report's aims of performance and possibilities (J. Foss, personal communication, 8 Jan. 2017).

Failsafe K

Corwin (1983) provided one way to test for the likelihood of publication bias. Using a fail-safe *K*, a calculation is performed to determine the number of studies that would have to report a null or negative effect. Corwin (1983) provided the formula:

$$k_{fs} = \frac{k_{obt}(\bar{d}_{obt} - d_c)}{d_c - \bar{d}_{fs}}$$

to determine the number of studies needed to nullify the effect seen in the literature. The *d*bar_obtained is the mean effect size from the analysis.

With the 58 effect sizes coded, the *d*bar_obtained was 0.26. The $d_c$ is the criterion level for a trivial effect size. In this case, $d_c$ will be 0.10. The failsafe *d*bar, typically 0.0, will be -0.30 for this analysis because negative effects are reported in the literature. Therefore, the failsafe *K* in this case is 23, the number of effect sizes that would have to be reported that fall within those bounds to obtain the critical *d* value. If the failsafe *d*bar were -0.50, the number of effect sizes needed would diminish to 15 studies.

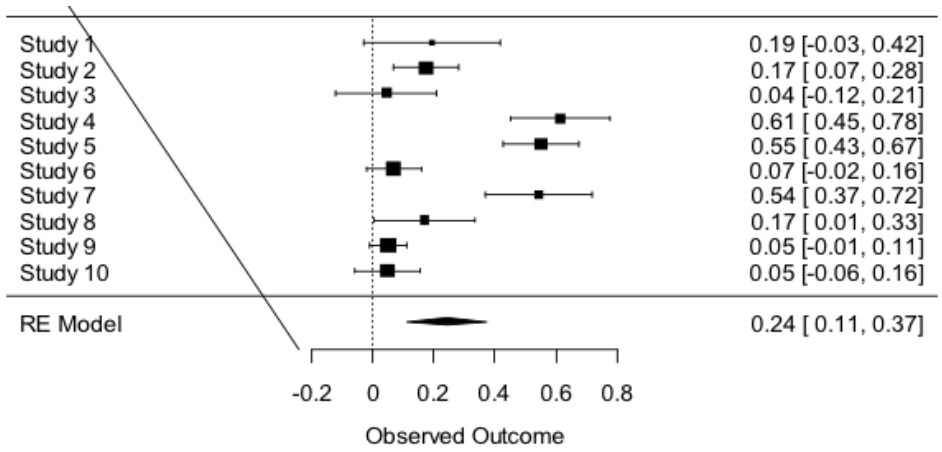
Using the 10 studies with an average effect size computed for each, the obtained *k* value is 10. Instead of Cohen's *d*, Hedge's *g* is used. The average obtained effect size from the ten studies is 0.25. Using the first scenario with the criterion *g* of 0.10 and the failsafe *g* of -.30, the failsafe *k*=4 studies. With the increased failsafe *g* of -0.50, the failsafe *k* rounds up to three. Therefore, it is very plausible that few studies are required that demonstrate effects that contradict the often-positive effect reported in most studies. Using Dahlke and Wiernik's

(in press) automatic settings in R with the Orwin approach and a target effect size of 0.10, the fail-safe *k* is fifteen studies.

Forest Plot

A forest plot indicates the mean, weight, and confidence levels of each of the studies in the meta-analysis. High variability is indicated by the spread of the confidence intervals and between study variability is indicated by the extent to which the confidence intervals overlap. The results of the random effects model are presented in Figure 1 with the mean effect size and the confidence intervals built into it. As can be seen, the first study has a wide range associated with it because of the little weight it carries, whereas the ninth study has a very tight range for its confidence intervals given its weight. Although the mean effect size is slightly above .20, most of the studies fall between the mean effect size and zero, with some confidence intervals indicating potential negative effects on academic achievement due to education through a foreign language in immersion programs relative to the students' first language in traditional classroom instruction.

*Figure 1.* **Forest Plot of Ten Studies**




Funnel Plot & Trim and Fill

Funnel plots describe the standard error and the observed effect size of a study. The pyramid that is formed has its apex at the mean effect size and studies closer to the top represent those with very small standard errors. This funnel plot and the accompanying trim and fill plot examine the ten principle studies in this analysis and the composite effect sizes for each study. The trim and fill plot is a way to determine the likelihood of publication bias. Where present, an open

circle would indicate the presence of a study that would indicate an unpublished study. Although there are approximately equal weights outside of the triangle, I am surprised to see that there are not any indicated missing studies inside the pyramid. It is important to note that many of the studies have a relatively high standard error, indicating the variability inherent with small effect sizes. These plots are shown in Figure 2.

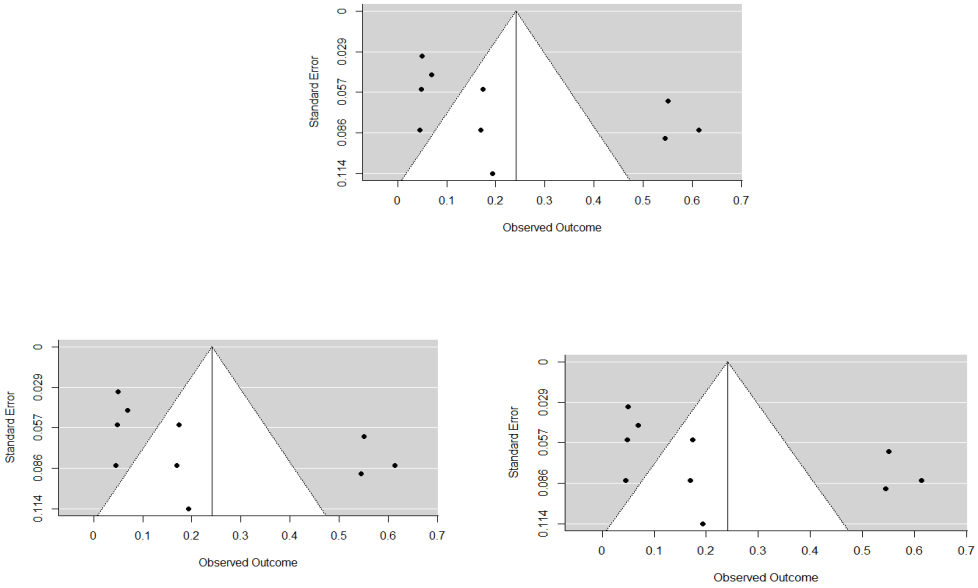



*Figure 2*. **Funnel Plot & Trim and Fill Plot of the Ten Studies**

Table 6. *Cumulative Record of Each Study Added to the Random Effects Model*

| Study | Estimated ES | SE | Z | P | CI | QE | QEp | $Tau^2$ | $I^2$ | $H^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.19 | 0.11 | 1.69 | 0.0905 | -0.03, 0.42 | 0 | 1 | 0 | 0 | 1 |
| 2 | 0.18 | 0.05 | 3.60 | 0.0003 | 0.08, 0.27 | 0.02 | 0.88 | 0 | 0 | 1 |
| 3 | 0.14 | 0.04 | 3.37 | 0.0007 | 0.06, 0.23 | 1.88 | 0.39 | 0 | 0 | 1 |
| 4 | 0.26 | 0.12 | 2.12 | 0.0340 | 0.02, 0.49 | 27.04 | 0 | 0.0512 | 88.90 | 9.01 |
| 5 | 0.32 | 0.11 | 2.81 | 0.0049 | 0.10, 0.54 | 44.83 | 0 | 0.0573 | 91.07 | 11.21 |
| 6 | 0.27 | 0.10 | 2.74 | 0.0061 | 0.08, 0.47 | 65.63 | 0 | 0.0541 | 92.38 | 13.13 |
| 7 | 0.31 | 0.09 | 3.31 | 0.0009 | 0.13, 0.50 | 76.69 | 0 | 0.0560 | 92.18 | 12.78 |
| 8 | 0.29 | 0.08 | 3.51 | 0.0004 | 0.13, 0.46 | 77.74 | 0 | 0.0499 | 91.00 | 11.11 |
| 9 | 0.26 | 0.07 | 3.55 | 0.0004 | 0.12, 0.41 | 103.55 | 0 | 0.0443 | 92.27 | 12.94 |
| 10 | 0.24 | 0.07 | 3.60 | 0.0003 | 0.11, 0.37 | 108.46 | 0 | 0.0395 | 91.70 | 12.05 |

Cumulative Effect

A cumulative effect record helps to understand how individual studies fit the specified model, with each being added at a time. Table 6 displays how each study changes the effect size and heterogeneity statistics of the random effects model.

Sensitivity analysis

A sensitivity analysis is a diagnostic tool that helps provide insight into which studies affect different aspects of the meta-analysis by way of a leave-one-out diagnostic. In particular, this analysis provides a way to quickly scan for outliers and for studies that have a large influence over the meta-analytic results. According to Viechtbauer (n.d.), one of the quick tells of a study's influence is if it has a DFBETAS value greater than one, among other checks. Figure 3 provides a visual display of the diagnostics. Studies 4 and 5 consistently appear to exert a larger influence than most of the other studies. Table 7 provides this information quantitatively.

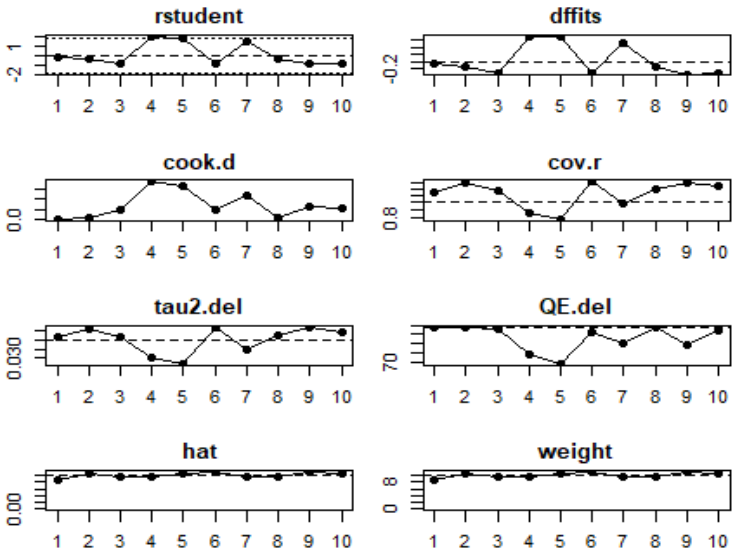



**Figure 3. Diagnostics for the Ten Studies in the Meta-Analysis**

Table 7. *Diagnostic Results for the Ten Studies in the Meta-Analysis*

| Study No. | Authors, Date | R Student External Standardized Residuals | Dffits Value | Cook's d | Cov r | Tau2. del | QE.del | Hat | Weight | Dfbetas inf. |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Artzer, 1990 | -0.22 | -0.07 | 0.01 | 1.14 | 0.04 | 108.40 | 0.09 | 8.53 | -0.07 |
| 2 | Fortune & Song, 2016 | -0.32 | -0.12 | 0.02 | 1.29 | 0.05 | 108.42 | 0.11 | 10.53 | -0.12 |
| 3 | Greene VonCannon, 2015 | -0.93 | -0.31 | 0.10 | 1.16 | 0.04 | 106.36 | 0.10 | 9.62 | -0.31 |
| 4 | Marian et al., 2013 | 2.05 | 0.69 | 0.37 | 0.86 | 0.03 | 78.06 | 0.10 | 9.62 | 0.69 |
| 5 | Mukai et al., 2005 | 1.89 | 0.69 | 0.33 | 0.79 | 0.03 | 67.59 | 0.10 | 10.29 | 0.68 |
| 6 | Strickland & Hickey, 2016 | -0.82 | -0.29 | 0.10 | 1.32 | 0.05 | 103.22 | 0.11 | 10.78 | -0.29 |
| 7 | Jacobsen, 2013 | 1.56 | 0.51 | 0.23 | 0.97 | 0.03 | 89.50 | 0.09 | 9.42 | 0.51 |
| 8 | Hill & Otani, 2014 | -0.34 | -0.12 | 0.01 | 1.18 | 0.04 | 108.46 | 0.10 | 9.62 | -0.12 |
| 9 | Thomas et al., 1993 | -0.93 | -0.33 | 0.13 | 1.30 | 0.05 | 89.26 | 0.11 | 11.05 | -0.34 |
| 10 | Padilla et al., 2013 | -0.94 | -0.33 | 0.12 | 1.24 | 0.04 | 103.55 | 0.11 | 10.53 | -0.33 |

## Discussion

The question of the utility of immersion education in improving the academic performance of students whose first language is English with considerable class time spent learning through the target language is an important consideration for parents as well as school leaders as the dual language model has expanded (Center for Applied Linguistics, 2011). Considering that accountability measures are often tied to school funding formulas, it is important for school administrators to understand the potential impacts on student achievement with the implementation of an immersion program. Although the results reported publicly are generally positive, the theoretical model and some examples in the literature indicate potential detrimental effects.

This meta-analysis examined ten studies that met the inclusion criteria outlined above. Two types of analyses—full inclusion of all effect sizes and one composite effect size per study—were conducted to examine the immersion effect. In both analyses, the overall effect was positive, albeit small. Immersion students may be predicted to rank a quarter standard deviation above students in traditional classrooms. When examining differential effects on English language arts or mathematics achievement, the positive effect on mathematics

achievements was slightly larger. However, there is considerable variability within these analyses and it is likely that other factors play a more important role in determining student achievement. For instance, a recent study that did not meet inclusion criteria, which examined the effects of three years of immersion education in Utah, did not demonstrate that immersion programming or the specific language of instruction predicted greater achievement on fourth grade mathematics performance (Watzinger-Tharp, Swenson, & Mayne, 2016). Other factors, such as gender and socioeconomic status as proxied by free and reduced lunch status, were important in predicting performance.

Unfortunately, the diagnostic analyses indicated the potential for publication bias in the reported results in the literature. It would only take a few studies with a moderate negative effect size to negate the small, positive effect sizes reported in this manuscript. Due to the accountability measures tied to public education systems, it is probable that unsuccessful or ineffective programs were quickly terminated. Gregg Roberts, one of the principal actors in Utah's immersion boom, had stated that he had never heard of the cancellation of an immersion program until I mentioned how the program I headed was terminated (personal communication, 18 November 2017). As stated earlier, failed programs are unlikely to be reported in the literature.

Further, the confusion in terminology between immersion and bilingual education programs and the many different program models involved can often lead to positive results reported for groups that do not coincide with the targeted students. For example, many positive results are reported for language minority students in bilingual or immersion programs, particularly in two-way or developmental bilingual programs. In addition to the recognition of heritage language diversity as an asset, California's Proposition 58 opens the possibility for students, particularly ELLs, to access two-way immersion programs. It is expected that these programs will yield positive academic results for students as they learn English. While these results are promising, they bolster the argument that students learn content best when taught in their first language or with other language supports.

It is important to note that the positive effects often reported in the immersion research can be traced to potentially unequal groups. These groups can be seen through the lens of demographic and sociolinguistic variables that may begin as small differences but yield greater influence on academic achievement over program sequences. Indeed, Hill and Otani (2014) found in one school that there were more students that qualified for special education programming in the group of students in traditional classrooms relative to immersion classrooms. Similarly, students' home language can mediate performance in school, and immersion programs may include large numbers of ELLs in two-way models. Interestingly, Watzinger-Tharp, Swenson, & Mayne (2016) differentiated the average growth percentiles of students by target language, but also explicitly

contrasted the growth in Spanish one-way and Spanish two-way programs. By definition, the two-way immersion programs consist of classrooms in which approximately half the students already speak Spanish. Although the overall Spanish two-way performance was summarily lower than other programs or target languages, the error bars overlapped. However, the authors indicated that "because of large amounts of variance among the schools, these differences are not statistically significant" (p. 11). Likewise, status as a linguistic minority or majority group in a two-way program can often imply important demographic information about a community with the potential to affect student achievement.

Williams (2017) explicitly states how parents see immersion programs as a way to increase the cultural capital of their children. Middle-class parents in particular have fought to enroll their children in these programs, even soliciting charter schools to create the model. Parents that take an active part in their children's education are likely to contribute to their children's academic achievement. Because parents traditionally elect to enroll their children in these programs, it can be posited that children of greater socioeconomic means tend to fill immersion seats relative to poorer children. In two-way immersion programs, where there are often two equal groups of language majority and language minority students, the language majority students tend to have lower rates of free or reduced lunch than the language minority students. Moreover, Williams (2017) noted that two-way dual language programs often evolve into one-way programs as middle-class parents put pressure on districts to offer more immersion options for their children. Schools that are schools-of-choice can experience an influx of out-of-district students whose affluent parents can provide private transportation to a school that offers these programs. Therefore, it is difficult to make the most basic of statistical assumptions when there are unequal groups. The comparisons that are most likely to be reported in the literature are most likely unequal prior to the beginning of formal education.

Lastly, immersion programs stand in contrast to another reality in public education—student attrition. Although all students can begin an immersion track in kindergarten, it is difficult for students that do not have prior language exposure or proficiency to enter an immersion program after the first grade. Attrition rates are often related to poverty as more transient students tend to live in less financially stable homes (Mehana & Reynolds, 2004; Temple & Reynolds, 1999). Therefore, it is probable that the worst performers in immersion programs, which can also be some of the lower performing students in general, leave the school district or the program itself. This can potentially leave a greater concentration of higher SES students or better achieving students in the immersion classroom, which tends to experience smaller class sizes with increasing grades (Arthur, 2004; Padilla, Fan, Xu, & Amado, 2013).

At the same time, the traditional classrooms can experience overcrowding and effectively leave teachers with less time to spend on individual children,

which also happen to belong to a lower social class or experience more learning difficulties. Because immersion education does not happen in a laboratory, it is difficult to control for the many confounding variables. Because of the mantra that immersion students tend to experience depressed performance early and later demonstrate equal achievement rates, some schools have instituted special programs to make sure immersion students do not fall behind. This was the case in a Maryland immersion program (Essama, 2007). However, expectations of lower performance from other Maryland lower-class students not in the immersion program might not have been met with a similarly intense effort to ensure academic success (Rist, 1970).

Although this analysis found a small, positive effect for immersion education relative to traditional classrooms, it is also likely that there can be a true null effect. Even if this is the case and there is no difference in the academic performance of students in immersion programs, then immersion education proves ultimately beneficial where students can perform on par with their peers. The successful immerson students perform in two languages whereas their peers in traditional classrooms can only achieve in one language.

## Limitations and Future Research Directions

As a meta-analysis, caution must be exercised in identifying wide implications due to this study's methodological limitations. Meta-analyses typically involve multiple coders whereas the articles used in this study were coded solely by the author. Future studies should employ multiple coders and present the interrater reliability of those coders. This is important because instability in the coding and determinations in the inclusion criteria may yield different results. Second, this meta-analysis only examines ten studies which met the inclusion criteria. As indicated by the failsafe *k* analysis, it would not take many studies to negate the small positive effect that was found. Although the inclusion criteria were focused on including studies from which an effect size could be determined from immersion programs, unsuccessful programs are unlikely to have been reported in the academic literature.

Although efforts were made to examine the literature for articles and dissertations, it is likely that studies that met the inclusion criteria were missed because of the multiple terms that are utilized in the field. Further research would include a more exhaustive literature review and should follow best practices (Moher, Liberati, Tetzlaff, & Altman, 2009) by including attempts to contact the eminent researchers in the field for datasets, file drawer articles, and leads to other potential sources for effect sizes to be determined. Because of the data reported in newsletter articles from the American Council on Immersion Education, multiple datasets should be available that could yield enough information for effect size calculation. It is important to note, however, that

utilizing unpublished work is inherently complicated, as increasing the amount of studies for meta-analysis may contrarily decrease the overall quality of the corpus and the validity of its conclusions—effectively "garbage in, garbage out" (Borenstein et al., 2009; Sharpe, 1997). There may be many legitimate reasons that unpublished manuscripts have not passed peer review.

In addition to the inclusion of more studies, another consideration for future research is the type of meta-analysis performed. Although briefly discussed previously, there are multiple techniques for conducting meta-analyses and some of those techniques are outdated or used principally in other fields. For example, vote counting was one of the original meta-analytical approaches (Borenstein et al., 2009). In this approach, significant results were simply tallied based on the direction of the *t* test, otherwise known as the sign test. Likewise, non-significant results were also considered in the tally. Because *p* values depend both on the size of the effect and the size of the sample, it is plausible that many studies with low levels of practical significance are discarded. Other approaches to meta-analysis failed to consider the weight of the population; hence, a few studies with many participants can overshadow the effects seen by multiple contradictory studies with lower *n*.

While the approach used in this analysis follows the steps outlined in Borenstein et al. (2009) and their parameters for heterogeneity of effect sizes and the inverse-variance method, other meta-analytic approaches are utilized in other fields. Schmidt and Hunter (2015) illustrated that within industrial/organizational (I/O) psychology's two principle journals, there were four different meta-analytical methods utilized. Within I/O psychology, though, the Hunter-Schmidt method was used in about 80% of publications. Schmidt and Hunter's (2015) approach to meta-analysis employs a series of artifact corrections based on the characteristics of the studies involved. In this way, they attempt to correct for bias in the effect sizes that are reported in the literature. Similarly, Borenstein et al. (2009) use Hedges' *g* instead of Cohen's *d* in order to account for the effect of population size in the effect size when entered into the analysis. This is an example of an artifact correction for bias because studies with fewer participants tend to produce larger effect sizes. Therefore, the Hunter-Schmidt method attempts to account for all bias inherent in the studies (samples) in the analysis. Future investigations may find different results utilizing the Hunter-Schmidt method than the Hedges-Olkin method outlined by Borenstein et al. (2009).

The final major limitation of this meta-analysis and the greatest potential for clarification in future research is the use of randomized data. Because the studies examined in this meta-analysis principally involved non-randomized students, the potential for bias from unequal groups limits the generalizability of these results. Parents often self-select to place students in these programs. As Williams (2017) explained, immersion programs are seen as a commodity to raise the cultural capital of middle-class students, and initial two-way immersion

programs designed to both encourage achievement in ELLs sometimes evolve into one-way programs as neighborhood demographics change with this highly sought-after educational program. Even though full lottery entry into these programs helps to mitigate initial unequal group differences, sociocultural factors tied to student transience will continue to affect academic achievement as well as continuing student participation in these programs. There is often a window of two years in which students are permitted to matriculate into an immersion program, and immersion class sizes tend to decrease with each consecutive grade level. Lastly, immersion programs do not operate in a bubble and educational leadership must always weigh the effects of immersion programs, whether neutral or directional, against the best interests of each building and district given the needs of the student population.

**Works Cited**

Arthur, G. (2004). "Partial Spanish Immersion Program Expands by Offering Academic Excellence for All." *The ACIE* Newsletter 7.2. Web.

Artzer, M. E. (1990). *The Effect of an Early Partial Immersion Program in Foreign Language on the Achievement and Academic Self-concept of Students in Grades Seven and Eight.* Dissertation, Miami University. Web.

Baker, K. (1987). "Comment on Willig's 'A meta-analysis of selected studies in the effectiveness of bilingual education.'" *Review of Educational Research* 57.3: 351-362.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-analysis*. West Sussex, UK: Wiley. Print.

Bournot-Trites M., & Reeder, K. (2001). "Interdependence Revisited: Mathematics Achievement in an Intensified French Immersion Program." *The Canadian Modern Language Review/La Revue Canadienne des Langues Vivants* 58.1: 27-43.

Center for Applied Linguistics. (2011). *Directory of Foreign Language Immersion Programs in U.S. Schools.* Web.

Corwin, R. G. (1983). "A Fail-safe N for Effect Size in Meta-analysis." *Journal of Educational Statistics* 8: 157-159.

Cummins, J. (1977). "Cognitive Factors Associated with the Attainment of Intermediate Levels of Bilingual Skills." *The Modern Language Journal* 61.1-2: 3-12.

Cummins, J. (1979). "Linguistic Interdependence and the Educational Development of Bilingual Children." *Review of Educational Research* 49: 222-251.

Dahlke, J. A., & Wiernik, B. M. (in press). "Psychmeta: An R package for Psychometric Meta-analysis." *Applied Psychological Measurement.*

Diaz, R. M. (1983). "Thought and Two Languages: The Impact of Bilingualism on Cognitive Development." *Review of Research in Education* 10.1: 23-54.

Essama, L. (2007). "Total Immersion Programs: Assessment Data Demonstrate Achievement in Reading and Math." *The ACIE Newsletter* 11. Web.

Fortune, T. W., & Song, W. (2016). "Academic Achievement and Language Development in Early Total Mandarin Immersion Education." *Journal of Immersion and Content-Based Language Education* 4.2: 168-197.

Felton, T. F. (1999). "Sink or Swim? The State of Bilingual Education in the Wake of California Proposition 227." *California University Law Review* 48: 843-880.

Greene VonCannon, S. (2015). *A Study of a Spanish Immersion Program and its Impact on the Academic Achievement of First Grade Students.* Dissertation, Wingate University. Web.

Haj-Broussard, M. (2005). "Comparison Contexts: African-American Students, Immersion, and Achievement." *The ACIE Newsletter* 8.3. Web.

Hall, C. J., Smith, P. H., & Wicaksono, R. (2011). *Mapping Applied Linguistics: A Guide for Students and Practitioners.* New York, NY: Routledge.

Hill, S. R. & Otani, H. (2014). *The Academic Effects of an Elementary Chinese Immersion Program*. Unpublished thesis. Central Michigan University.

Hopkinson, A. (2017, January 6). "A New Era for Bilingual Education: Explaining California's Proposition 58." *EdSource.* Web.

Jacobson, S. (2013). *A Comprehensive Evaluation of a K-5 Chinese Language Immersion Program*. Dissertation, Gardner-Webb University. Web.

Jones, C. T. (2005). "Spanish Immersion and the Academic Success of Alamo Heights Students." *The ACIE Newsletter* 9.1. Web.

Kennedy, B., & Medina, J. (2017, September). *Dual Language Education: Answers to Questions from the Field*. Washington, D.C.: Center for Applied Linguistics. Web.

Kirkici, B. (2004). "Foreign Language-medium Instruction and Bilingualism: The Analysis of a Myth." *Sosyal Bilimler Dergisi* 2: 109-122.

Lee, A. (2018, January 17). "Parents, Teachers Praise Utah's Dual Language Immersion." *The Daily Universe*. Web.

Lo, Y. Y., & Lo, E. S. C. (2014). "A Meta-analysis of the Effectiveness of English-medium Education in Hong Kong." *Review of Educational Research* 84.1: 47-73.

Macnamara, J. (1966). *Bilingualism and Primary Education.* Edinburgh, Scotland: Edinburgh University Press.

Marian, V., Shook, A., Schroeder, S. R. (2013). "Bilingual Two-way Immersion Programs Benefit Academic Achievement." *Bilingual Research Journal* 36.2: 167-186.

Mehana, M., & Reynolds, A. J. (2004). "School Mobility and Achievement: A Meta-analysis." *Children and Youth Services Review* 26.1: 93-119.

Moher D., Liberati A., Tetzlaff J., Altman D. G., The PRISMA Group (2009). "Preferred Reporting Items for Systematic Reviews and Meta-analyses: The PRISMA Statement." *PLoS Med* 6.7. Web.

Mukai, A., Downes, S. M., & Sato, J. (2005). "Academic Achievement of English-speaking Students in a Japanese Immersion Program." *Educ. Technol. Res.* 28: 53-58.

Office of English Language Acquisition. (2015). *Dual Language Education Programs: Current State Policies and Practices.* Washington, D.C.: U.S. Dept. of Education. Web. 1

Padilla, A. M., Fan, L., Xu, X., & Silva, D. (2013). "A Mandarin/English Two-way Immersion Program: Language Proficiency and Academic Achievement." *Foreign Language Annals* 46.4: 661-679.

Rist, R. (1970). "Student Social Class and Teacher Expectations: The Self-fulfilling Prophecy in Ghetto Education." *Harvard Educational Review* 40.3: 411-451.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of Meta-analysis: Correcting Error and Bias in Research Findings. 3rd ed.* Thousand Oaks, CA: Sage.

Sharpe, D. (1997). "Of Apples and Oranges, File Drawers and Garbage: Why Validity Issues in Meta-analysis Will Not Go Away." *Clinical Psychology Review* 17: 881-901.

Strickland, K. & Hickey, T. M. (2016). "Using a National Dataset to Explore Sub-groups in Irish Immersion Education." *Journal of Immersion and Content-Based Language Education* 4.1: 3-32.

Swain, M., & Lapkin, S. (1982). *Evaluating Bilingual Education: A Canadian Case Study*. Avon, England: Multilingual Matters.

Tedick, D. J. & Wesely, P. M. (2015). "A Review of Research on Content-based Foreign/Second Language Education in US K-12 Contexts." *Language, Culture and Curriculum* 28.1: 25-40.

Temple, J. A., Reynolds, A. J. (1999). "School Mobility and Achievement: Longitudinal Findings from an Urban Cohort." *Journal of School Psychology* 37.4: 355-377.

Thomas, W. P., Collier, V. P., & Abbott, M. (1993). "Academic Achievement through Japanese, Spanish, or French: The First Two Years of Partial Immersion." *The Modern Language Journal* 77.2: 170-179.

Viechtbauer, W. (n.d.). "Outlier and Influential Case Diagnostics for 'rma.uni' Objects. Metafor-Project." Retrieved from http://127.0.0.1:20161/library/metafor/html/influence.rma.uni. html.

Watzinger-Tharp, J., Swenson, K., & Mayne, Z. (2016). "Academic Achievement of Students in Dual Language Immersion." *International Journal of Bilingual Education and Bilingualism* 21.8: 913-928. Web.

Williams, C. (2017, December 28). "The Intrusion of White Families into Bilingual Schools." *The Atlantic*. Web.

Willig, A. C. (1985). "A Meta-analysis of Selected Studies on the Effectiveness of Bilingual Education." *Review of Educational Research* 55.3: 269-317.