Educational Research Review 27 (2019) xxx–xxx

Contents lists available at ScienceDirect

Educational Research Review

journal homepage: www.elsevier.com/locate/edurev

Review

A systematic review on evidences supporting quality indicators of bilingual, plurilingual and multilingual programs in higher education

Fernando D. Rubio-Alcalá[a,*], José Luis Arco-Tirado[b], Francisco D. Fernández-Martín[b], Rocío López-Lechuga[c], Elvira Barrios[d], Víctor Pavón-Vázquez[e]

[a] English Language Department, Avda. Tres de Marzo, s/n, 21071, Huelva, Spain
[b] Department of Psychology and Education, Facultad de Ciencias de la Educación, Campus de Cartuja, s/n, Avda. Pulianas, 18071, Granada, Spain
[c] Department of Developmental Psychology, Facultad de Ciencias de la Educación, Campus de "El Carmen", Avda. de las Fuerzas Armadas s/n, 21007, Huelva, Spain
[d] Department of Didactics of Languages, Arts and Sports, Facultad de Ciencias de la Educación, Campus de Teatinos, Bulevar Louis Pasteur s/n, 29071, Málaga, Spain
[e] Dpto. Filologías Inglesa y Alemana, Facultad de Filosofía y Letras, Plaza del Cardenal Salazar 3, 14071, Córdoba, Spain

ARTICLE INFO

Keywords:
Bilingual programs
Higher education
Quality indicators
Evidence-based practices
Systematic review

ABSTRACT

This systematic review intends to report on the strength of evidences supporting the quality indicators (predictors) attributed to higher education bilingual, plurilingual or multilingual practices and programs across four key dependent variables (outcomes) analyzed (i.e., student performance, second language proficiency, employment, and motivation and attitudes). The rapid growth of both offer and demand of this type of higher education and learning worldwide requires the implementation of high-quality evaluation strategies and techniques to measure potential causal links between interventions and results. To do so, a pre-specified systematic review protocol following the Campbell Collaboration (2015) recommendations is designed and implemented. The results suggest the urgent need to increase the primary research quality standards in this sub-discipline by reducing bias in the processes of designing, implementing and reporting research. Despite the scarcity of results sustained on statistical conclusions with the higher statistical power found in this review, specific results of the dependent variables indicate that this type of education benefits students' performance and second language proficiency, with a higher impact on receptive skills. Although no results were obtained concerning student employment, other results point out that there is general satisfaction of participation with the programs. Finally, several recommendations on how to scale up those quality research standards in this sub-discipline are provided.

1. Introduction

Internationalization, globalization, *Englishization*, and other marketization forces have spurred, all around the world, the delivery of bilingual, plurilingual or multilingual programs at all educational levels, including Higher Education (HE). At this point, it has to

* Corresponding author. English Language Department, University of Huelva, Avda. Tres de Marzo, s/n, 21071, Huelva, Spain.
E-mail addresses: fernando.rubio@dfing.uhu.es (F.D. Rubio-Alcalá), jlarco@ugr.es (J.L. Arco-Tirado), fdfernan@ugr.es (F.D. Fernández-Martín), rocio.lopez@dpsi.uhu.es (R. López-Lechuga), elvira.barrios@uma.es (E. Barrios), victor.pavon@uco.es (V. Pavón-Vázquez).

https://doi.org/10.1016/j.edurev.2019.03.003
Received 2 August 2018; Received in revised form 4 December 2018; Accepted 6 March 2019
Available online 10 March 2019
1747-938X/

be said that we differentiate between 'multilingual', the existence of several languages, and 'plurilingual', where the emphasis when using languages is on the cultural dimension, as defined by the Common European Framework of Languages (Council of Europe, 2001). In this context, an increasing number of universities offer undergraduate and postgraduate programs through the medium of English (Lasagabaster, Doiz, & Sierra, 2014) and other languages. Additionally, the number of students enrolled outside their country of citizenship has increased enormously over the past three decades, from .8 million worldwide in 1975 to 4.6 million in 2015, with inflows towards European countries and the United States increasing by 5.0% and 7.5% respectively (Organization for Economic and Cooperation Development, 2017). In this vein, Wächter and Maiworm's (2014) study reveals 239% growth in Bachelor and Master programs over a 7-year period: from 2,389 in 2007 to 8,089 in 2014.

This rapidly emerging phenomenon, particularly in non Anglo-Saxon countries, has led to a new educational paradigm under diversity of terms like bilingual degree programs, bilingual or plurilingual learning, or bilingual MOOCs (Arco-Tirado et al., 2018). In this regard, exploratory research shows that variability found in bilingual, plurilingual or multilingual practices and programs (BPMPPs) ranges from those associated to the label English as a Medium of Instruction (EMI), which basically entails the delivery of instruction in English, to those within the framework of other approaches such as the Integrated Content and Learning in Higher Education (ICLHE), which is a variation of the form of bilingual education known as Content and Language Integrated Learning (CLIL), which has developed in compulsory education (Arco-Tirado, Fernández-Martín, & Hernández-Moreno, 2016). Concurrently, this variability in the use of these terms could arguably be attributed to the absence of a commonly shared definition of the concept, as it is discussed in the Council of Europe's (2007) document entitled "From linguistic diversity to plurilingual education: Guide for the development of language education policies in Europe". Indeed, this document entails a comprehensive effort to justify and frame the development of plurilingual education policies by emphasizing consensus around linguistic, sociological or economic arguments rather than theoretical and/or empirical evidence from evaluation research.

As far as curricular development is concerned, BPMPPs encompass the same curricular components as monolinguals ones, plus a few additional elements stemming from the use of more than one language as a means of instruction, which adds extraordinary complexity from the teaching, learning and research perspective. In this regard, whereas some authors like Marsh, Pavon, and Frigols (2013) identify several levers conditioning BPMPPs' quality at the macro level (university language policy, program objectives, program language plan, English language fluency, staff incentives, role of language specialists, linking program to research, technologies for learning, student intake, voluntary involvement of teaching staff, coordinated staff dialogue, English language communication objectives, learning success benchmarking, concept formation, English language program input, plagiarism management, program support staff, international networking, cooperation and publishing, cooperative ventures, interactional methodologies, conceptual scaffolding, quality assurance and accreditation, digitized learning environments, social media, studio and virtual environments); other authors like Soltero and Ortiz (2012) enumerate a similar set of strands accompanied each by a set of specific principles (assessment and accountability, curriculum, instruction, quality of staff, facilitators, and administrators and professional development, program structure, community partnerships, support and resources). Yet other authors, like Short (2006), identify the following components for effective bilingual content lessons at the micro level: (a) lesson planning (learning objectives –language and contents–, instructional adjustments –materials, atmosphere, teaching functional language, activity plan, support strategies like peer learning or mentoring or tutoring activities, native language support–, and assessment adjustments); and (b) student academic behaviors (e.g., engagement, verbal interactions).

All in all, these components and their potential interactions are pointed as determinants of the quality and effectiveness of BPMPPs and therefore represent the research targets for evaluation studies. These studies intend, consequently, to identify the extent to which those "high-quality" components embedded in these programs are supported by credible evidence. In this vein, although all those key components are justifiable from a theoretical perspective, the discussions are not settled when it comes to arguing their importance or contribution to BPMPPs' effectiveness from an empirical evaluation standpoint.

In this context, regardless of the type of bilingualism, plurilingualism or multilingualism adopted, from the quality of teaching and learning perspective, the university faculty has been left in front of basic educational and instructional decisions around those key dimensions of the curriculum for which no systematic reviews on evidence-based practices or very little on practice-based evidence were available (Arco-Tirado et al., 2016). For example, while authors like Dafouz, Camacho, and Urquia (2014) suggest that plurilingual education programs (which make use of some native language instruction) do not significantly differ from English-only programs in their impact on standardized test performance, other authors like Arco-Tirado et al. (2018) provide empirical evidence, using a counterfactual impact evaluation design, that there is a cost for bilingual students in academic performance compared to their monolingual counterparts. Therefore, in this paradoxical and pressing context of delivering high-quality educational practices without enough evaluation research data and results, faculty and staff have had to turn their research efforts toward those studies that allow them to summarize intervention effects from BPMPPs accurately and reliably (Arco-Tirado & Fernández-Martín, 2018).

Systematic reviews are widely believed to provide the best evidence to inform decision-making (Stewart, Moher, & Shekelle, 2012). In this study we adopt the definition of systematic reviews provided by The Cochrane Collaboration, an international and independent not-for-profit organization aimed at making up-to-date, accurate information about the effects of BPMPPs in HE. That is, a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies included in the review (Higgins. & Green, 2011).

Evidence-based practice is achieving widespread recognition across disciplines and policies (Coalition for Evidence-Based Policy, 2012; Commission on Evidence-Based Policymaking, 2017). The movement on evidence-based practice originated in the medical sciences in the late 1960s and early 1970s (Cochrane, 1972), and has since taken root in other scientific fields, such as occupational therapy (Ottenbacher & Maas, 1999), management (Rousseau, 2006), social work (Bellamy, Bledsoe, & Traube, 2006), criminal justice (Mears & Barnes, 2010), and education (Buskist & Groccia, 2011; Slavin, 2008a). In medicine, for example, evidence-based is

defined as the "conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996, p. 71). Basically, the process of establishing scientific evidence starts with different studies on a particular educational practice (or problem) yielding different results, followed by the use of the "best evidences" or results to claim such practice (including its components) as evidence-based practice. By "best evidence" we refer to those results sustained on statistical conclusions with the higher statistical power, i.e., those most successful in avoiding Type I error (concluding that a treatment has an effect when it does not) or Type II error (failing to detect the effect of a treatment); the latter being a major threat to the statistical conclusion validity of educational research studies (Arco-Tirado et al., 2018), although other factors threatening internal and ecological validity are equally important. Ideally, the series of studies results in an empirical consensus regarding the effectiveness of a program or practice, which is essential to establish guidelines for evidence-based practice (Ottenbacher & Maas, 1999). In the education field, the term evidence-based practice or program is used to describe proven programs or practices that present evidence that students who use them will learn more than other students who so not (Slavin, 2008b).

The objective of this study is, therefore, to gather, summarize and integrate the quality of empirical evidences supporting causal links between quality indicators or components of BPMPPs in HE and results through the implementation of a pre-specify research plan or protocol based on international high-quality standards for systematic reviews (Brunton et al., 2011; The Campbell Collaboration, 2015). We intend to respond to the following research questions:

1. Are the current evidences supporting quality indicators or components of BPMPPs in HE reliable? And if not, what are the main sources of bias limiting that reliability?
2. Can the current quality level of evidence inform future evidence-based policies? And if not, what changes should be introduced in the process of establishing and using evidences should be implemented to advance this movement in BPMPPs?

2. Method

2.1. Inclusion and exclusion criteria

The review team adapted a systematic review protocol from Campbell Systematic Reviews: Policies and Guidelines (The Campbell Collaboration, 2015). The access to this protocol (Arco-Tirado & Fernández-Martín, 2018) is possible under request to the contact author.

The eligibility criteria were defined in relation to the objectives of the systematic review. First, the operational characteristics of independent (predictor) and dependent (outcome) variable(s) were established. The current variety of practices and programs (independent variables-predictors) across countries has been operationally defined as follows: (a) bilingual education practices and programs refer to education in which two (or sometimes more) languages are used as medium of instruction (Council of Europe, 2007); dual language practices and programs also refer to this, and the term is used in the North-American context; (b) plurilingual education practices and programs refer to a manner of teaching, not necessarily restricted to language teaching, which aims to raise awareness of the language repertoire of each individual, to emphasize its worth and to extend this repertoire by teaching lesser used or unfamiliar languages (Council of Europe, 2007); and finally, (c) multilingual education practices and programs is used to describe the situation in a geographical area where several languages coexist; speakers in this geographical area may not be proficient in each of the different varieties represented (Council of Europe, 2007).

Similarly, the operational definition of dependent variables-outcomes were: (a) student performance: academic performance (i.e., Grade Point Average –GPA, achievement test scores on bilingual courses, attendance, dropout, retention, repetition and graduation), second language proficiency (i.e., English language proficiency, academic English proficiency), and employment (i.e., employment rate), measured through standardized and/or objective quantitative procedures, usually a questionnaire, official reports, a structured interview, or language or content tasks (e.g., reading, listening, speaking, and writing); and (b) student attitude or motivation, all of them defined as students' perception and opinion after their participation on bilingual, plurilingual or multilingual practices and programs, measured through quantitative and/or qualitative procedures such as questionnaires, focus groups, and/or structured or semi-structured interviews.

Second, the eligible research designs for this research study were, following the classification of Campbell and Stanley (1963), Pre-experimental, Quasi-experimental, Experimental, Correlational, and Ex Post Facto.

Third, eligible participants were undergraduate students from HE, universities, and college institutions.

Fourth, no time restriction was applied to this study.

Fifth, no geographical and/or cultural restrictions were included. The following publication languages were eligible: English and Spanish.

2.2. Search strategies

The search aims to arrive at a comprehensive and unbiased set of relevant studies. To this end, the review team systematically tested and screened potentially relevant sources to identify pertinent sources and develop customized search strategies. Proquest, Web of Science, and Scopus were selected as search engines for different databases.

The search for relevant literature was based on a variety of sources in order to ensure that published and unpublished studies ("grey literature") relevant to the review question are included in the search process. Thus, the search process included a primary search, searching of electronic platforms and databases, and a complementary search, searching other resources and hand searching

of relevant websites, literature snowballing, and contacting experts. Through this comprehensive search process, the review sought to arrive at a comprehensive and unbiased set of studies. This search was conducted during October 2016.

Primary search was performed using Proquest, Web of Science, and Scopus in three different universities (i.e., University of Granada, University of Málaga and University of Huelva), in order to strengthen representativeness and reliability of data. The primary search of electronic platforms and databases included can be found at Appendix A.

The primary search was supplemented by a complementary search to comprise further studies for inclusion. The complementary search included searching other resources, hand searching of relevant websites and associations, literature snowballing and contacting experts (see Appendix B for details).

The search (terms) strategy was modified according to the specifications of each electronic platform and database. Also, when appropriate, synonyms were used. The search terms reflect the inclusion criteria defined above and try to strike a balance between sensitivity (i.e., finding all articles in a topic area) and specificity (i.e., finding only relevant articles).

For electronic platforms and databases with advanced search functions, we classified search terms according to three categories (independent variables-predictors, dependent variables-outcomes, and participant population), which were combined using the Boolean operator "AND" to identify potentially relevant studies in each electronic platform and database in title, abstract and keywords. To ensure inclusion of papers that do not specifically report their research design or geographical/cultural restriction in their title or abstract, the search excluded methodology and geographical or cultural restrictions terms.

For websites or databases with basic search functions, the review team adjusted the search terms due to limited functionality of search functions. The preferred search strategies were based on keyword searches and/or topic/theme searches. For databases/websites that do not allow the combination of keywords, separate keyword searches were conducted for the terms (see Appendix C for full search terms).

The review team used Refworks to manage and document the process. The software allows decision tracking for each identified citation throughout the search. Bibliographic information of studies from electronic platforms and databases was imported into Refworks as well as databases with compatible formats.

To enable transparency and reproducibility, the review team kept records of the search process. The search log includes the database, the database interface, the type of database, the customized search strategy, the language of search terms, the search string, the number of records obtained, the date of search and the initials of the researcher.

Three screening levels were conducted to complete the selection process: (a) the first screening level was aimed at identifying and removing duplicate registers and studies which, based on their titles, were clearly related to other fields or topics; (b) the second screening level involved identifying and removing those studies that, after further examination of the title and abstract, did not meet the remaining inclusion criteria (independent variables-predictors, dependent variables-outcomes, and participant population); and (c) at the third screening level the full text versions of the studies were read to ascertain eligibility based on both the inclusion criteria and exclusion criteria, such as the language of publication (e.g., Chinese), type of publication (e.g., book reviews or theoretical studies), studies not based on bilingual, plurilingual or multilingual practices and programs, and/or studies that did not provide data on student performance, attitude, or motivation. Based on the review's inclusion and exclusion criteria, discrepancies were resolved by further review of the respective titles, abstracts and full text, and discussion by the review team.

From the selected sample of studies data and information were coded on variables related to: (a) study methods (i.e., sampling technique and procedure, response rate/attrition, representativeness, instruments, research design, data analysis, and bias); (b) independent-predictor variables (i.e., bilingual, plurilingual or multilingual practices and programs); (c) outcome variables (i.e., students' performance, attitude, or motivation); (d) characteristics of the subject samples of analysis (i.e., sample size, mean and range age, and gender); (e) contextual features (i.e., reference and country); and (f) results and conclusions.

The approach adopted for the data analysis and reporting was a narrative content analysis (Dochy, 2006). This decision was based on Petticrew and Roberts' (2006) recommendation not to undertake a meta-analysis when studies are too heterogeneous in terms of study designs or the set of dependent variables or outcomes analyzed. Other authors like Garg, Hackman, and Tonelli (2008) align with this idea and suggest that when the primary studies differ in the design, populations studied, interventions and comparisons used, outcomes measured, etc., it is "appropriate for the review team simply to report the results descriptively using text and tables" (p. 257).

3. Results

The overall search and screening process is depicted in Fig. 1.

In terms of total sample, 202,685 participants were examined in these studies (M = 3,118.23). The sample size ranged from 1 participant to 191,948 participants. Gender distribution varies among studies, with 26 studies including samples composed by males and females, 2 including only female participants, and 37 studies that do not mention these data. The studies selected come from 21 countries: 23 from Spain, 7 from Turkey, 6 from Taiwan, 5 from Korea, 3 from China and Sweden, 2 from Bangladesh and Greece, and 1 from Botswana, Bulgaria, Germany, Italy, Japan, Lithuania, Qatar, Rwanda, Serbia, Switzerland, Ukraine, United Arab Emirates, Spain and Belgium, and Spain and Japan. The studies were published between 1995 (N = 1) and 2016 (N = 6) (i.e., 2000 = 1, 2006 = 1, 2007 = 1, 2008 = 3, 2009 = 1, 2010 = 2, 2011 = 5, 2012 = 4, 2013 = 13, 2014 = 13, 2015 = 14). The publication languages were English (N = 58) and Spanish (N = 7). 30 studies revolved around EMI, 23 CLIL, 3 multilingual, 3 plurilingual, 3 CLIL and English for Specific Purposes (ESP), 1 English as a Second Language (ESL), 1 EMI and ESP, and 1 English-medium education, according to the information reported by the authors.

Sampling techniques were distributed as follows: 26 studies did not explicitly mention this information but it was inferred from

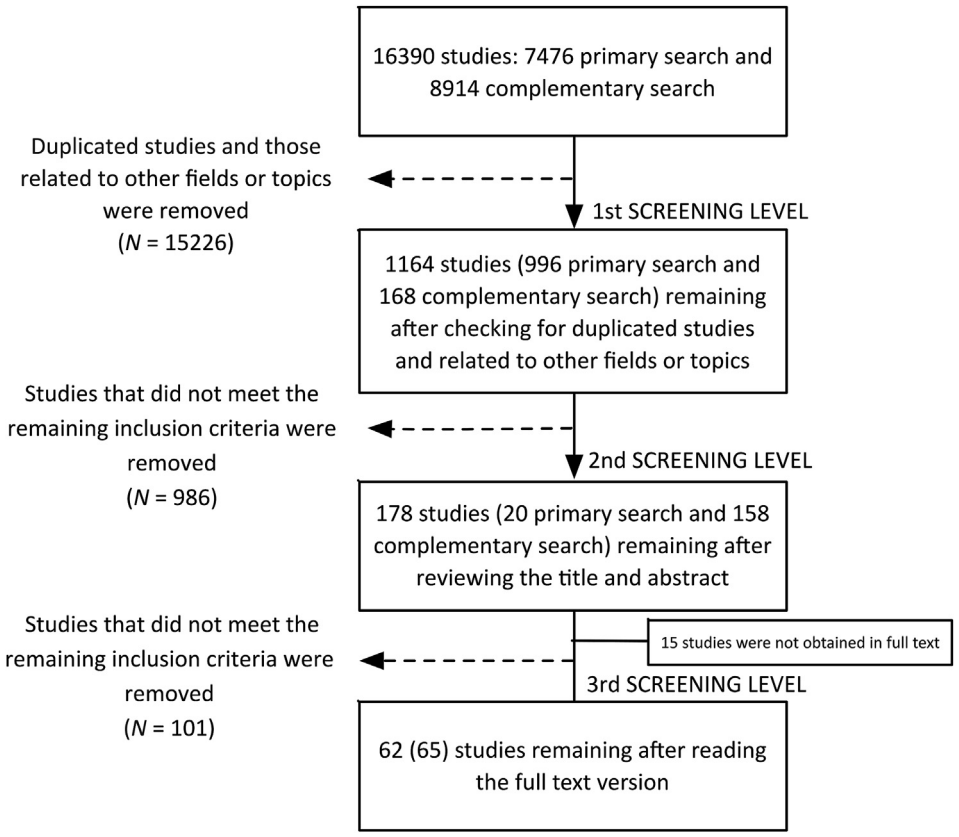

Fig. 1. Flow chart for the literature search and screening.

the text (i.e., non-probabilistic), 2 were unclear, 35 non-probabilistic, and 2 probabilistic. Out of those reviewed studies, 51 belonged to journal publications, 5 were reported in dissertation theses, and 9 came from other sources such as conference proceeding ($N = 5$) or book chapter ($N = 4$). Finally, reporting bias was included in 42 studies (64.61%) and missing in 23 studies (35.39%).

In relation to the research design, 23 studies did not explicitly mention this information, although it was inferred from the text and the conclusion. Thus, our search yielded the following results according to the classification adopted: (a) 3 studies (4.62%) included quasi-experimental designs, with all of them including a non-equivalent control group design; (b) 4 (6.15%) adopted a pre-experimental design, with 2 including one-group pretest-posttest design and 2 adopting one-shot case study design; and (c) 58 (89.23%) adopted an ex post facto design, with 11 studies including static-group, 9 including one-group pretest-posttest, and 38 adopting one-group posttest.

Tables 1 and 2 show the distribution of studies when grouped by outcomes analyzed. For the outcome "student performance", a total of 12 (18.46%) studies were found, with 2 of them including a quasi-experimental design with a non-equivalent control group. Regarding "second language proficiency" a total of 21 (32.31%) studies were found, with 1 (1.53%) study adopting a quasi-experimental design with a non-equivalent control group. No studies focusing on students' employment were found. Finally, in relation to the outcome "student motivation or attitudes" a total of 51 (79.46%) studies were found, with 3 (4.61%) including a quasi-experimental design with a non-equivalent control group.

In terms of "weight" of evidence measure, when the "students' performance" is measured as GPA, two studies show favorable (although non-statistically significant) results for the EMI group (i.e., Dafouz et al., 2014; Hernández-Nanclares & Jiménez-Muñoz, 2015), and one study shows non-significant differences, even after controlling for the covariates "university access grade" and "achievement level" (i.e., Dafouz & Camacho, 2016). Interestingly, if those outcomes are measured through different tasks or instruments, the results change: positive non-statistically significant differences for the EMI group when results are measured through teaching units (Madrid & Madrid, 2015); no differences when measured through lectures on the post-test effect (Joe & Lee, 2013); and negative differences on students' performance for EMI groups with effects size between 0.45 with matching and 0.55 without matching (Vinke, 1995).

In relation to the outcome "second language proficiency", significant differences from CLIL experiences were reported on listening skills (Aguilar & Muñoz, 2014) and receptive linguistic skills in the post-language test (Yang, 2014), although no significant differences were found when compared to other GEPT test-takers. Moreover, no significant differences were reported in grammar, vocabulary and reading by studies using CLIL (Bosisio, 2015) and full immersion (Ament & Pérez-Vidal, 2015); the intervention group outperformed the monolingual group in terms of reading skills and content (Chostelidoua & Grivab, 2014) and the semi immersion

Table 1
Studies classification based on their research design methodological features: Ex post facto and pre-experimental.

Outcomes	One-shot case study			One-group pretest-posttest			Static-group		
	$p > .05$	$p < .05$	Descriptive	$p > .05$	$p < .05$	Descriptive	$p > .05$	$p < .05$	Descriptive
Student performance									
Academic performance									
GPA	25						14, 56	14	24
Achievement test scores			41[c]				4, 33	33	
Content tasks								4	
Attendance						43			
Dropout			40						
Second language performance									
Language level tests	25	1, 30	1	57	27, 37, 57, 58	43	27	4	
Language tasks			8[b], 13, 31[d], 32[d], 47	15, 52	15, 24, 52	15	33	4, 33	10, 24
Employment									
Student motivation or attitude									
Positive perceptions	1, 3, 26, 28, 30	1, 2, 5, 23, 26, 28, 30, 34, 54	1, 2, 3, 5, 6, 7, 16, 17, 18, 19, 20, 21, 26, 28, 35, 38, 40, 41[a], 41[c], 44, 45, 50, 54, 55, 59, 61, 62	9[d], 37	9[d], 37	9[d], 22, 36[d], 37, 43, 58	42	4, 27, 42, 48, 56	27, 39, 48, 60
Negative perceptions		2, 23, 28, 34, 54	2, 6, 7, 8[a], 12, 18, 19, 26, 28, 29, 35, 38, 40, 41[a], 41[b], 44, 45, 46, 50, 53, 54, 55, 59, 62			36[d], 37	48	4, 27, 48, 56	27, 39, 48, 60

Note. Numbers represent references as listed below.
[a] Study 1.
[b] Study 2.
[c] Study 3.
[d] Pre-experimental design.

Table 2
Studies classification based on their research design methodological features: Quasi-experimental and experimental.

Outcomes	Quasi-experimental: Non-equivalent control group									Experimental
	Control group pretest-posttest		Experimental group pretest- posttest		Inter-groups pretest			Inter-groups posttest		
	$p > .05$	$p < .05$	$p > .05$	$p < .05$	$p > .05$	$p < .05$	Descriptive	$p > .05$	$p < .05$	Descriptive
Student performance										
Academic performance										
GPA										
Achievement test scores					49			51	49	
Content tasks										
Attendance										
Dropout										
Second language performance										
Language level tests										
Language tasks		11		11	11				11	
Employment										
Student motivation or attitude										
Positive perceptions								49	49	11, 51
Negative perceptions								49	49	51

Note. Numbers represent references as listed below.

group in the grammar task (Ament & Pérez-Vidal, 2015), concluding that an integrated content and language (ICLHE) approach is more effective than a solely content based EMI model for university level content courses, if linguistic gains are the desired outcomes of the program. Interestingly, when real gains are measured through objective assessment of learners' skills, their scores on CEFR level are half of those self-reported (Hernández-Nanclares & Jiménez-Muñoz, 2015).

In relation to the "employment" outcome, as mentioned before, no results were found.

Finally, in relation to the dependent variable "student attitude or motivation", the enormous variability of methodological

features found (e.g., researcher-made vs. independent measures, sample size, research design, lack of information on correlations among factors, significance criteria) makes it unfeasible to provide a more detailed account of this outcome (Cheung & Slavin, 2016). Alternatively, two blocks of results are reported. Part one refers to the students' positive perceptions toward BPMPPs' impact in HE, mainly related to level of satisfaction, usefulness of English for their academic and future professional life, and impact on their academic and personal competences (see Appendix D for further details).

Part two refers to the students' negative perceptions or criticism toward BPMPPs' impact on HE, usually due to a limited linguistic competence in the L2, lack of support, poor linguistic and pedagogical practices on the part of the teachers, fear of a potential detriment in the use of the L1, and ideological opposition to these programs (see Appendix D for further details).

Interestingly, in relation to the students' strategies for overcoming language problems, our data reveal the following: use of L1 when making definitions for comprehensibility, asking the teacher during and after lessons, translating, using examples and interacting with other classmates in English, note taking, consulting L1 resources, dictionary use, translation integration of L2 into daily life, use of authentic materials (e.g., music, movies, books) in L2, changing study habits so that they no longer take notes in class, reading sections of work before class, using references/text-books in their L1 to make sense of their English lectures and English course books, translating contents from L2 to L1, preparing for tests by memorizing answers based on L1 and English textbooks).

4. Discussion

This systematic review reports on the strength of the evidences supporting causal inferences between the quality indicators (predictors) or components attributed to HE BPMPPs, and the four key dependent variables (outcomes) analyzed (i.e., student performance, second language proficiency, employment, and motivation and attitudes).

In relation to the first part of the research question one on the reliability level of the evidences supporting quality indicators or components of BPMPPs in HE, considering that a Randomized Control Trial (RCT), if appropriately designed, conducted, and reported, is the ideal way to study the net effects of educational practices and programs (Slavin, 2008b), the skewed distribution of studies sampled showed in Tables 1 and 2 demonstrates the significant scarcity of high-quality and therefore credible research and evaluation designs endorsing the current selection of quality indicators for BPMPPs, measured against the four outcomes analyzed. Whilst an RCT evaluation approach is probably not applicable to estimate the contribution of all indicators for ethical reasons (e.g., plurilingual education policy or teachers' preparation), our findings suggest the insufficiency of the evidence available to either support or reject the prospective association of any of the indicators (predictors) with the consequential outcomes analyzed (e.g., students motivation or attitude, L2 proficiency and progress). In other words, the number of studies based on robust evaluation research designs and statistical analysis is clearly insufficient to demonstrate from a causal perspective which programs or practices lead to better outcomes. Therefore, this invites to take both lists as starting points for future research in order to focus on their efficacy and efficiency towards promised outcomes.

These results on the current state of research in this sub-discipline align with those obtained by Macaro, Curle, Pun, An, and Dearden (2018), and Kremer and Valcke (2014), and clearly unveil the need to strengthen the quality of the evaluation research designs and, as Arco-Tirado et al. (2018) indicated, the need to carry out more impact evaluation studies in this sub-discipline at both the macro (e.g., using counterfactual impact evaluation), and micro levels (e.g., more research designs using equivalent control groups). Furthermore, answering the second part of the research question one, our results suggest that limitations in research design, measurement, implementation, and reported statistical analyses are the main sources of failure in demonstrating potential effects in the studies analyzed.

In this regard, for example, a methodological factor explaining the research quality of the studies analyzed could be what Cohen (1988) calls low statistical power. According to this author, the statistical power of a study is defined as its ability to detect a phenomenon of specified magnitude given the existence of that phenomenon. Low-power emerges from an inappropriate use or interpretation of statistical tests or procedures, a form of statistical conclusion invalidity according to Cook and Campbell (1979), which may result in a higher probability of Type II errors in the sample of studies examined, as it was pointed out in the introduction. In practical terms, it means failing to detect that a treatment has an effect when the true treatment effect is nonzero.

In our case, most of the studies failed to provide complete descriptions of that critical information (e.g., the mean differences among treatment and non-treatment groups, the required Type I error probability, experimental-control comparison with evidence of no pre-test differences among the two groups, and/or the sample size), along with other data (including subjects heterogeneity or error of measurement), as Lipsey (1990) and Slavin (2008a) recommends. In this regard, knowing the complexity and uncertainties behind rating the strength of an evidence (e.g., without knowing the anticipated effects for each program), we discarded the idea of justifying or comparing programs effectiveness according to the research base found. In this line, rating the evidences found as foundational and/or exploratory actually means to limit their reliability and validity and, therefore, external validity from the evidence-based perspective. In a similar vein, Garg et al. (2008) point out that, when the primary studies sampled are quite different in terms of the design, populations studied, interventions and comparisons used or outcomes measured, and provide insufficient information on the true effect being estimated, then the decision about what model of "fixed" or "random" effects to use in order to combine the obtained results is compromised. As Borenstein, Hedges, Higgins, and Rothstein (2009) point out, although the number of studies to undertake a meta-analysis depends on the concrete conditions of each meta-analysis, a reasonable minimum number of studies required to adopt a random effects model is about 30. Furthermore, Petticrew and Roberts (2006) argue that, in social science systematic reviews, the studies are sometimes too heterogeneous in terms of study design or the set of dependent variables or outcomes analyzed to permit a statistical summary. Specifically, these authors warn against the possibility of making inappropriate comparisons, especially if the interventions received by the control groups are clearly different between studies (even if the

intervention group is the same), which applies to our case. For example, while in study 51 (Joe & Lee, 2013) the control group (with Korean as L1) received three successive class hours (about 150 min) in a single day in Korean and outcomes were assessed in both languages (Korean and English), in study 49 (Vinke, 1995) the control group (with Dutch as L1) received a single class instruction in Dutch and outcomes were measured in Dutch. Additionally, these authors point out that "meta-analysis should only be applied when a series of studies has been identified for review that addresses an identical conceptual hypothesis", which is not possible to determine under the current reporting conditions of the studies analyzed.

Most of these problems found on BPMPPs in HE have also been reported in health sciences (Schulz, Altman, Moher, & Fergusson, 2010), as well as in recent meta-analyses on the effectiveness of bilingual over submersion programs, though these analyses concerned children in Europe (Reljic, Ferring, & Martin, 2015), and in US public schools (Chin, 2015).

Regarding the first part of research question two on whether the current level of evidence could inform future evidence-based policies, our results show that the current state-of-the-art in BPMPPs is actually putting at risk the development of the relevant statistical consensus to establish guidelines for evidence-based practice (Rosenberg & Donald, 1996) in the field, which in turn prevents the accumulation of reliable evidence to inform future high-quality indicators or components for BPMPPs. Unfortunately, we cannot discuss the reasons underlying the results found from our data; but alternatively, the hypotheses set by other systematic reviews range from implementation procedures, which create more challenges than opportunities for students' and instructors' linguistic academic needs (Williams, 2015, pp. 1–23), to the fairly small number of studies reporting enough information to calculate statistical power (Kremer & Valcke, 2104), to the political will of policy makers and particularly university managers to implement new policies as a direct response to the findings on the current state of research regarding teacher preparation and resourcing that are evidently and urgently needed (Macaro et al., 2018).

In relation to the second part of the second research question, expanding the movement of evidence-based reform in the field of BPMPPs requires changes in the process of preparing, building, mediating and using evidences. In this regard, as Bobrovnikov, Sahni, and Bozzi (2013) argue, although not all projects are ready nor is it feasible to conduct a fully rigorous evaluation, project coordinators, staff, and evaluators who are seeking clear, practical advice on how to report on evaluations they conduct do wish to learn more about the requirements of such an evaluation or to make their evaluation more rigorous. In this line, following the "pipeline" classification suggested by the Institute of Education Sciences and National Science Foundation (2013), most of the studies reviewed correspond to the IES categories of "foundational" (type #1, to advance the frontiers of education and learning; develop and refine theory and methodology; and provide fundamental knowledge about teaching and/or learning), and "early-stage or exploratory research" (type #2, to investigate approaches to education problems to establish the basis for design and development of new interventions or strategies, and/or to provide evidence for whether an established intervention or strategy is ready to be tested in an efficacy study), using bivariate and multivariate analysis yielding significant results. This provides the fundamental knowledge contributing to set the theoretical and methodological bases to inform, guide, support and conduct more causal research studies in the following years, and examines relationships (usually correlational rather than causal) among relevant constructs in learning and education to pave the way to design future interventions in order to improve educational outcomes, respectively. If the field moves in that direction the next levels categories named "design and development" (type #3, to develop new or improved interventions or strategies to achieve well-specified learning goals or objectives, including making refinements on the basis of small-scale testing), "efficacy research" (type #4, to determine whether an intervention or strategy can improve outcomes under what are sometimes called "ideal" conditions), "effectiveness research" (type #5, to estimate the impacts of an intervention or strategy when implemented under conditions of routine practice) and, eventually, "scale-up research" (type #6, to estimate the impacts of an intervention or strategy under conditions of routine practice and across a broad spectrum of populations and settings), will allow researchers, educators and practitioners to distill more effective strategies and interventions on BPMPPs in HE. From a more refined methodological perspective, it is recommended to refocus educational research from the lowest level of evidence, that is, programs with a rationale based on high-quality research or a positive evaluation that are likely to improve student performance or other relevant outcomes and that are undergoing evaluation, to the next level of "promising evidence", meaning at least one correlational study with pre-tests as covariates, to the next level up -"moderate evidence"-, meaning supported by at least one quasi-experimental study, to finally, the "strong evidence" meaning supported by at least one randomized study (Slavin, 2016). For example, whilst identifying well-done syntheses of evaluative studies and integrating individual expertise with external evidence represent two key strategies in the mid-term of this endeavor, improving the quality of reporting of primary studies in BPMPPs arises as a priority in the short-term.

Such process of disseminating and using rigorous, systematic, and objective methodologies to obtain reliable and valid knowledge requires, according to the American Educational Research Association (AERA) (2006): (a) development of a logical, evidence-based chain of reasoning; (b) methods appropriate to the questions posed; (c) observational or experimental designs and instruments that provide reliable and generalizable findings; (d) adequate data and analysis to support findings; (e) clear and detailed explanation of procedures and results, including specification of the population to which the findings can be generalized; (f) adherence to professional norms of peer review; (g) dissemination of findings to contribute to scientific knowledge; and (h) access to data for reanalysis, replication, and the opportunity to build on findings. Additionally, the quality of reporting of primary studies and syntheses in BPMPPs can benefit from the development of new more detailed reporting standards available at the Institute for Education Sciences (US Department of Education) (see www.ies.ed.gov/ncee/wwc), the Best Evidence Encyclopedia (see www.bestevidence.org), the Campbell Collaboration (see www.campbellcollaboration.org), the American Psychological Association (see www.apa.org), or the Evidence for Policy and Practice Information and co-ordinating Centre (EPPI-Centre) (see www.eppi.ioe.ac.uk), to name just a few.

Despite the scarcity of studies based on robust evaluation research designs and statistical analysis found, this review can offer suggestions for administrators, program organizers and teachers to improve BPMPPs. For instance, teachers and students demand

support to improve their language and teaching/learning competences. Language courses can thus be offered to both by the language center of the institution with a focus on academic purposes (Cognitive Academic Language Proficiency –CALP) rather than the traditional BICS (Basic Interpersonal Communicative Skills) approach (Cummins, 1979), so that language has a functional purpose. According to the results, courses for teachers should be focused on pronunciation and oral skills, and should also include methodological training, so that teachers were better prepared to facilitate language comprehensibility by including scaffolding techniques (i.e., pre-reading texts before class, using glossaries, etc.) and discourse adaptations (i.e., paraphrasing, defining, over articulating words, etc.).

Results have also shown that academic performance of students improves and that they are able to access primary sources of information. This provides an opportunity for teachers to implement an active methodology (i.e., student-centred), in which discovery and exploration tasks are deployed for students to further develop heuristic competences.

Finally, since students develop strategies for overcoming language or comprehensibility problems, teachers can enrich their methodology by promoting pre-reading texts before class, or using authentic materials (e.g., watching documentaries).

5. Limitations

Researchers and publishers underreporting practices identified in the results and discussion sections may threaten the desired impact of this systematic review. For example, some studies were classified within the subtypes of the pre-experimental category based on our inferences from the rest of the information reported. For some other studies, information on pre-test and randomization is included, although, paradoxically, the corresponding statistical analyses are missing. Still, other studies include information on pre-test and randomization, but no comparisons on post-test measures are reported. For most of the studies, information was not reported on actual efficacy or impact of the practices or programs, with underreporting data problems affecting effect sizes in particular. This is particularly troublesome as it complicates attempts to demonstrate the extent to which the observed correlation between certain methodological features and effect sizes is present in this field (Cheung & Slavin, 2016).

The search of electronic databases was completed in October 2016, so literature published since then has not been included in the systematic review. Additionally, 'grey literature' published after the cut-off date has not been included. In spite of the efforts made during the search of "grey literature", the "file drawer problem", which refers to the bias introduced into the scientific literature by selective publication of positive results but not negative or non confirmatory results, can be another source of bias for this research, as other authors suggest for this type of studies (Macaro et al., 2018).

6. Conclusions

The development of standards and reviews are building blocks of the Evidence-based educational reform movement. In this research we have tried to summarize how much remains to be done on designing, conducting and reporting effective BPMPPs and the results we have synthesized reveal that there is much to be done. It is particularly urgent to increase the quality of the research designs (e.g., large-scale, randomized, longitudinal evaluations, experimental-control comparison, with evidence of no pre-test differences among the two groups, counterfactual impact evaluations), in order to recommend more confidently effective BPMPPs.

The current significant variation of components of BPMPPs recommends not to use the meta-analysis as a way to derive a more precise estimate of the effects, which unveils the need to use both qualitative and quantitative methods more consistently to better understand how different interventions combining different components affect the development and effectiveness of bilingual, plurilingual or multilingual skills. In this vein, the use of (international) standardized instruments, if available, should be prioritized over local *ad hoc* questionnaires for example, as is also recommended in other fields (e.g., Duckworth & Yeager, 2015).

In response to the first of the research questions posited, our results also show that the field still lacks the capacity to generate a body of studies designed according to the experimental model, as Sloane (2008) ascertains for the field of Education in general. Additionally, it cannot be stated that the majority of the evidences found are reliable from a purely scientific point of view. On the contrary, only a small percentage meets the technical requirements for evidence. However, it is noteworthy that many of these evidences show interesting experiences and disclose relevant areas of analysis. Following this reasoning, synthesizing the current findings from BPMPPs into numerical ratings to make causal inferences requires not only a finer level analysis on how a cause produces an effect, but also a collection of randomized experiments with considerable heterogeneity among the students, treatments, and outcomes across studies. As for the second research questions, the analysis of the evidences reveals that there are important decisions that have to be taken at different levels (organizational, economic, methodological, etc.) in order to implement the indispensable changes that BPMPPs require.

Finally, the implementation of more studies focused on estimating causal effects of BPMPPs will allow the field to move from "good practices" and "practice-based evidence" to "evidence-based practices" and, consequently, to promote more sensible and effective bilingual, plurilingual or multilingual educational policies for students in HE.

Funding source declaration

This work was supported by the Junta de Andalucía-funded Proyecto de Excelencia: "Análisis y Garantía de Calidad de la Educación Superior Plurilingüe en la Educación Superior de Andalucía (AGCEPESA) [grant number P12-SEJ – 1588]. The funding source had no involvement in study design, in the collection, analysis and interpretation of data, in the writing of the report or in the decision to submit the article for publication.

Conflicts of interest

We hereby confirm that there's no financial/personal interest or belief that could affect our objectivity.

Acknowledgement

Thanks are due to the following AGCEPESA members for their participation in the search for relevant literature: Javier Ávila-López, Aurora Carretero-Ramos, Sonia Casal-Madinabeitia, Candela Contero-Urgal, M. Carmen Fonseca-Mora, Manuel Hermosín-Mojeda, M Concepción Julián-de-Vega, M. Carmen Méndez-García, Patricia F. Moore, Jesús Nieto-García, Ana M. Ramos-García, Elena Romero-Alfaro, Francisco Rubio-Cuenca, Mercedes Vélez-Toral and Francisco Zayas-Martínez.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.edurev.2019.03.003.

Appendix A

Electronic Platforms and Databases Included in the Primary Search.

Platforms	Databases
Proquest	ABI/INFORM Complete, ProQuest Accounting & Tax, ARTbibliographies Modern (ABM), Arts and Humanities Full Text, Avery Index to Architectural Periodicals, Banking Information Source, Bibliografía de la Literatura Española (Bibliography of Spanish Literature), Ebrary e-books EconLit, Index Islamicus, International Index to Music Periodicals Full Text, International Index to Performing Arts Full Text, International Pharmaceutical Abstract, Library and Information Science Abstract Medline, MLA International Bibliography, Periodical Archive Online Periodical Index Online, Proquest Aquatic Science Collection, Proquest Computer Science Collection, ProQuest Entrepreneurship, ProQuest Health & Medical Complete, ProQuest Library Science, ProQuest Nursing & Allied Health Source, ProQuest Psychology Journals, ProQuest Social Science Premium Collection (i.e., ASSIA, ERIC, IBSS, LLBA, PAIS International & PAIS Archive, PILOTS, ProQuest Criminal Justice, ProQuest Education Journals, ProQuest Political Science, ProQuest Politics Collection, ProQuest Social Science Journals, ProQuest Sociology, ProQuest Sociology Collection, Social Services Abstracts, Sociological Abstracts, Worldwide Political Science Abstracts), PsycARTICLES, PsycCRITIQUES, PsycEXTRA, PsycINFO, PsycTEST, and RILM Abstracts of Music Literature
Web of Science	Social Science Citation Index, Science Citation Index Expanded, Arts & Humanities Citation Index, Emerging Citation Index, Book Citation Index BIOSIS, Chinese Science Citation Database, Current Contents Connect, Derwent Innovations Index, Korean Journal Database, and SciELO Citation Index
Scopus	

Appendix B

Complementary Search Strategies and Resources.

Search strategies	Resources
Hand searching	Reference lists of included studies and reference lists of relevant reviews were searched
Web search	A general web search was conducted using Google Scholar to identify potential unpublished studies. Advanced search options were used to refine the grey search strategy
Open access (grey literature)	OpenGrey (EAGLE-European Association of Grey Literature Exploitation), GreyNet International-Grey Literature Network Service, National Technical Information Service (NTIS), OpenSIGLE, Directory of Open Access Repositories (OpenDOAR), Open Access Scholarly Information Sourcebook (OASIS), Science Commons, COPAC, and Urbadisc
Ongoing research	OpenGrey (EAGLE-European Association of Grey Literature Exploitation), GreyNet International-Grey Literature Network Service, National Technical Information Service (NTIS), OpenSIGLE, Directory of Open Access Repositories (OpenDOAR), Open Access Scholarly Information Sourcebook (OASIS), Science Commons, COPAC, and Urbadisc
Personal contacts	Personal contacts with national and international researchers were made to identify unpublished reports and on-going studies
Relevant institutions and networks	American Institutes for Research, What Works Clearinghouse, EPPI Centre, Educational Evidence Portal (EPP), IZA World of Labor, Social Science Research Network (SSRN), Center for Applied Linguistics (CAL), National Clearinghouse for English Language Acquisition (NCELA), Office of English Language Acquisition (US Department of Education), National Council of Teachers of English (NCTE), National Association for Bilingual Education (NABE), The Association of Language Testers in Europe (ALTE), European Centre for Modern Languages of the Council of Europe, and European Confederation of Language Centres in Higher Education
Key journals	

Reading Research Quarterly, American Educational Research Journal, Journal of Educational Research, Journal of Adolescent and Adult Literacy, Journal of Educational Psychology, Bilingual Research Journal, Reading and Writing Quarterly, International Journal of Bilingual Education and Bilingualism, Innovation in Language Teaching and Learning, International Journal of Multilingualism, Language and Education, Linguistics and Education, Annual Review of Applied Linguistics, Bilingual Education and Bilingualism, Bilingualism: Language and Cognition, Cine Qua Non-Bilingual Arts Magazine, International Journal of Bilingualism, Linguistic Approaches to Bilingualism, Bilingual Review, and Bilingualism

Appendix C

Search Terms: Independent Variables (Predictors), Dependent Variables (Outcomes), and Participant Population.

Variables	Search terms
Independent variables (Predictors)	
	("bilingual education" OR "bilingualism" OR "biliteracy" OR "multilingual education" OR "multilingualism" OR "plurilingual education" OR "dual language" OR "English medium instruction" OR "EMI" OR "Spanish medium instruction" OR "French medium instruction" OR "German medium instruction" OR "Italian medium instruction" OR "content language integrated learning" OR "CLIL" OR "integrated content learning higher education" OR "ICLHE" OR "English academic purpos*" OR "Spanish academic purpos*" OR "French academic purpos*" OR "German academic purpos*" OR "Italian academic purpos*" OR "EAP" OR "sheltered instruction observation protocol" OR "SIOP" OR "immersion program*" OR "two-way instruction" OR "one-way instruction" OR "English mediated course" OR "Spanish mediated course" OR "French mediated course" OR "German mediated course" OR "Italian mediated course" OR "English content instruction" OR "Spanish content instruction" OR "French content instruction" OR "German content instruction" OR "Italian content instruction" OR "additional language" OR "vehicular language") AND
Dependent variables (Outcomes)	
	("performance" OR "achievement" OR "outcome" OR "grade point average" OR "GPA" OR "dropout" OR "drop-out" OR "retention" OR "repetition" OR "graduation" OR "language proficiency" "academic English proficienc*" OR "academic Spanish proficienc*" OR "academic French proficienc*" OR "academic German proficienc*" OR "academic Italian proficienc*" OR "language skill*" OR "language competenc*" OR "quality" OR "effectiveness" OR "success" OR "employ*" "motivation" OR "attitude") AND
Participant population	
	("university" OR "college" OR "tertiary education" OR "learning higher education" OR "higher education" OR "higher educational language policy" OR "grade" OR "undergraduate*")

Appendix D

Students' Perceptions Towards BPMPPs' Impact on HE.

Variables	Perceptions
Positive perceptions	
	- High satisfaction level with their bilingual learning experience
	- Willingness to repeat the bilingual, plurilingual or multilingual experience
	- Acknowledgement of the importance of English for their future careers and job opportunities and employability
	- Improvement in their English language skills
	- Access to primary sources of information
	- Improvement in their abilities to perform academic tasks in English
	- Benefits in content knowledge for those with higher English proficiency level
	- Doubts on the necessity of ESP in the presence of EMI lessons
	- Benefits of support systems based on mentoring programs on skills and competencies for bilingual mentors and students
	- Improved attitudes toward internationalization and foreign language - learning and use in other contexts
	- Improved perception about their mobility
	- Favorable shifting in their motivational discourse
	- Benefits in the acquisition of disciplinary contents
	- Local integration
	- Cultural awareness
	- Open-mindedness
	- Self-concept
	- Flexibility
	- Global vision
Negative perceptions	
	- Lack of English skills to understand subject contents or limited language skills
	- Disregard for the importance of second language proficiency
	- Defense of content instruction in L1 to improve content performance
	- Detrimental effects on content learning in L1 and classroom performance
	- Increased study load
	- Increased surface learning
	- Limited comprehension and misunderstanding of contents
	- Teachers' low English command in terms of pronunciation and comprehensibility
	- Need to improve bilingual, plurilingual or multilingual curricula and assessment systems including understanding of examination questions
	- Lack of support systems in terms of English language learning before and during lessons
	- Lack of sufficient resources in English
	- Need to "adjust" classroom management practices
	- Teachers' lack of communicative-didactic skills

- Difficulty in understanding disciplinary knowledge –particularly details-
- Reduced number of bilingual, plurilingual or multilingual courses
- Impact of low level of English command on psychological functioning and level of awareness
- Participation concerns due to the presence of international students with higher English language command
- Reduced attention span and frequent attention gaps, concerns with the ideological
- Marginalization impact effects tied to the imposition and/or overuse of English

References

Asterisks (*) indicate studies included in this review.

American Educational Research Association (2006). Standards for reporting on empirical social science research in AERA publications: American educational research association. *Educational Researcher, 35*(6), 33–40.

Arco-Tirado, J. L., & Fernández-Martín, F. D. (2018). *Systematic review protocol example on effective plurilingual Higher Education programs. Unpublished manuscript.* Granada, Spain: Department of Developmental and Educational Psychology, University of Granada.

Arco-Tirado, J. L., Fernández-Martín, F. D., Ramos-García, A. M., Littvay, L., Villoria, J., & Naranjo, J. A. (2018). A counterfactual impact evaluation of a bilingual program on students' grade point average at a Spanish university. *Evaluation and Program Planning, 68*, 81–89. https://doi.org/10.1016/j.evalprogplan.2018.02.013.

Bellamy, J. L., Bledsoe, S. E., & Traube, D. E. (2006). The current state of evidence-based practice in social work: A review of the literature and qualitative analysis of expert interviews. *Journal of Evidence-Based Social Work, 3*(1), 23–48. https://doi.org/10.1300/J394v03n01_02.

Bobrovnikov, E., Sahni, S. D., & Bozzi, L. (2013). *A guide for reporting on evaluations for the US department of education mathematics and science partnerships (MSP).* Cambridge, MA: Abt Associates Inc. Retrieved from http://www.ed-msp.net/public_documents/document/resource/Guide for Reporting on MSP Evaluations.pdf.

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

*Bosisio, N. (2015). CLIL in the Italian university. A long but promising way to go. *Elle, 4*(1), 133–154.

Brunton, G., Green, S., Higgins, J. P. T., Kjeldstrøm, M., Jackson, N., & Oliver, J. S. (2011). Preparing a Cochrane review. In J. P. T. Higgins, & S. Green (Eds.). *Cochrane handbook for systematic reviews of interventions. Version 5.1.0*London, United Kingdom: The Cochrane Collaboration. Retrieved from http://training.cochrane.org/handbook.

Buskist, W., & Groccia, J. E. (2011). Evidence-based teaching: Now and in the future. *New Directions for Teaching and Learning, 128*, 105–111.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston, MA: Houghton Mifflin Company.

Cheung, A., & Slavin, R. E. (2016). Howe methodological features affects effect sizes in education. *Educational Researcher, 45*(5), 283–292.

Chin, A. (2015). Impact of bilingual education on student achievement. *IZA World of Labor, 131*. https://doi.org/10.15185/izawol.131.

Coalition for Evidence-Based Policy (2012). *Rigorous program evaluations on a budget: How low-cost randomized controlled trials are possible in many areas of social policy. Coalition for Evidence-Based Policy.* Retrieved from http://coalition4evidence.org/wp-content/uploads/uploads-dupes-safety/Rigorous-Program- Evaluations-on-a-Budget- March-2012.pdf.

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services.* London, UK: Nuffield Provincial Hospitals Trust.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Commission on Evidence-Based Policymaking (2017). *The promise of evidence-based policymaking. Report of the commission on evidence-based policymaking*Washington, DC: Commission on Evidence-Based Policymaking. Retrieved from https://www.cep.gov/cep-final-report.html.

Cook, T. D., & Campbell, D. T. (1979). *Quasi experimentation: Design and analysis issues infield settings.* Chicago, IL: Rand McNally.

Council of Europe (2001). *Common european framework of reference for languages.* Cambridge, UK: Cambridge University Press.

Council of Europe (2007). *From linguistic diversity to plurilingual education: Guide for the development of language education policies in Europe.* Strasbourg, France: Language Policy Division. DGIV Council of Europe.

Cummins, J. (1979). BICS and CALP: Origins and rationale for the distinction. *Sociolinguistic, 1–6*.

Dochy, F. (2006). *A guide for writing scholarly articles or reviews for the Educational Research Review. Educational Research Review.* Retrieved from http://www.journals.elsevier.com/educational-research-review/.

Duckworth, A., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237–251. https://doi.org/10.3102/0013189X15584327.

Garg, A. X., Hackman, D., & Tonelli, M. (2008). Systematic review and meta-analysis: When one study is just not enough. *Clinical Journal of the American Society of Nephrology, 3*, 253–260. https://doi.org/10.2215/CJN.01430307.

Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions. Version 5.1.0*London, United Kingdom: The Cochrane Collaboration. Retrieved from http://training.cochrane.org/handbook.

Institute of Education Sciences and National Science Foundation (2013). *Common guidelines for education research and development.* Washington, DC: United States Department of Education and United States National Science Foundation. Retrieved from http://goo.gl/8sjXtg.

Kremer, M., & Valcke, M. (2014, july). Teaching and learning in English in higher education: A literature review. *Paper presented to 6th international conference on education and new learning technologies, Barcelona, Spain*. Abstract retrieved from https://biblio.ugent.be/publication/5818549/file/5838141.pdf.

Lasagabaster, D., Doiz, A., & Sierra, J. M. (2014). Motivation: Making connections between theory and practice. In D. Lasagabaster, A. Doiz, & J. M. Sierra (Eds.). *Motivation and foreign language learning: From theory to practice* (pp. 173–183). Amsterdam: John Benjamins.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching, 51*(1), 36–76. https://doi.org/10.1017/S0261444817000350.

Marsh, D., Pavon, V., & Frigols, M. J. (2013). *The Higher Education languages landscape: Ensuring quality in English language degree programmes.* Valencia, Spain: Valencian International University.

Mears, D. P., & Barnes, J. C. (2010). Toward a systematic foundation for identifying evidence-based criminal justice sanctions and their relative effectiveness. *Journal of Criminal Justice, 38*(4), 702–710.

Organization for Economic and Cooperation Development (2017). *Education at a glance 2017: OECD indicators.* Paris, France: OECD Publishinghttps://doi.org/10.1787/eag-2017-en.

Ottenbacher, K., & Maas, F. (1999). How to detect effects: Statistical power and evidence-based practice in occupational therapy research. *American Journal of Occupational Therapy, 53*(2), 181–188.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences. A practical guide.* Malden, MA: Blackwell Publishing.

Reljic, G., Ferring, D., & Martin, R. (2015). A meta-analysis of the effectiveness of bilingual programs in Europe. *Review of Educational Research, 85*(1), 92–128. https://doi.org/10.3102/0034654314548514.

Rosenberg, W., & Donald, A. (1996). Evidence based medicine: An approach to clinical problem solving. *British Medical Journal, 310*, 1120–1126.

Rousseau, D. M. (2006). Is there such a thing as “evidence-based management”? *Academy of Management Review, 31*(2), 256–269.

Sackett, D., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence-based medicine: What is it and what it is not. *British Medical Journal, 312*, 71–72.

Schulz, K. F., Altman, D. G., Moher, D., & Fergusson, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *The Lancet, 375*(9721), 1144–1146.

Short, D. J. (2006). Content teaching and learning and language. In (2nd ed.). K. Brown (Vol. Ed.), *Encyclopedia of language and linguistics: Vol. 3*, (pp. 101–105). Oxford, England: Elsevier. https://doi.org/10.1016/B0-08-044854-2/00685-4.
Slavin, R. E. (2008a). Perspectives on evidence-based research on education. What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5–14.
Slavin, R. E. (2008b). Evidence-based reform in education: Which evidence counts? *Educational Researcher, 37*(1), 47–50. https://doi.org/10.3102/0013189x08315082.
Slavin, R. E. (2016, April 19). *Evidence and the every student succeeds act [Blog post]*. Retrieved from http://www.huffingtonpost.com/robert-e-slavin/evidence-and-the-essa_b_8750480.html.
Sloane, F. (2008). Through the looking glass: Experiments, quasi-experiments, and the medical model. *Educational Researcher, 37*(1), 41–46.
Soltero, S., & Ortiz, T. (2012). Discipline-based dual language immersion model in higher education (position paper). Retrieved from AGMUS Ventures website: http://goo.gl/5HCviS.
Stewart, L., Moher, D., & Shekelle, P. (2012). Why prospective registration of systematic reviews makes sense. *Systematic Reviews, 1*(7), https://doi.org/10.1186/2046-4053-1-7.
The Campbell Collaboration (2015). Campbell collaboration systematic reviews: Policies and guidelines [supplement]. *Campbell Systematic Reviews*. https://doi.org/10.4073/csr.2015.1.
Wächter, B., & Maiworm, F. (Eds.). (2014). *English taught programmes in European Higher Education. The state of play in 2014*Born, Germany: ACA Papers on International Cooperation in Education, Lemmens Medien GmbH. Retrieved from http://www.aca-secretariat.be/fileadmin/aca_docs/images/members/ACA-2015_English_Taught_01.pdf.
Williams, D. G. (2015). *A systematic review of EMI and implications for the south Korean HE context. English language teaching world online.* Retrieved from https://cpb-us-w2.wpmucdn.com/blog.nus.edu.sg/dist/7/112/files/2015/04/EMI-in-South-Korea_editforpdf-1gmsyy5.pdf.
*Yang, W. (2014). Content and language integrated learning next in Asia: Evidence of learners' achievement in CLIL education from a Taiwan tertiary degree programme. *International Journal of Bilingual Education and Bilingualism*. https://doi.org/10.1080/13670050.2014.904840.

Selected publications

*Kym, I., & Kym, M. H. (2014). Students' perceptions of EMI in higher education in Korea. *The Journal of Asia TEFL, 11*(2), 35–61.
*Huang, D. F. (2015). Exploring and assessing effectiveness of English medium instruction courses: The students' perspectives. *Procedia - Social and Behavioral Sciences, 173*, 71–78.
*Toledo, I., Rubio, F. D., & Hermosín, M. (2012). Creencias, rendimiento académico y actitudes de alumnos universitarios principiantes en un programa plurilingüe [Believes, academic performance and attitudes of first year students attitudes in a plurilingual program]. *Porta Linguarum, 18*, 213–229.
*Madrid, D., & Madrid, M. (2015). Efectos de la instrucción monolingüe y bilingüe en la enseñanza universitaria. Estudio de casos [Effects of monolingual and bilingual instruction on university teaching]. *Revista Mexicana de Investigación Educativa, 20*(67), 1255–1279.
*Hermosín, M., Alonso, P., & Cruz, R. (2011, march). *The improvement of students' attitudes in front of bilingualism in university context. Paper presented to INTED 2011 - international Technology*. Valencia, Spain: Education and Development Conference.
*Keiko, T., & Pérez, M. D. (2015). Comparing the language policies and the students' perceptions of CLIL in tertiary education in Spain and Japan. *Latin American Journal of Content & Language Integrated Learning, 8*(1), 25–35. https://doi.org/10.5294/laclil.2014.8.1.3.
*Aguilar, M., & Rodríguez, R. (2012). Lecturer and student perceptions on CLIL at a Spanish university. *International Journal of Bilingual Education and Bilingualism, 15*(2), 183–197. https://doi.org/10.1080/13670050.2011.615906.
*Pereira, S. I. (2016). Conciencia metacognitiva y estrategias de lectura en un contexto pro-AICLE a nivel universitario [Reading strategies and metacognitive awareness in a pro-CLIL context at the university level]. *Íkala, Revista de Lenguaje y Cultura, 21*(1), 81–97 1017533.udea.ikala.v21n01a06.
*Arco-Tirado, J. L., Fernández-Martín, F. D., & Hernández-Moreno, N. (2016). Skills learning through a bilingual mentors program in higher education. *International Journal of Bilingual Education and Bilingualism*. https://doi.org/10.1080/13670050.2016.1228601.
*Chostelidoua, D., & Grivab, E. (2014). Measuring the effect of implementing CLIL in higher education: An experimental research project. *Procedia - Social and Behavioral Sciences, 116*, 2169–2174. https://doi.org/10.1016/j.sbspro.2014.01.538.
*Arnó-Macià, E., & Mancho-Barés, G. (2015). The role of content and language in content and language integrated learning (CLIL) at university: Challenges and implications for ESP. *English for Specific Purposes, 37*, 63–73. https://doi.org/10.1016/j.esp.2014.06.007.
*Chimbganda, A. B. (2000). Communication strategies used in the writing of answers in biology by ESL first year science students of the University of Botswana. *English for Specific Purposes, 19*, 305–329.
*Dafouz, E., & Camacho, M. M. (2016). Exploring the impact of English-medium instruction on university student academic achievement: The case of accounting. *English for Specific Purposes, 44*, 57–67. https://doi.org/10.1016/j.esp.2016.06.001.
*Ament, J. R., & Pérez-Vidal, C. (2015). Linguistic outcomes of English medium instruction programmes in higher education: A study on economics undergraduates at a Catalan university. *Higher Learning Research Communications, 5*(1), 47–68. https://doi.org/10.18870/hlrc.v5i1.239.
*Sierra, L., & López, A. (2015). CLIL en la formación inicial del profesorado de educación infantil y primaria: La experiencia del CES Don Bosco [CLIL on infant and primary initial teacher training: The Don Bosco CES experience]. *Educación y Futuro, 32*, 83–114.
*Lasagabaster, D. (2012). El papel del inglés en el fomento del multilingüismo en la Universidad [The role of English on the promotion of multilingualism at the University]. *Estudios de Lingüística Inglesa Aplicada, 12*, 13–44.
*Goodman, B. A. (2014). Implementing English as a medium of instruction in Ukrainian university: Challenges, adjustments, and opportunities. *International Journal of Pedagogies and Learning*. https://doi.org/10.1080/18334105.2014.11082026.
*Maíz-Arévalo, C., & Domínguez-Romero, E. (2013). Students' response to CLIL in tertiary education: The case of business administration and economics at Complutense University. *Revista de Lingüística y Lenguas Aplicadas, 8*, 1–12.
*González, J. B. (2013). (In)compatibility of CLIL and ESP courses at university. *Language Value, 5*(1), 24–47.
*Tzoannopoulou, M. (2015). Rethinking ESP: Integrating content and language in the university classroom. *Procedia - Social and Behavioral Sciences, 173*, 149–153.
*Jiménez, A. J. (2014). Measuring the impact of CLIL on language skills: A CEFR-based approach for higher education. *Language Value, 6*(1), 28–50. https://doi.org/10.6035/Language.2014.6.4.
*Popović, M., Vagić, M., Kuzmanović, M., & Andelković, J. (2016). Understanding heterogeneity of students' preferences towards English medium instruction: A conjoint analysis approach. *Yugoslav Journal of Operations Research, 26*(1), 91–102. https://doi.org/10.2298/YJOR140915009P.
*Hernández-Nanclares, N., & Jiménez-Muñoz, A. (2015). English as a medium of instruction: Evidence for language and content targets in bilingual education in economics. *International Journal of Bilingual Education and Bilingualism*. https://doi.org/10.1080/13670050.2015.1125847.
*Jiménez-Muñoz, A. (2015). Flipping lectures: Analysing student workload in EMI contexts. *Procedia - Social and Behavioral Sciences, 212*, 35–41. https://doi.org/10.1016/j.sbspro.2015.11.295.
Yang, W., & Gosling, M. (2014). What makes a Taiwan CLIL programme highly recommended or not recommended? *International Journal of Bilingual Education and Bilingualism, 17*(4), 394–409. https://doi.org/10.1080/13670050.2013.808168.
*Studer, P. (2015). Coping with English: Students' perceptions of their teachers' linguistic competence in undergraduate science teaching. *International Journal of Applied Linguistics, 25*(2), 183–201. https://doi.org/10.1111/ijal.12062.
*Mancho-Barés, G., & Llurda, E. (2013). Internationalization of business English communication at university: A three-fold needs analysis. *Iberica, 26*, 151–169.
*Moore, E. (2016). Conceptualising multilingual higher education in policies, pedagogical designs and classroom practices. *Language Culture and Curriculum, 29*(1), 22–39. https://doi.org/10.1080/07908318.2016.1132652.
*Moore, E., & Dooly, M. (2010). “How do the apples reproduce (themselves)?” How teacher trainees negotiate language, content, and membership in a CLIL Science

Education Classroom at a Multilingual University. *Journal of Language, Identity and Education, 9*(1), 58–79. https://doi.org/10.1080/15348450903523591.
*Lindgrén, S. A., & Laine, M. (2011). Cognitive linguistic performances of multilingual university students suspected of dyslexia. *Dyslexia, 17*, 184–200. https://doi.org/10.1002/dys.422.
*Kim, A., Son, Y. D., & Sohn, S. Y. (2009). Conjoint analysis of enhanced English Medium Instruction for college students. *Expert Systems with Applications, 36*, 10197–10203. https://doi.org/10.1016/j.eswa.2009.01.080.
*Vilkancienė, L. (2011). CLIL in tertiary education: Does it have anything to offer? *Studies About Languages, 18*, 111–116.
*Valcke, J., & Pavón, V. (2013). A comparative study on the use of pronunciation strategies for highlighting information in university lectures. In R. Wilkinson, & M. L. Walsh (Eds.). *Integrating content and language in higher education. From theory to practice* (pp. 323–341). Frankfurt, Germany: Peter Land Edition.
*Kung, F. W. (2013). The more the merrier? Bilingualism in an academic perspective: Exploring the implementation of English-medium instruction in Taiwanese tertiary education. *Asian EFL Journal, 15*(4), 8–35.
*Bozdoğan, D., & Karlıdağ, B. (2013). A case of CLIL practice in the Turkish context: Lending an ear to students. *Asian EFL Journal, 15*(4), 89–110.
*Ginesta, X., Coll-Planas, G., & De San Eugenio, J. (2013). La aplicación del método AICLE en los estudios de comunicación de la Universidad de Vic [The application of the CLIL method on the communication studies at the university of Vic]. *Estudios sobre el Mensaje Periodístico, 19*, 813–821.
*Chapple, J. (2015). Teaching in English is not necessarily the teaching of English. *International Education Studies, 8*(3), 1–13. https://doi.org/10.5539/ies.v8n3p1.
*. Arkın, I. E. (2013). *English-medium instruction in higher education: A case study in a Turkish University context*(doctoral dissertation). Gazimağusa, North Cyprus, Turkey: Eastern Mediterranean University.
*Karabinar, S. (2008). Integrating language and content: Two models and their effects on the learners' academic self- concept. In R. Wilkinson, & V. Zegers (Eds.). *Realizing content and language integration in Higher Education* (pp. 53–63). Maastricht, Netherlands: Maastricht University.
*Tarasheva, E. (2008). Media as content for language learning at university: Impact on language proficiency. In R. Wilkinson, & V. Zegers (Eds.). *Realizing content and language integration in Higher Education* (pp. 179–190). Maastricht, Netherlands: Maastricht University.
*Airey, J., & Linder, C. (2006). Language and the experience of learning university physics in Sweden. *European Journal of Physics, 27*, 553–560. https://doi.org/10.1088/0143-0807/27/3/009.
*Byun, et al. (2010). English-medium teaching in Korean higher education: Policy debates and reality. *Higher Education, 62*(4), 431–449. https://doi.org/10.1007/s10734-010-9397-4.
*Evans, S., & Morrison, B. (2011). Meeting the challenges of English-medium higher education: The first-year experience in Hong Kong. *English for Specific Purposes, 30*, 198–208. https://doi.org/10.1016/j.esp.2011.01.001.
*Kagwesage, A. M. (2013). Coping with English as language of instruction in higher education in Rwanda. *International Journal of Higher Education, 2*(2), 1–12. https://doi.org/10.5430/ijhe.v2n2p1.
*Sert, N. (2007). The language of instruction dilemma in the Turkish context. *System, 36*, 156–171. https://doi.org/10.1016/j.system.2007.11.006.
*. Vinke, A. (1995). *English as the medium of instruction in Dutch engineering education*(Doctoral dissertation). Delft, Netherlands: Delft University of Technology.
*Airey, J. (2011). The relationship between teaching language and student learning in Swedish university physics. In B. Preisler, A. Fabricius, & I. Klitgård (Eds.). *Language and Learning in the international university: From English Uniformity to Diversity and hybridity (3–18)*. Bristol, UK: Multilingual Matters.
*Joe, Y., & Lee, H. K. (2013). Does English-medium instruction benefit students in EFL contexts? A case study of medical students in Korea. *Asia-Pacific Education Research, 22*(2), 201–207. https://doi.org/10.1007/s40299-012-0003-7.
*Tai, H. Y. (2015). Writing development in syntactic complexity, accuracy and fluency in a content and language integrated learning class. *International Journal of Language and Linguistics, 2*(3), 149–156.
*Pessoa, S., Miller, R. T., & Kaufer, D. (2014). Students' challenges and development in the transition to academic writing at an English-medium university in Qatar. *International Review of Applied Linguistics in Language Teaching, 52*(2), 127–156. https://doi.org/10.1515/iral-2014-0006.
*Kim, J., Tatar, B., & Choi, J. (2014). Emerging culture of English-medium instruction in Korea: Experiences of Korean and international students. *Language and Intercultural Communication, 14*(4), 441–459. https://doi.org/10.1080/14708477.2014.946038 441–459.
*Hu, G., & Lei, J. (2014). English-medium instruction in Chinese higher education: A case study. *Higher Education, 67*(5), 551–567. https://doi.org/10.1007/s10734-013-9661-5.
*Dafouz, E., Camacho, M. M., & Urquia, E. (2014). 'Surely they can't do as well': A comparison of business students' academic performance in English-medium and Spanish-as-first-language-medium programmes. *Language and Education, 28*(3), 223–236. https://doi.org/10.1080/09500782.2013.808661.
*Aguilar, M., & Muñoz, C. (2014). The effect of proficiency on CLIL benefits in Engineering students in Spain. *International Journal of Applied Linguistics, 24*(1), 1–18. https://doi.org/10.1111/ijal.12006.
*. Rogier, D. (2012). *The effects of English-medium instruction on language proficiency of students enrolled in higher education in the UAE*(Doctoral dissertation). United Kingdom: University of Exeter.
*Islam, M. M. (2013). English medium instruction in the private universities in Bangladesh. *Indonesian Journal of Applied Linguistics, 1*(3), 126–137.
*Kirkgöz, Y. (2014). Students' perceptions of English language versus Turkish language used as the medium of instruction in higher education in Turkey. *Turkish Studies, 9*(12), 443–459.
*Gao, X. (2008). Shifting motivational discourses amongst mainland Chinese students in an English medium tertiary institution in Hong Kong: A longitudinal inquiry. *Studies in Higher Education, 33*(5), 599–614.
*Sultana, S. (2014). English as a medium of instruction in Bangladesh's higher education: Empowering or disadvantaging students? *Asian EFL Journal, 16*(1), 11–52.