The Impact of International Speaking and Listening Assessments on Primary School Bilingual Learning: Insights from Survey Research

Mark Griffiths
Trinity College London
mark.griffiths@trinitycollege.com

Abstract

256 primary school teachers from the bilingual Spanish-English teaching project Autonomous Region of Madrid, Spain, were recruited to complete an online survey regarding what impact the use of external international speaking and listening assessments has had on their students and their learning, and how they as teachers use speaking and listening exam materials as a learning tool in the classroom. By utilising best practice in the design and execution of the survey, and by analysing the data using descriptive and inferential statistics, it was found that a large majority of teachers believe that the use of the external oral exams has improved learning outcomes, particularly with regard to their students' confidence and skills in spoken communication. Correlating areas of improvement in the students' English skills were identified. The teachers also reported that they used the exam materials for a wide range of purposes. A factor analysis of the teachers' responses revealed that there were five purposes, only two of them related directly to exam training, the other three related to other areas of English teaching, skills development, monitoring and evaluation. These five purposes are presented here as an activity and support planning framework for the bilingual classroom. The impact data and activity and support planning framework should be of interest to teachers, education professionals and publishers, and to those responsible for teacher support and education policy decisions.

13.1. Introduction

13.1.1. Context

This research was conceived and conducted with the aim of drawing to the surface evidence regarding how international speaking and listening assessments are used in and impact on the bilingual learning context. Since the early 2000's state-initiated bilingual projects have been introduced in different autonomous communities in Spain, including Andalucía, Murcia, Canarias and Cantabria. In this research, the focus is on the bilingual project of the Comunidad de Madrid, the local government organisation for the autonomous municipality of Madrid, Spain. The bilingual project in Madrid was one of the first of its kind in Spain, set up to improve the English language ability of primary age learners in Madrid, using content and language integrated learning (CLIL)/English as the medium of instruction (EMI) methodologies as the vehicle for the improvement. One of the early targets for the project was for students aged seven and eight to achieve a level of A2.1 on the Common European Framework of Reference (CEFR) in their oral communication skills.

13.1.2. Introduction of External Assessments

From 2004, the Comunidad de Madrid chose to introduce an external assessment to the bilingual project. It was felt that external language assessments could provide an objective and independent measure of student progress and serve to audit the students' overall English oral communication skills. The exams would take place towards the end of each school year (around May), thereby providing an oral communication target for teachers and students to aim for that complemented the overall aims of the bilingual project. In order to introduce a measurable target using a respected and objective system of evaluation, the Comunidad went into partnership with Trinity College London. Trinity administered the Graded Examination in Spoken English (GESE) with lower learners taking GESE Grade 2 (A1 on the CEFR) and stronger learners taking GESE Grade 3 (A2.1 on the CEFR). Details of these exams and GESE Examination Specifications can be found on the Trinity College London (2017) website.

For the academic year 2006-2007, the external speaking and listening assessment of learners in the bilingual project was expanded. The first learners to have taken an external assessment in English oral communication skills were now two years older, and the Comunidad felt that a further measurement of progress was required, with a target of the learners achieving A2.2 or, ideally, B1 on the CEFR. Again, working in partnership with Trinity College London, the Comunidad introduced assessments at higher levels in GESE, with nine and ten year olds now taking either GESE Grade 4 (A2.2 on the CEFR) or GESE Grade 5 (B1 on the CEFR). A further expansion of the project was introduced in 2008 with students in the final years of primary being set the goal of achieving A2 or B1 in all four skills. For this, the Comunidad worked with Cambridge ESOL, introducing their Key English Test (KET) and Preliminary English Test (PET) for students aged eleven and twelve. Although the Cambridge assessments are four-skills assessments, they nevertheless include the assessment of speaking and listening skills and are therefore included in this research. Details of Cambridge language assessments can be found on the Cambridge ESOL (2017) website.

By 2016, the bilingual project in the Comunidad de Madrid had grown to include 391 public and 71 subsidised private bilingual teaching centres, with more than 33,000 students being enrolled in external English assessments for the 4th and 6th years of primary.

13.2. Research Aims

This research concerns itself with two lines of investigation. The first looks to explore any positive impacts that can result from integrating examination preparation into the bilingual school curriculum and using exam-like exercises and materials in the classroom. One might suggest that a view onto possible impacts can come from the certified achievement of language levels in the form of exam results. However, results are but one view of learning outcomes. They do not reveal any details of the learning process or washback - the impact of an exam on teaching and learning in the classroom. As an alternative to focusing on exam results, one could also ask the learners about their experience of the exams. However, there are both ethical and practical constraints inherent in such an approach: students in this study would be primary age pupils with a lack of awareness of the learning process and an inability to articulate their learning experience. The focus therefore shifts to the teachers to provide their reports of any positive impacts these assessments and exam materials may have on teaching and learning. This brings us to research questions (RQ)1 and 2:

RQ1: What positive impacts on learners can result from integrating examination preparation into the curriculum and using exam-like exercises in the classroom?

The second line of investigation in the present research concerns how the teachers use exam materials in the bilingual primary classroom. One might logically expect the answer to such an enquiry to be something along the lines of, 'I use exam materials to prepare for exams'. But is this really the case? Is that simply an assumption without evidence? Do teachers use the exam materials for more than just exams? This brings us to RQ2:

RQ2. In the most recent year of teaching, in what ways did teachers use the international speaking and listening exam materials in the classroom?

This question was general and did not seek to specifically name activity types as it was felt that the range of materials available in books, video and other formats was so broad that to specify which materials had been used risked creating a false negative with teachers reporting that because their type of activity was not listed, they would report no use of exam materials at all.

13.3. Method

13.3.1. Considerations in Survey Design

The research was conducted using survey techniques. All too often, considerations in good survey design are omitted from the method sections of social research papers. Perhaps the proliferation of professional-looking, easy-to-access online survey tools leads researchers to presume that there are few theoretical considerations in achieving reliable social survey results. It is no coincidence, however, that where survey techniques are present but there is little evidence in the method section of an understanding of survey design theory or best practice, there is often a poorly-designed and executed survey close by. The construction of valid and reliable survey questionnaires and the collection of accurate and reliable data requires an understanding of good survey design practice, possible sources of measurement error, and of how a respondent's answer may differ from the conceptual 'true value' (Olsen and Parkhurst 2013). The present survey was constructed utilising best practice in survey design, in full awareness of the common pitfalls that may arise. Here I provide a brief summary of some of the considerations that have informed the construction of the questionnaire used in this research.

i) Concept/construct specification: 'The research objectives of many survey studies are ill-defined' (Schwartz, 1997). Only when one has identified the underlying construct or concepts

that the survey is attempting to access and measure, should one consider the creation of survey items, by translating concepts from a construct specification into questions. These must then be piloted to ensure satisfactory levels of validity and reliability. The present survey started life as a concept specification and the survey items were created from this blueprint.

ii) Respondent behaviour: respondents are a major source of measurement error. It is commonly assumed by researchers unfamiliar with survey theory that if a respondent provides an answer, their response is honest or accurate. But the reality is that respondents are humans, and as humans, we are vulnerable to cognitive biases. Our responses can be mediated by psychological factors such as topic sensitivity and personal vulnerability. Motivated misreporting –providing socially desirable, face-supporting or self-protecting responses– is an ever-present threat to eliciting reliable survey data as we look for clues regarding researcher-desired responses and acquiesce to what we believe is desired of us.

iii) Respondent cognitive burden: our performance is compromised by heavy cognitive load, high demand on working memory and lengthy, taxing questionnaires. Respondent fatigue or faulty recall are a common consequence of a heavy cognitive burden, and if the respondents do not drop out completely, they may look to get through a questionnaire by 'satisficing' (Krosnick and Alwin, 1987) – using an answering strategy that is less demanding of mental effort. This leads individuals to be less than thorough and select the first items that are 'good enough'.

The researcher in this study gave no indication of any desired outcomes, and avoided probing for sensitive information. The survey was sensitive to the danger of respondent burn-out and was designed not to require a large amount of cognitive processing or effort to retrieve judgements from memory.

iv) Question wording and sequencing: question wording and sequencing are sources of measurement error. Good survey practice tells us that questions should be clear and unambiguous, using terms that all respondents will interpret the same way. Each item should contain only one single question and there should be no double-barrelled questions. Negative questions should generally be avoided to prevent misunderstanding of the question by the respondent and misunderstanding of the respondent's answer by the researcher. The sequencing of items can influence how respondents behave, and the layout and even spacing of scales can significantly affect the responses given. The present research was sensitive to all of these issues.

v) Accessing attitudes and opinions: Researchers must be aware that the public do not conduct their lives with a set of fixed, stable and readily available attitudes and we do not hold an opinion on everything. However, by the age of two or three, we have lost the ability not to answer and will always provide replies, whether we hold an opinion or not. As a general rule, the greater the abstraction, irrelevance or unfamiliarity to the respondent, the greater the chances of respondents simply inventing opinions on the spot, yielding unreliable survey data. The problem here is that researchers commonly mistake spontaneously created researcher-pleasing responses for real-world, strongly held views, and come to inaccurate and unreliable conclusions regarding the meaning of their data. The present research was designed around constructs that are real, relevant and available to the audience of teachers using concepts that they could readily express in their own words if required.

vi) Ethics: at all points, the teachers were assured of their anonymity, that any data they provided would be handled securely and that they could leave the research at any time with no consequence to themselves. They were told who was conducting the survey, where the data would be stored and given a contact email address. The whole questionnaire was designed in accordance with the guidelines for good practice set out by the British Association for Applied Linguistics (2017).

13.3.2. Participants

The bilingual project in the Comunidad de Madrid has grown significantly since its inception, with hundreds of teachers being involved in the delivery of the Madrid primary school syllabus using both English and Spanish as the medium of instruction. In total, 256 of these teachers responded to the various questions in the research and Table 1 gives a breakdown of which exams they reported that they have taught. In administering the survey, the questionnaire did not mention the exams by name, as it was anticipated that to do so could entail teachers responding to knowledge of or loyalty to/dislike of an exam brand rather than the focus on the learning outcomes and methods.

Unsurprisingly, given the history of external assessments in the bilingual project, the distribution shows that more teachers have experience of teaching Trinity's GESE exams than the Cambridge KET/PET exams, with 88% of all respondents having taught GESE, and 51% having taught only GESE, never KET/PET. 49% of teachers have some experience of teaching the Cambridge KET/PET exams and 12% of teachers have only experience of preparing for Cambridge exams. 37% of teachers responded that they have experience of preparing both Trinity and Cambridge exams.

TABLE 13.1. Which exams teachers who participated in the research have taught 2011-2016

	I taught only GESE	I taught both KET/PET and GESE	I taught only PET/KET
256 teachers responded that they have taught in the bilingual programme from 2011-2016	132 (51%)	94 (37%)	30 (12%)

13.3.3. Structure of the Survey

Both research questions in the survey used their own measurement scale. For RQ1, the teachers were given a list of nine categories representing possible areas of impact. These impact areas were chosen to reflect some of the common improvements that we know from anecdotal evidence teachers report. The teachers were able to indicate their responses using three categories:

➢ no positive effect / some improvement / a big improvement

As RQ1 gave answers that were potential representations of the single underlying concept - the impact of the language exam on the students' communication and study skills. The nine impact variables were also checked for reliability, to ensure that each variable had the same meaning for each respondent, and to confirm that the question items and scores are internally consistent. The reliability of the items in this section was checked using Cronbach's coefficient alpha. It is generally accepted that a value greater than $\alpha = .8$ is appropriate for cognitive surveys. The result of the Cronbach's alpha was $\alpha = .928$, and there was no increase in reliability if any individual item were deleted. This shows us that this section of the questionnaire was reliable and addressing one underlying construct.

For RQ2, the teachers were provided with a list of sixteen possible uses of exam materials in their classes and asked to indicate if, in the year they had most recently taught on the bilingual programme, they had or had not used the exam materials in these ways. The most recent year was chosen to reduce cognitive and memory demand. They were provided with a binary choice of response:

➢ I did this in my class / I didn't do this in my class

As these sixteen categories were not seen as dimensions of the same concept, it was not appropriate to run a reliability analysis in the form of Cronbach's alpha. However, as we will see in the Results section, an alternative analysis was run to investigate the meaning of the reported behaviours.

13.3.4. Administering the Survey

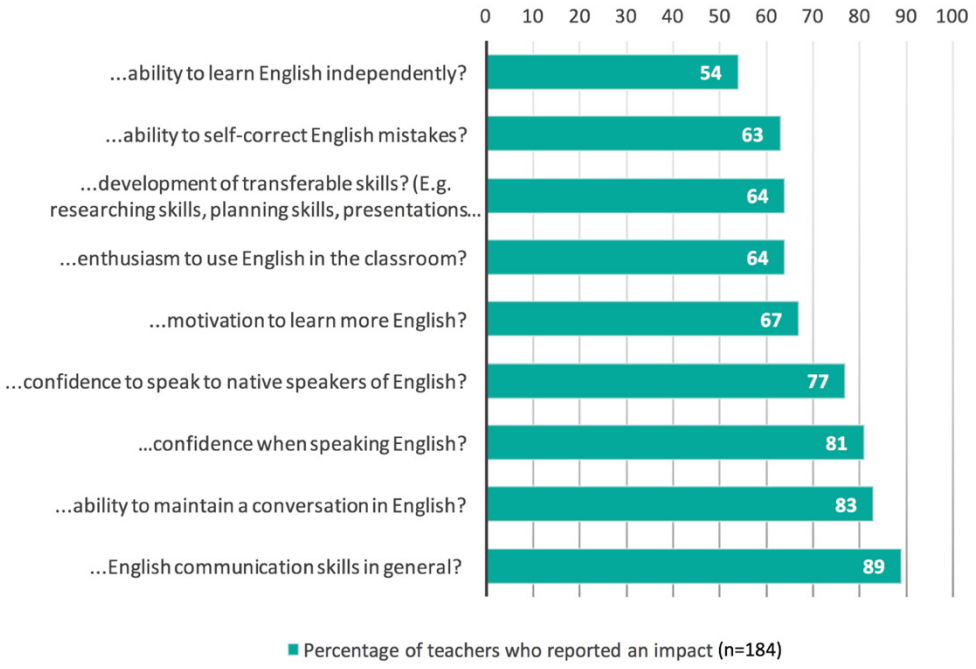
The survey took the form of an online questionnaire and was administered to the teachers locally in Spain using an online survey tool. The email with the survey invite and link was sent to each school in the Comunidad bilingual project and the survey could be filled in by one or many teachers from each school that received the invitation to the survey. It was sent out to schools in October 2016 and the survey remained open for three weeks.

13.4. Results

13.4.1. RQ1: Impact of Preparing for the External Speaking and Listening Assessment

For RQ1, the teachers were provided with nine areas of learning and development to consider on which preparing for an international speaking and listening exam may have had an impact. The teachers' responses indicated an emphatic belief that the exam preparation had had an impact, as seen in Figure 13.1.

FIGURE 13.1. Teachers' reports of the impact of preparing for an international speaking and listening exam

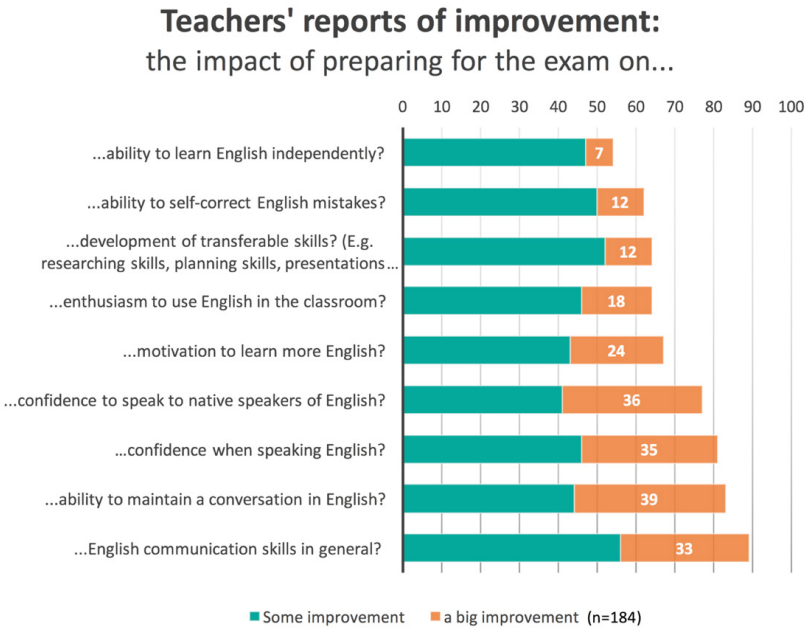


Whilst teachers reported a strong belief that the exam preparation has had an impact on the given areas of learning and development, that reported impact was not uniform across all variables. Only 54% of teachers reported an improvement in the students' ability to learn English independently, whilst 89% of teachers reported that the exam preparation has had an impact on their learners' English communication skills in general. At the lower end of the scale, it comes as no great surprise that exam preparation has not resulted in improvements for all students in the areas of independent learning, self-correction and transferable skills, given that they are primary school learners as young as seven and eight. Yet while this is the smallest reported impact, it should not be dismissed: 54% of teachers reporting that they have observed at least some improvement in an area ordinarily associated with older learners is an impact worthy of note. Likewise, two thirds of teachers reporting improvements in their students' abilities to self-correct, their development of transferable skills, and an increase in their enthusiasm and motivation to learn more English and use it in the classroom, all as a result of the preparation for an external speaking and listening exam, is a substantial finding.

Further up the scale, over three quarters of teachers report improvements in their students' confidence when speaking and when speaking to native speakers. These reports depict an impact on learning that contrasts enormously with the historical, text book-focused, didactic teaching approach in Spain, a system that traditionally relied heavily on written grammar practice at the expense of developing oral communicative competence, and a system that the bilingual project was devised to move away from. Just how far things have moved on is perhaps most vividly illustrated by over 80% of teachers reporting an increase in the students' abilities to maintain a conversation in English and nearly 90% reporting an improvement in their students' communication skills in general as a result of preparing for the international speaking and listening exams.

The results above give us a very clear view of the teachers' belief that exam preparation has had an impact on the students, but there is more to the story: also built into the survey was the option to report the scale of the impact, by providing teachers with three choices: 'no improvement', 'some improvement' and 'a big improvement'. The figures for 'a big improvement' are contained in Figure 13.2.

FIGURE 13.2. Teachers' reports of impact: 'some improvement' and 'big improvement'



In the areas associated more with learner independence, reports of big improvements are not as frequent. This is no great surprise, given that these students are still of primary school age. However, we see in Figure 13.2 that a quarter of teachers report having seen 'a big improvement' in their students' motivation to learn more English, and as many as 35% of teachers chose to emphasise that they have seen 'a big improvement' in their students' learning and communication skills as a direct result of preparing for the international speaking and listening exams: *confidence to speak to native speakers*, *confidence when speaking English*, *ability to maintain a conversation in English* and *English communication skills in general*, are all reported to have seen a big improvement, with almost 40% of teachers reporting a big improvement in their students' ability to maintain a conversation in English. These reported improvements are apart from any other improvements that may have resulted from following the Spanish curriculum or improvements in the quality of the teaching, which are ongoing, year-round sources of learning and improvement. These data indicate a strong impact of exam preparation on learning and development and present a remarkable example of positive washback from the oral focus of the external exam onto the classroom.

A second area of focus in the analysis of these data was the possible identification of correlations between the variables: did certain variables behave in the same way, according to the teachers? A two-tailed Spearman's correlation was run to assess the relationship between the nine question variables and the 184 teacher responses. The result of the Cronbach's alpha had previously confirmed that this section is most likely measuring one underlying construct. It is unsurprising, therefore, that all nine items correlated with each other. There were nineteen moderate correlations (r_s = .48 to .59, ρ = .01), and seventeen strong correlations (r_s = .60 to .72, ρ = .01). All correlations were positive. Of particular interest was the identification of three groups within which each variable strongly correlated with all others.

- improved ability to maintain a conversation in English
- improvement in English communication skills in general
- greater confidence to speak to a native speaker of English
- increased confidence when speaking English

- increased motivation to learn more English
- enthusiasm to use English in the classroom
- increased confidence when speaking English

- increased motivation to learn more English
- increased ability to learn English independently
- increased ability to self-correct English mistakes

When handling correlation data, we must remember that no causation is implied by the results. What the correlation data tell us is that certain reported impacts of the exams on the students co-occur in patterns – when one is high, another will likely be high; if one is low, another will likely be low, etc. Observing these correlations gives us an insight into the strong links between certain variables and helps us to predict the impact of groups of variables in given circumstances. Moreover, identifying groups of variables that behave in a similar way has the potential to inform the focus of future academic support materials for teachers and the measurement of improvements in areas of learning and development.

13.4.2. RQ2: Teachers' Use of Exam-Like Materials in the Classroom

Unlike RQ1, which asked the teachers to describe their view of the students, RQ2 focused on teachers' self-reports of their own professional activities: their use of exam materials in their

classes. The data analysis first considered what percentage of teachers did or did not do any of these activities. These activities were then tabulated in order of most common to least common as shown in Table 13.2. 185 teachers completed this section, and Table 13.2 tells us what percentage of the teachers reported doing a particular activity.

TABLE 13.2. Percentage use of exam materials for different purposes in class

Purpose	%
to familiarise my students with the exam format	93
to provide my students with a model of how to do the exam	92
to reduce student anxiety about what is going to happen in the exam	90
in the class, as part of my exam preparation classroom activities	87
as a reference to help plan my exam preparation classes	83
as a tool to encourage more communication in the classroom	77
to help my students practise understanding of a native speaker	75
in the class, as part of my general English classroom activities	73
as a reference to help plan my general English classes	69
to help monitor students' progress.	66
to show my students examples of successful spoken communication in general	65
as a tool for improving my students' grammatical accuracy	64
as a tool to encourage dynamic and spontaneous communication	63
to develop learner autonomy	54
to help evaluate my students	50
to accelerate classroom learning	42

Almost every teacher that responded indicated that they use the exam materials for exam familiarisation, to provide a model of how to do the exam, to reduce anxiety about what is going to happen in the exam, as part of the exam preparation activities and as a reference to help plan exam preparation classes. It is perhaps no surprise to see that the most common uses of exam materials were exam-related: the teachers have an external international speaking and listening exam to prepare their students for as part of the bilingual programme and it would rather mystifying if teachers chose to ignore the exam materials. However, the use of exam materials is not confined to exam preparation alone. To explore this further, a factor analysis was conducted.

A factor analysis identifies and examines clusters of large correlation coefficients between subsets of questions. It can show us that the questions in a survey do not necessarily each investigate an individual dimension. It may be that there are in fact fewer underlying dimensions than questions; some of the questions may be related, measuring different aspects of a smaller number of underlying dimensions. These underlying dimensions are known in statistical terms as factors. By reducing the large number of questions into smaller sets of factors, a factor analysis explains the largest amount of variance using the smallest number of explanatory concepts.

The procedure for performing the factor analysis was as follows: the ideal number of cases for a factor analysis is over 300, and in this section of the survey, we have only 185 responses. A Kaiser-Meyer-Olkin measure of sampling adequacy was therefore performed to ensure the sample was adequate. The value of the KMO should be greater than 0.5 for the sample to be adequate, and in the case of our data, the value was 0.8 and sufficient to proceed to the next stage. An examination of correlations was performed next to ensure that no question variables correlated very strongly, which would be an indicator that two questions were measuring identical

concepts. There were no strong correlations, indicating that each question was measuring a different variable. Next, an exploratory factor analysis using principal component analysis was conducted. This produces a scree plot with Eigenvalues, which tell us how many factors the number of question variables can be reduced to. The result of this test tells us that although there were 16 questions that all appeared to be different, the 16 questions load onto just five factors. That is, the 16 questions are measuring different aspects of five underlying concepts, behaviours or activities. The results were rotated using Varimax and Kaiser normalisation to produce a rotated component matrix grouping the question variables into the five related sub-groups identified as existing in the scree plot. At this point, the informed human judgement of the researcher is required to identify a common link that unites all of the features in each factor or sub-group. Below are given the results of the factor analysis, with each question variable grouped into its respective factor, and a name provided to identify each factor based on the theme that links the questions contained within it.

FACTOR 1: Using exam materials for exam familiarisation and modelling
- To familiarise my students with the exam format (93%)
- To provide my students with a model of how to do the exam (92%)
- To reduce student anxiety about what is going to happen in the exam (90%)

FACTOR 2: Using exam materials for exam preparation and planning
- In the class, as part of my exam preparation classroom activities (87%)
- As a reference to help plans my exam preparation classes (83%)

FACTOR 3: Using exam materials for general English classes and class planning
- In the class, as part of my general English classroom activities (73%)
- As a reference to help plan my general English classes (69%)

FACTOR 4: Using exam materials for improving real-life independent communication
- As a tool to encourage more communication in the classroom (77%)
- To help my students practise understanding of a native speaker (75%)
- As a tool for improving my students' grammatical accuracy (64%)
- As a tool to encourage dynamic and spontaneous communication (63%)
- To show my students examples of successful spoken communication in general (65%)
- To develop learner autonomy (54%)
- To accelerate classroom learning (42%)

FACTOR 5: Using exam materials to monitor and measure progress
- To help monitor students' progress (66%)
- To help evaluate my students (50%)

The five factors that emerged from the factor analysis show us something perhaps never seen before. They provide insight into how teachers are really working; how, according to their own reported practices, they use exam materials in the classroom, and how they conceptualise the use of exam materials in the bilingual learning context. We see in Factor 1 and in Factor 2, that teachers are using the exam materials for exam-related activities, but for different aspects of exam preparation. Where Factor 2 shows what we might expect teachers to be doing with exam materials – explicit exam preparation and planning, Factor 1, the largest factor, shows us that there is more to exam preparation than explicit exam practice. Almost all teachers also use exam materials for exam familiarisation and modelling, demonstrating to the learners what the exam looks like and how it works.

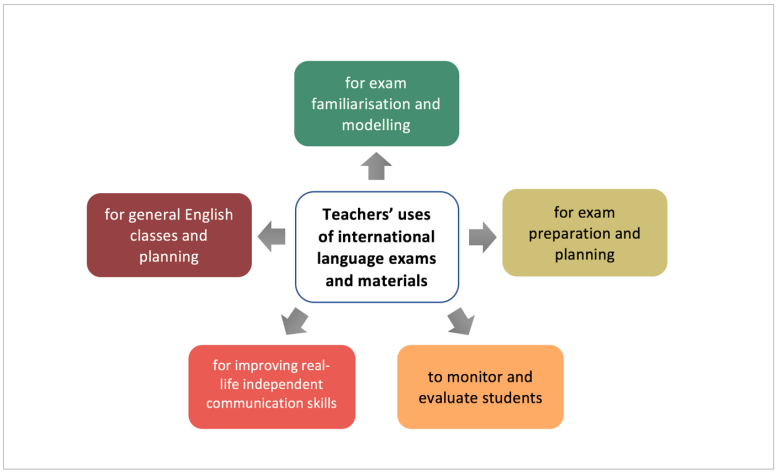
It is when we examine the remaining three factors that we see a pattern of usage that might not have been predicted. Factor 3 identifies a relationship between the exam content

and the content of the general English lessons, with around 70% of teachers using the exam materials as part of their general English classes and for class planning. Here we see clearly that the utility of the exam materials extends beyond the exam preparation itself. It appears that the international speaking and listening exam content and aims have synthesised with the general English teaching aims of the teachers and the Spanish bilingual primary curriculum overall, and it is no longer clear where exam teaching and general English teaching begin and end. This implication is perhaps more sharply defined by Factor 4, in which we see the largest number of non-exam directed activities, showing unequivocally that the exam materials are used for more than exam preparation. No less than seven different activities load onto the factor summarised as 'developing real-life independent communication skills', with three-quarters of the teachers reporting that they use the exam materials to encourage more communication in the classroom and develop understanding of a native speaker. This suggests that the specific speaking and listening focus of the international assessments is very much complementing the oral focus of the bilingual classroom. The reader may also be surprised to see that the great majority of teachers reported using the oral exam materials for improving students' grammatical accuracy. Unlike the traditional approach to teaching grammar, the teachers appear to be providing their students with models of interaction and interactional competence, illustrating how the grammar and functions the students learn as part of their linguistic focus are operationalised in real life interactions.

The final factor identified by the factor analysis, Factor 5, is that of using the exam materials to monitor and evaluate the students, with the majority of teachers using the exam materials presumably to guide and audit general teaching and learning. The research was not designed with any expectation that numbers on this scale would be high, due to teachers having at their disposal the main Spanish curriculum, past exam papers and general coursebooks that they use year-round. Moreover, teachers are not trained examiners and are not in possession of the examiner guidance or marking criteria. Yet still, despite the availability of other tools, the teachers use the exam materials for monitoring and evaluating their students.

Pulling the research evidence together regarding teachers' use of international speaking and listening exam materials, there is no doubt that there may be other activities that may load onto the five dimensions and future research will add greater detail to the factors. It is not certain, however, that there are any more than the five factors reported here – again, this will be a matter for future research. At this stage, I propose a framework for designing and planning class activities and providing teacher support for the bilingual classroom using speaking and listening exam materials:

FIGURE 13.3. Activity and support planning framework for the bilingual classroom



13.5. Closing Remarks

The data collected in this research indicate that the use of external international speaking and listening assessments has had a major impact on the Comunidad de Madrid primary school bilingual project. First, the teachers reported that the use of the external oral exams has resulted in a range of improvements to their students' English, particularly in the areas of communication skills and speaking confidence. Naturally, these data will be of interest to those working in Autonomous Region of Madrid itself; but the percentage scores and the results of the correlation analysis in RQ1 should provide teachers, parents and education policy makers further afield with greater predictive power regarding possible outcomes of employing these external oral exams, and these data may assist in the evaluation of whether the pedagogical benefits external oral assessments may provide are reconcilable with the financial costs involved in providing them.

A second insight resulting from this research comes from RQ2: teachers reported having integrated the exam materials into their classes for both exam and non-exam related activities, blurring the line between exam preparation and general English teaching in the bilingual classroom. The percentage usage data and the factor analysis offer us a unique view of what is happening in the classroom, reported by those at the chalkface. The factorial model demonstrates how teachers conceptualise these exam materials and put them into use for a range of learning purposes, providing us with an evidence-based understanding of real-life classroom practice. The activity and support planning framework for the bilingual classroom can be utilised in future to re-evaluate how support and learning materials can be packaged in ways that match teachers' own practice. This should be of interest not only to teachers, but also to those who plan curricula, provide teacher support, to publishers, education planners and politicians.

The methodology used in this research demonstrates that if used diligently, with reference to theory and best practice, survey research can be highly effective at gaining access to respondents' opinions, practice and conceptualisations. Similarly, the use of statistical treatments such as correlation and factor analysis can provide highly illustrative data, beyond the simple picture painted by percentage scores. In addition, concurrent to the quantitative data collection in the present research, qualitative data was also collected from the teachers regarding each research question (not reported here) which assisted the Comunidad de Madrid in contextualising the quantitative data.

To conclude, whilst it is acknowledged that the population sample in this research is limited to teachers of the Madrid primary bilingual programme, it is anticipated that in many primary and secondary compulsory education contexts in Spain and elsewhere in the world, there are likely to be parallels in terms of how teachers conceptualise and use exam materials in their classrooms. If this is the case, there is likely to be a benefit to others involved in planning bilingual projects in choosing to replicate this survey approach, complemented by a qualitative strand involving classroom observation and face-to-face interviews with selected staff, augmenting the methodology as required with locally-relevant adaptations. In doing so, we can provide a bottom-up understanding of what really happens when we introduce speaking and listening assessments to the bilingual classroom.

13.6. References

British Association for Applied Linguistics. (2017). *Recommendations on good practice in Applied Linguistics* [Online]. Retrieved from: http://www.baal.org.uk/goodpractice_full.pdf [Accessed 28/02/17].

Cambridge ESOL. (2017a). *Cambridge English: Key (KET)* [Online]. Retrieved from: http://www.cambridgeenglish.org/exams/key/ [Accessed 28/02/17].

Cambridge ESOL. (2017b). *Cambridge English: Preliminary (PET)* [Online]. Retrieved from: http://www.cambridgeenglish.org/exams/preliminary/ [Accessed 28/02/17].

Krosnick, J. A. & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51,* 201–219.

Olsen, K. & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter (Ed.), *Improving surveys with paradata: analytic uses of process information* (pp. 43-72). Hoboken: Wiley.

Schwartz, N. (1997). Questionnaire design: the rocky road from concepts to answers. In L. Lyberg, P. Biemer, M. Collins, E. DeLeeuw, C. Dippo & N. Schwarz (Eds.), *Survey Measurement and Process Quality* (pp. 29-45). Chichester: Wiley.

Trinity College London. (2017). *Graded Examinations in Spoken English (GESE)* [Online]. Retrieved from: http://www.trinitycollege.com/site/?id=368 [Accessed 28/02/17].