EXAMINING THE CONTENT ALIGNMENT BETWEEN LANGUAGE CURRICULUM AND A LANGUAGE TEST IN CHINA

Matthew P. Wallace[a], Haijiao Ke[b]
([a]mpwallace@um.edu.mo; [b]jackke2006@foxmail.com)

[a]University of Macau
[b]Rongtai Elementary School, China

Abstract: This study examined the content alignment between an English as a foreign language skills curriculum and a provincial language test in China. When there is misalignment in the content between the standards of a curriculum and a test, conclusions about student abilities and teaching effectiveness can be questioned. To examine this, three categories of alignment were investigated using document analysis and expert judgment: categorical concurrence, range of knowledge correspondence, and balance of representation. Eight reviewers coded the curriculum and test items. Results showed that the curriculum aligned across the three criteria for the listening and reading skills. For the writing skills, the range of knowledge correspondence and balance of representation criteria were met, but categorical concurrence was not. The test did not include speaking items, so there was complete misalignment with that curriculum. The findings showed that the test partially aligned with the curriculum, suggesting that performance may not fully represent students' ability to meet the curricular standards. We recommend that future tests should comprehensively cover all of the content in the curriculum and when doing so to ensure there is a sufficient number of items measuring each objective. This would improve how accurately interpretations of student performance can be made.

Keywords: alignment, standards, language test

DOI: http://dx.doi.org/10.15639/teflinjournal.v34i1/116-135

An overlooked aspect of language assessment is the alignment between a language test and the educational curriculum that it purportedly measures. Alignment refers to the degree of overlap among the three types of curriculum in a learning context: intended curriculum, enacted curriculum, and assessed curriculum (Anderson, 2002; Webb, 1997). The intended curriculum describes the knowledge, skills, and competences that students are expected to achieve. It is presented at two levels: standards and objectives (Webb, 2007). Standards are broader descriptions of what learners are expected to do, and objectives are more specific and measurable descriptions of the content under each standard. In traditional learning contexts, like public schools, the intended curriculum is designed by experts at the national or local level (Kurz, 2011). Kurz (2011) explains that the enacted curriculum refers to what teachers do in the classroom to

achieve the intended curriculum, inclusive of their instructional methods and teaching materials. He defines the assessed curriculum as the content included on tests to evaluate how well students have learned the enacted curriculum. Kurz emphasizes that to ensure that students are given adequate opportunities to learn, it is important for teachers to teach what students are expected to learn and that the assessments accurately measure the content described in the curriculum standards. In other words, there should be alignment across all three types of curriculum.

There has been limited research attention given to alignment in language learning contexts, with few studies examining the alignment between the intended and assessed curriculum (Kong, 2015; Rouffet et al., 2022), the intended and enacted curriculum (Tekir & Akar, 2018; Umar, 2018; Wotring et al., 2021), and the enacted and assessed curriculum (Papageorgiou et al., 2020; Timpe-Laughlin, 2018). A consistent finding across these studies is that there is partial alignment among the curricula investigated, meaning that learners may not be taught or assessed on all of the knowledge, skills, and competences they are expected to develop. To our knowledge, only two studies (e.g., Kong, 2015; Rouffet et al., 2022) have examined the alignment between the intended and assessed language learning curriculum. However, the assessed curriculum in both of these studies was operationalized as classroom tests administered within schools. There is scant research on the alignment between an intended curriculum and a compulsory assessment designed to measure the foreign language learning standards, such as the Grade 9 English language proficiency test administered in mainland China, the current study's context. This could be problematic because the inferences that are drawn from test performance may not accurately reflect how well students meet the curriculum standards. This in turn could have strong effects on the decisions made about test performance (e.g., if applicants are admitted to a competitive university program based on scores from a misaligned test). Therefore, it is essential to ensure that the tests used for these high-stakes exams align closely with the curriculum standards they are expected to measure.

The number of studies examining curriculum alignment in language learning is extremely limited, though we speculate that it is frequently performed as part of an organization's (individual school or school system) internal curriculum review process. Keeping the results of these analyses internal may be a due to concerns about confidentiality of identities and proprietary material, a lack of confidence in the rigor with which the alignment examination is performed, or the potential harm that results showing misalignment may cause. Because the expectation is for curriculum to be closely aligned, any finding outside of that could cast doubt on the quality of the educational process in that context. We feel that despite all of these concerns, curriculum alignment should be consistently examined and publicly shared (including through research studies like this one). Doing so would ensure that learners are assessed on the content of the intended curriculum, allowing for interpretations about their knowledge and skills to be accurately made. Increasing the transparency of curriculum alignment would also hold key decision-makers accountable for their work to educational stakeholders. Motivated by these views, the current study examined the alignment between the intended and assessed curriculum

in the English as a foreign language (EFL) context in China, home to one of the largest populations of learners receiving formal EFL education in the world.

LITERATURE REVIEW

Curriculum Alignment

The pedagogical literature offers a number of models for examining curriculum alignment, most of which have been used to investigate alignment in traditional educational settings (e.g., public schools). One of the earliest frameworks proposed by Biggs (1996), called constructive alignment, theorized alignment as a key feature of curriculum design. Originally proposed to enhance the depth of student learning for university courses, the constructive alignment model has also been applied as a post hoc analytical model in varied contexts, including EFL (e.g., Rouffett et al., 2022; see below). To evaluate alignment using the model, content of the intended, enacted, and assessed curriculum is mapped onto one another and when misaligned elements are found (e.g., an assessment item does not match an objective), they are adjusted to match (Huet et al, 2009).

Alternative models of alignment expand beyond the single content dimension to include additional dimensions, including cognitive demand, depth of knowledge, accessibility, equity and fairness, and pedagogical implications, among others (see La Marca et al., 2000; Porter & Smithson, 2001; Webb, 1997, 1999, 2002, for alternative models). Though these models are more comprehensive in their treatment of alignment, a common dimension across all of them is content focus. Clearly, the alignment of the content across the curriculum is of high importance when determining degree of alignment. Therefore, the current study will focus on that dimension in examining alignment between the intended and assessed curriculum. At this point we must acknowledge that one limitation of the alignment literature is that the frameworks are quite dated, with a majority of them being proposed in the late 1990s and early 2000s. This may be due to the complexity of the models, with some having up to five dimensions and several sub-dimensions (e.g., Webb, 1997, 1999, 2002) making it challenging to empirically test them. Despite this limitation, the models continue to be frequently used to examine alignment in varied contexts.

The theoretical model supporting the current study is Webb's (2002) model of alignment, specifically the content focus dimension. This model has been used in several educational contexts to examine alignment across curriculum. We attribute this popularity to its comprehensive coverage of alignment and the explicit criteria that are laid out in the model to examine it. For the alignment of the content, Webb suggests that four criteria be met:

- Categorical concurrence: whether the assessment content corresponds with the content described in the curriculum standards.

- Depth of knowledge consistency: whether the cognitive difficulty of the assessment corresponds with that described in the curriculum standards.
- Range-of-knowledge correspondence: whether the breadth of knowledge measured on the assessment corresponds with that described in the curriculum standard.
- Balance-of-representation: whether the assessment covers a sufficient number of content objectives within each curriculum standard.

When there is sufficient overlap between the curriculum and assessment across these criteria, then there is a high degree of alignment[1].

Intended and Assessed Curriculum Alignment in EFL

Few studies have examined the alignment between language assessments and the intended curriculum. In the Chinese EFL context, Kong (2015) examined the alignment between English language standards (including listening, reading, and writing skills) and 10 achievement tests for Grade 10 students in five high schools across three semesters in mainland China. Kong used Webb's (2002) model of alignment and focused on the content focus dimension. He found that the categorical concurrence criterion was met for the listening and reading skills, but it was not met for writing skills. The results further showed that the depth of knowledge and balance of representation criteria were met for all three skills. Finally, Kong reported a weak alignment between objectives and test items for listening and reading skills on the range of knowledge correspondence criterion, but writing was fully aligned. Based on these results, Kong concluded that there was partial alignment of the content between the classroom tests and curriculum standards. Interviews with teachers in these schools revealed that test development was guided by a combination of the curriculum standards, their preferred pedagogical approach, and their attempt to improve student weaknesses.

Similar results were reported by Rouffet et al. (2022) in the Netherlands, who showed there was partial alignment among the curriculum standards, teaching practices, and classroom tests in the lower secondary foreign language learning context (approximated to be around the Common European Framework of References for Language A2 level). Using the constructive alignment model as a post hoc analytical framework, Rouffet et al. found that tests from 10 different schools predominantly focused on reading skills, which did not align well with Communicative Language Teaching principles outlined in the Dutch national curriculum. Their results further showed that language knowledge (e.g., measured by decontextualized vocabulary and grammar exercises) was more frequently tested than language skills, but that within the skills that were tested, productive skills (i.e., speaking and writing) were rarely evaluated. When they were assessed, tasks measuring productive skills did not meet the standards focusing on

[1] Curriculum alignment determined using the constructive alignment model (Biggs, 1996) appears to correspond with the Categorical Concurrence criterion in Webb's (2002) model.

communication that were outlined in the intended curriculum. Interviews with teachers revealed that they lacked the resources (time and expertise) to adequately develop their teaching and testing materials, forcing them to rely on their assigned teaching materials (e.g., the textbook) for support.

Another explanation offered by Rouffet et al. for the misalignment is that the nature of the classroom tests was the result of washback from national foreign language tests in the Netherlands, on which language knowledge and reading skills are of great importance. Teachers are tasked with helping their students succeed on these tests, so teaching and assessing them on similar skills and knowledge in preparation for the tests is an understandable approach. Both of these explanations offered by Rouffet et al. may also help explain Kong's findings because he also examined classroom tests designed by teachers. It is possible that a high stakes assessment, like the one we examine in the current study, would be better aligned with the curriculum standards because the test designers have more resources available to them than teachers would in individual schools.

Another line of inquiry has examined the alignment between a large-scale proficiency test and a set of benchmarks describing language performance (e.g., CEFR; Harsch & Hartig, 2015). The benchmarks represent the intended curriculum, though they are seldom adopted wholly as curricular standards and objectives, mainly because they are too general to be applied to any one context. While this research is helpful for benchmarking test scores with the external criteria, it does little to determine how well tests used in traditional educational settings match the curriculum they are designed to measure.

Two studies recently attempted to address this issue, but ended up examining the alignment between the enacted and assessed curriculum instead. Timpe-Laughlin (2018) examined the alignment in content between a large-scale English language proficiency test and the mandated English language curriculum in Berlin, Germany. The test examined was the TOEFL Junior Standard test, designed for 11 to 17 year old EFL learners, but it was not directly tied to a specific learning curriculum. The curriculum that was examined came from textbooks used throughout the learning context because the intended curriculum was not specific enough for alignment judgments to be made. Papageorgiou et al. (2020) examined the content alignment between the TOEFL Primary test (for 8-12 year old learners) and that of an online learning program for Chinese elementary school EFL learners studying English one-on-one with an instructor in the US. Similar to Timpe-Laughlin, the intended curriculum was not clear enough for alignment judgments to be made, so the learning materials, representing the enacted curriculum, were examined as operationalized versions of the curriculum. Both studies reported good alignment between the enacted and assessed curriculum. However, a limitation of these studies is that the examined test was a large-scale proficiency test that was developed by an external organization and not specifically for the learning curriculum in that specific context.

Context of the Current Study

The aim of the current study was to examine the alignment between English language curriculum and the skills measured on a language test in mainland China. At the time of the current study in 2021, the curriculum regulations in China, such as the Basic Education Curriculum Reform Outline (Ministry of Education of the People's Republic of China, 2001) clearly state that the instruction and assessment in schools should follow the guidance of the national curriculum standards. Similarly, the National English Curricula Standards (NECS) for Compulsory Education (Ministry of Education of the People's Republic of China, 2011) indicates that English assessments should be in line with the objectives and requirements listed in NECS[2]. This means that all educational contexts throughout mainland China should follow the nationally mandated curriculum and that assessments that are developed to evaluate it should closely align with it. Individual provinces are allowed to develop their own tests for this purpose, but the test content should closely align with the NECS standards. Despite these goals being explicitly stated in the educational literature, the degree of alignment between the tests and curriculum has been under-researched.

To address this limitation, the current study analyzed the alignment between NECS (Ministry of Education of the People's Republic of China, 2011) and the provincial English test for Grade 9 students in Guangdong province of mainland China. Guandong was selected because it is a well-populated (about 127 million people) and economically developed province in southern China. Many resources are devoted to language education there, so it is essential to ensure that tests mandated by the school system closely align with the curriculum there so as to ensure accurate interpretations of student abilities can be made.

In Guangdong, students begin to formally learn English as a foreign language from Grade 3 in primary school (8 years old) with an average instruction period ranging from three hours to four hours per week throughout their academic career. The intended curriculum is measured with a provincial test at three points during the educational process—6th grade, 9th grade, and 12th grade. By grade 9, students are expected to meet the curriculum standards for Level 5 of the NECS (see Appendix 1). To evaluate how well students meet the standards, they sit a provincial test before their graduation to senior high school. The test result has an important impact on students' future academic development because it affects their high school enrollment. If they perform well, they may be admitted to a strong high school, but if they perform poorly, they may be admitted to schools with a lesser reputation for learning. Therefore, it is important to determine the degree of alignment between the test and curriculum to ensure that the inferences made about the students' language ability are accurate.

[2] At the time of this study in 2021, the 2011 standards were still being used in the school system.

Research Questions

To address these limitations, the current study aimed to answer the following research question: How well does the National English Curricula Standards in mainland China align with the content of the Guandong Province language test for Grade 9 students? Specifically, how does the content align based on three criteria of categorical concurrence, range of knowledge correspondence, and balance of representation[3]?

METHOD

This study examined the alignment between the language curriculum and language test for Guangdong province in southern China. To do so, the language curricular standards for mainland Chinese schools and a provincial language test were collected from publicly available sources. Characteristics of the two documents are described in the following sections.

Curriculum Standards

The curriculum standards for this study are the National English Curricular Standards (NECS; Ministry of Education of the People's Republic of China, 2011) for Compulsory Education. The 2011 version of NECS has been used as a guideline for English education in elementary schools and secondary schools across mainland China since its inception. The aim of NECS is to develop students' overall language competency and build the foundation for their all-round development. The overall English language competency goals were divided into nine levels. The NECS listed out the descriptions of the goals from Level 1 to Level 5 (see Appendix 2). These descriptions are general goals of each level of English competency and are further divided into five subcategories of standards. These subcategories include:

1. Language skills
2. Language knowledge
3. Learning attitudes
4. Learning strategy
5. Cultural awareness

Students are expected to attain certain levels of English competency at different grades from elementary school to high school. Most schools in mainland China provide formal English education from Grade 3 in primary school, giving students three years to acquire the second level of English competency by the end of Grade 6. Students at Grade 9 should be able to reach Level 5 of English competency and students at Grade 12 should be able to reach Level 7.

---

[3] The depth of knowledge consistency criterion was not examined because the intended curriculum standards (see Appendix 1) did not contain sufficient information for this to be investigated.

The current study focused on the first aspect, the language skills standards. The description of the language knowledge standard is too broad to evaluate its alignment with the test accurately and the attitudes, strategies, and cultural awareness are not directly measured by the test. Therefore, this study only examined the alignment between the language skills standards in the NECS and the language test for students at Grade 9. The skill standards for Level 5 are shown in Table 1.

Table 1. Skill Standards for Level 5 (Ministry of Education of the People's Republic of China, 2011)

Skill	Standard
Listening	Students would be able to listen to and understand the narrations on familiar topics and join in the discussion
Speaking	Students would be able to exchange information and express their opinions on related topics of daily life
Reading	Students would be able to read and understand the English books, newspapers, magazines at appropriate level, to overcome the barriers of unfamiliar words, and get the main idea. Students would be able to use appropriate reading strategies according to different reading purpose
Writing	Students would be able to write and correct short essay independently

The language skill standards include four categories of objectives for listening, speaking, reading, and writing skills. Each category lists out several specific objectives. Table 2 shows the descriptions of the listening skill objectives at Level 5. The objectives for the speaking, reading, and writing skills are provided in Appendix 1.

Table 2. Listening Objectives at Level 5 (translated from the figure in NECS, Ministry of Education of the People's Republic of China, 2011)

Level	Description of the listening objectives
Level 5	Students: 1. Can understand the intention of the speakers according to the pronunciation and intonation. 2. Can listen and understand conversations on familiar topics, and get information and viewpoints from the conversations. 3. Can deal with unfamiliar words and get the main idea using contextual clues. 4. Can listen and understand stories and narratives spoken at almost natural speed, and understand the cause and effects relationships of the story. 5. Can respond to what they hear in a suitable way. 6. Can take simple notes on paragraphs they hear.

Provincial Language Test

The test that was analyzed is the provincial English language test for students at Grade 9 in a province of mainland China. The test is administered across the province every June except in cities that use their own tests. This test can be used as a high-school entrance exam; therefore, it is also considered as a high-stakes test. The test information is open to the public after its administration (see Appendix 3 for detailed information about the test).

The examined test was administered at the end of the second semester for students at Grade 9 in 2019. It consisted of six parts and a total number of 86 test items with a total score of 120 points. Students were given 100 minutes to complete the test. The key characteristics of the test are presented in Table 3. Parts 1 and 2 measure listening (30 points) and grammar skills (15 points), respectively, with one point per item. Reading skills were measured in Parts 3-6 and made up a large portion of the score (60 points). Writing skills were assessed with a single task item in Part 6, worth 15 points.

Table 3. Provincial Test Characteristics

Parts	**Sections/Response Format**	**Total Number of items/Score**
Part 1: Listening	Section A: Multiple choice (5 items) Section B: Multiple choice (10 items) Section C: Multiple choice (10 items) Section D: Fill in the blanks (5 items)	30 items/30 points (one point for each item)
Part 2: Grammar	Single section: Multiple choice (15 items)	15 items/15 points (one point for each item)
Part 3: Reading	Single section: Multiple choice (10 items)	10 items/10 points (one point for each item)
Part 4: Reading	Section A: Multiple choice (5 items) Section B: Multiple choice (5 items) Section C: Matching answers to questions. (5 items)	15 items/30 points (two points for each item)
Part 5: Reading	Single section: Fill in blanks. (10 items)	10 items/15 points items (1.5 points for each item)
Part 6: Reading and writing	Section A: Fill in blanks (5 items) Section B: Open question. (1 item)	5 items/5 points 1 task item, 15 points

Alignment Procedures

This study examined the content alignment between language skill standards and their objectives in the NECS with a provincial language test based on three criteria: categorical concurrence, range of knowledge correspondence, and balance of representation. After receiving ethical clearance for the study, eight coders (3 males, 5 females) were recruited to match the

items on the test with the standards and objectives in the curriculum. All of the coders had prior experience with language teaching and assessment in the target context (mainland China) and attained a Master's degree in second language acquisition or English linguistics. Because of the researchers' knowledge and experience with the specific learning context in which data was collected, they were not included as coders to reduce any potential biases and avoid unintentional influence on the coding panel scores.

The coders were first introduced to the test and curriculum and then completed a normalization process. To introduce the content, the coders were provided with a video explaining the standards and test, after which, the researcher held an online discussion to answer questions and clarify understanding of the documents. The normalization process was undertaken to ensure that the coders would use similar criteria to evaluate the match between the test items and the curriculum standards and objectives. Working individually, the coders matched five test items selected from each section of the test (items 1, 31, 46, 56, and 86) with their corresponding objective(s). To indicate which objective(s) was related to each test item, they used a matrix developed for this study (See Appendix 4 for details). If a test item was associated with an objective, the reviewers would give it a tick. One test item could be related to more than one objective. Table 4 shows an example of coding. After the individual coding, they discussed differences in their codes until they came to an agreement.

After the normalization process, the remaining items on the test were coded using the coding matrix. The coding work was carried out on a digital document and completed individually. Upon completion, the coders sent the document back to the researchers for data analysis.

Table 4. Example of the Coding Form

	Listen to the first conversation, answer question 6. 6. Who are the speakers going to help? A. A little kid B. A young lady C. An old man
Listening Objective 2: Can listen and understand conversations on familiar topics, and get information and viewpoints from the conversations.	√

Alignment Judgments

Categorical Concurrence

For the assessment to have categorical concurrence with the standards, the test items should measure the corresponding content described in the content standards. For the current study's test, this meant determining how many items measured each of the language skill standards (i.e.,

how many items measured the listening, speaking, reading, and writing standards). In line with Webb (1999), calculating the categorical concurrence was a multi-step process. First, the number of test items per standard was totaled for each examiner. This meant that the items that were matched with individual objectives under each standard (e.g., the seven listening objectives) were totaled for that standard (e.g., the listening standard). Webb (1999) proposed that at least six items should be related to the content from one standard to achieve sufficient categorical concurrence.

After the total number of items per standard was determined for each coder, the minimum level of coder agreement was inspected. The coder agreement was set to be a majority, or at least five out of eight coders agreeing that an item (e.g., item 1) measured a standard (e.g., listening skills standard). For example, if five reviewers considered item 18 to measure a standard, then this item would be deemed to have measured that standard. If less than five reviewers determined that item 18 measured a standard, then it was not considered to have measured that standard. The categorical concurrence was determined by totaling the number of items for each standard that met the 5/8 coder agreement criteria. If a standard had more than six items, then the test would be considered to have categorical concurrence of that standard.

Range of Knowledge Correspondence

For the test to have range-of-knowledge correspondence with the language skill standards (listening, speaking, reading, writing), items should measure at least 50 % of objectives under one standard. For example, for a standard that has six objectives, the test should have at least one item for at least three of the objectives to have range-of-knowledge correspondence. To determine this, a similar coder agreement procedure as the categorical concurrence described above was followed. This time the items measuring the individual objectives under each standard were tallied. If there was a 5/8 coder agreement that an item measured an objective, then that item was deemed to measure that objective. The range-of-knowledge correspondence was met if at least 50% of the objectives under a standard were measured by at least one test item.

Balance of Representation

For the assessment to have a balance-of-representation with the standards, the test items related to the objectives should be equally distributed. This means that if all of the objectives within a standard have an equal number of test items measuring them, then the balance of representation is very good. To determine this, the number of items for each objective calculated for the range-of-knowledge evaluation (described above) are inserted into a revised formula proposed by Flowers et al. (2006). The original formula for this index is:

$$Balance = 1 - \left(\sum_{i=1}^{k} \left|\frac{1}{O} - \frac{I_k}{H}\right|\right)/2, \text{(Flowers et al, 2006).}$$

Here, O refers to the total of objectives which have corresponding test items. I_k refers to the numbers of items which are related to the objective *k*. *H* refers to the total number of items related to the content standard. The index ranges from 0 to 1. Webb (1999) has stated that when the index is bigger than 0.7, there is an acceptable level of balance of representation. However, a limitation of this formula is that it assumes that each item on a test shares the same point value. However, the values of items differ on the provincial test examined in this study (and most language tests). On our test, the single writing task item is worth 15 points, while a fill-in blank item is worth one point. Though they are measuring related objectives in the standards, the writing task item is given more value on the test (15 points > 1 point). Therefore, we adapted the formula to account for this important variable:

$$Balance = 1 - (\sum_{i=1}^{k} \left|\frac{1}{O} - \frac{I_k}{H}\right|)/2, I_k$$

(the number of items measuring the objective k) and *H* (the total number of items measuring objectives in the standard) would be replaced as the total score of these test items. The interpretation of the result is still in line with Webb's (1999) recommendation, that an index greater than 0.7 indicates an acceptable level of balance-of-representation.

Reviewer Agreement

After the data was collected, Fleiss' kappa was calculated to assess the agreement among the 8 reviewer's coding on 71 items (See Appendix 4). Fleiss' kappa (Fleiss et al., 2003) is used to measure the agreement level between two or more raters when the assessment method uses categorical scaled data. Fleiss' kappa ranges from -1.0 to +1.0 (+1 indicates perfect agreement). The interpretation of the magnitude of agreement shown by Fleiss' kappa is similar to the Cohen's kappa (Fleiss et al., 2003). Agreement is excellent when it is above 0.75, fair to good between 0.40 to 0.75, and poor below 0.40.

FINDINGS AND DISCUSSION

The objectives in the speaking and performing standards were automatically considered as not measured by and subsequently misaligned with the assessment because the test did not include any items that required speaking or performance. Therefore, the results presented and discussed in the following sections will be for the listening, reading, and writing standards. The results of the Fleiss' kappa examining reviewer agreement showed that there was good agreement among the reviewers' judgments, with most of the overall kappa ranging from 0.40 to 0.75 (see Appendix 5).

Categorical Concurrence Alignment

The number of items on the test that measured each discrete language skill standard are presented in Table 5. To achieve categorical concurrence alignment, there would need to be at least six items per standard. The results show that the mean number of items measuring the listening and reading standards is 30.00 and 38.50, respectively; far exceeding the minimum number of 6 items required by Webb's (1997, 1999, 2002) model. This means that the criterion of categorical concurrence between listening and reading skills standards and the test is met. We may therefore claim that the test has enough items covering the listening and reading skills standards, meaning that students' listening and reading skills are assessed adequately enough for accurate inferences about their listening and reading abilities to be made.

However, the mean number of items corresponding to the writing standard is only 1.00. This indicates that the test has an insufficient number of test items (less than six) measuring students' writing skills, and therefore does not meet the criterion of categorical concurrence. If we take a closer look at the test, there are six items requiring students to write (items 81 to 86). Items 81 to 85 are short-answer productive items, which require students to demonstrate their comprehension of a reading text by writing the answers in short sentences or only sentence fragments (Appendix 3). Within the writing skills standards, three of the objectives (Objectives 2, 4, 5) require students to write paragraphs, simple descriptions or passages, use linking devices, and organize materials and use them in writing (see Appendix 1). However, successfully completing the short-answer test items (81 – 85) would not require the skills in the objectives to be demonstrated. The only test item that was able to successfully evaluate these objectives was item 86, the task item.

An insufficient number of test items covering writing objectives means that students' writing skills are not assessed sufficiently. This underrepresentation might undermine the validity of the test, as it would lead to an inaccurate interpretation of what students are able to do with their writing skills. Further, the test performance may insufficiently reflect the teaching of writing skills because students did not have enough opportunities to demonstrate that they could meet the standards. This misalignment between the standards and the test might have some washback effect on the teaching of writing skills. Studies (e.g., Rouffet et al., 2022; Stecher et al., 2004) have shown that teachers, particularly in mainland China (Kong, 2015; Qi, 2005; Zou & Xu, 2017), tend to focus their instructional time on what is evaluated on assessments, reflecting this washback. If a skill is not evaluated, teachers in these studies reportedly do not give it as much focus as the skills that are measured. Therefore, the misalignment on the current study's test could lead to teachers giving writing skills insufficient classroom attention, thereby limiting students' opportunity to learn the intended curriculum (Kurz & Elliott, 2011). For our study, we may only speculate about this effect based on the misalignment results. We do not know whether teachers were teaching based on the standards or the assessments. To see whether

teachers are teaching according to the curricular standards, researchers are encouraged to examine the alignment between intended curriculum and enacted curriculum.

Overall, these results generally align with those reported by Kong (2015) in the mainland Chinese context. In his study (and ours), both listening and reading skills met the categorical concurrence criterion, but the writing skill did not. Though the context of the two studies was quite different—ours of a provincial test in southern China and Kong's of classroom achievement tests in northern China—both contexts used the same curriculum standards as the intended curriculum. Therefore, it is reasonable to expect that the provincial test in Kong's study's context measured similar skills as the provincial test in our study. Interpreting the two studies together suggests that there may be some evidence of a washback effect from the provincial test to classroom tests. The writing skills not meeting the categorical concurrence criterion on the provincial test may have caused it to receive less attention on classroom tests throughout mainland China.

Table 5. Means and standard deviations for number of test items for each language skill standard, percentage of standards covered by test items, and balance index for the standards

Standards (# objectives per standard)	**# items per standard**	**% of standards covered by items**	**Balance index**
	Mean (SD)	Mean (SD)	Mean (SD)
Listening skills (7)	29.38 (1.17)	75.00% (12.60)	0.72 (0.10)
Reading skills (7)	38.50 (2.45)	64.28% (10.80)	0.78 (0.36)
Writing skills (5)	1.00 (0.00)	57.50% (7.07)	1.00 (0.00)

Range of Knowledge Correspondence Alignment

The means for the percentage of standards that are covered by the test items (representing range of knowledge correspondence) are presented in Table 5. The mean percentage of objectives having corresponding items on the test for listening, reading, and writing skills standards is 75.00%, 64.28% and 57.50% respectively. These results show that the test covered more than 50% of the objectives in the listening, reading, and writing skills standards, which meets Webb's (1997, 1999, 2002) minimum requirement of 50%. This means that the test required students to meet more than half of the objectives in the listening, reading, and writing standards. From these results, we can claim that the students were required to have a wide range of listening, reading, and writing skills in order to perform well on this test. These findings contrast with Kong's (2015) results, who reported weak range of knowledge correspondence for tests measuring listening and reading skills (40% - 50%), and one year's test measuring writing, but did not meet the range of knowledge correspondence for two years of writing tests. An explanation for this divergent result is that the teachers in Kong's study decided that certain

aspects of the intended curriculum deserved more attention than others. Interviews with the teachers revealed that test content was determined by the teacher philosophy and student needs, in consultation with the intended curriculum. So, it is possible that the teachers generating the classroom tests intentionally focused on specific aspects of the curriculum, while not assessing others. The difference in test purpose between ours and Kong's study may also explain the difference in result. Kong's assessments were classroom achievement tests designed to evaluate what had been taught in the class that term, and not the full intended curriculum, like our study's provincial test was intended to do. Classroom tests are designed to evaluate learning done in the classroom, and not to assess the full curriculum.

Because the range of knowledge correspondence criterion is met for the listening, reading, and writing skill standards in the current study, the stakeholders of this test can get more valid information about students' acquirement of these language skills. These results may also accurately reflect how well students were taught in developing different language skills. For example, the results show that more than 60 percent of reading skills objectives were covered by the test. This means that at least four objectives in the reading skills standards were measured. Therefore, if students perform well on the reading section, the test score can validly reflect whether students have met the objectives of the intended curriculum through their language learning.

Though the results showed that the criterion of range of knowledge correspondence is well met, not all of the objectives under each of the three skill standards were covered on the test. For example, the fifth objective in the reading skills standard states "Students can use dictionaries and other reference material to carry out reading." The sixth objective in the same category states "Students can read material other than that included in the textbook, totaling over 150, 000 words" (See Appendix 1). A reason for this may be that the curriculum designers did not make a distinction between enabling objectives (process-oriented description of how learners can develop their skills) and terminal objectives (product-oriented description of what learners should be able to do with the skill; Brown & Lee, 2015). Enabling objectives would be more appropriately assessed in daily teaching practice rather than on a test because they are process-oriented objectives. The terminal objectives measured on the test are those that are more product-oriented. Not accounting for this distinction between objective types highlights another limitation in the alignment literature. Future studies examining alignment are encouraged to make such a distinction when examining the curriculum standards. Alternatively, the curriculum standards can be revised to make this difference explicit so that test developers can ensure which objectives should be measured on a test and which others should be evaluated in the learning context.

Balance of Representation Alignment

The balance indices for the listening (0.72), reading (0.78), and writing skills (1.00) standards are presented in Table 5. All three indices were above 0.70, which met the minimum requirement. These results indicate that for the standards that have at least one item representing one objective on the test, there is a balanced distribution of test items for the objectives within each standard. This means that for this study's test, students' listening, reading, and writing skills were assessed in a balanced way. This is especially notable given the lack of a distinction between enabling and terminal objectives in the curriculum we noted earlier. These findings are consistent with Kong's (2015) results showing that listening, reading, and writing skills standards were acceptably represented on classroom achievement tests.

The balance of representation improves the validity of the test because the stakeholders can receive a more accurate interpretation of what students can do with their language skills from the test. It is important for a test to have good balance of representation because only when multiple objectives are represented by the test can it give a more comprehensive measure of a standard. If only one objective per standard were represented, the conclusions regarding whether the students sufficiently met that standard would be inaccurate.

CONCLUSIONS

Overall, the results showed the assessed curriculum partially aligned with the intended curriculum of the language skill standards. The listening and reading skills met the categorical concurrence, range of knowledge, and balance of representation criteria for alignment. The writing skills met the range of knowledge and balance of representation criteria, but failed to meet the categorical concurrence criterion. There was complete misalignment between the test and the speaking skills and performance standards because there was no speaking or performance section on the test. To our knowledge, this is the first study that has examined the alignment between a provincial English language test and the national language curriculum in mainland China. Only one other study, Kong (2015), has examined alignment in the Chinese EFL context, but Kong examined alignment between classroom tests and the intended curriculum for grade 10 students in a different province than that of the current study. The differing aims, contexts, and year of education between our study and Kong's make comparisons challenging to make, but the dearth of research highlights the need for more alignment research in the Chinese EFL context; especially given the high-stakes nature of the tests delivered throughout the educational process (provincial tests delivered at grades 6, 9 and 12 to determine educational advancement) and the impact that interpretations of test performance may have on academic futures of students.

Our findings have important implications for future test design. Firstly, to more completely align with the intended curriculum, we reinforce the advice made in the literature that all standards of the curriculum be evaluated on an assessment. Incomplete coverage of the content

would make it challenging to accurately determine if students were able to meet curricular standards. For the current study, this involves adding a speaking section to the test. Second, we suggest that tests include a sufficient number of items for each standard in the intended curriculum. This provides learners with adequate opportunities to demonstrate that they can meet the standards of the intended curriculum. A final implication of our study is that there should be a clear distinction between enabling and terminal objectives within the standards and to ensure that tests of the curriculum align with the terminal objectives. To do this, we recommend utilizing Huet et al.'s (2009) curriculum mapping procedure (also see Biggs & Tang, 2011, for illustration) during the curriculum design process, which involves matching the standards and objectives of the intended curriculum with the course content and assessments.

We acknowledge that there are limitations in the study. Firstly, this study focused solely on the alignment between skill standards and the test. Of the other four categories of standards—language knowledge, learning attitude, learning strategies, and cultural awareness—future studies may consider examining the alignment with language knowledge. The test alignment with the other three categories may be challenging because they are less directly observable than skills or knowledge. Another limitation is that the study did not include perspectives of stakeholders involved in the curriculum design, teaching, and assessment. Doing so may have provided some insider views on the degree of alignment examined in the study. A final limitation is the narrow scope of the study only examining alignment of the national curriculum from one provincial test. Future studies may consider expanding the scope to include alignment examination of tests at earlier or later stages of learning and in other areas of China. A comparison of alignment among multiple locations may elicit interesting insight into how different provinces approach test development to measure the curricular standards.

Despite these limitations, the current study makes an important contribution to the language teaching and assessment literature. Ensuring that tests align with the intended curriculum allows accurate interpretations to be made about the students (if they have the requisite abilities) and the quality of the instruction (if teachers are effectively teaching the content) and test design (if the test adequately measures what it should be measuring). These interpretations can directly affect the decisions made about the learners (if they should progress to the next stage of learning), teachers (if they should require more training and if they are effectively teaching the content), and test designers (if the test should be revised). In sum, alignment can give important evidence regarding the consequences associated with a test administration (Messick, 1989).

REFERENCES

Anderson, L. W. (2002). Curricular alignment: A reexamination. *Theory into Practice, 41*(4), 255–260. https://doi.org/10.1207/s15430421tip4104_9

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*, 347-364. https://doi.org/10.1007/BF00138871

Biggs, J. & Tang, C. (2011). *Teaching for quality learning at university*. McGraw-Hill.

Brown, H. D., & Lee, H. (2015). *Teaching by principles: An interactive approach to language pedagogy* (4th ed.). Pearson.

Fleiss, J. L., Paik, M. C., & Levin, B. (2003). *Statistical methods for rates and proportions* (3rd ed.). John Wiley & Sons, Inc.

Flowers, C., Browder, D. & Ahlgrim-Delzell, L. (2006). An analysis of three states' alignment between language arts and mathematics standards and alternate assessments. *Exceptional Children, 72*, 201-215. https://doi.org/10.1177/001440290607200205

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly*, *12*(4), 333–362. https://doi.org/10.1080/15434303.2015.1092545

Huet, I., Oliveira, J. M., Costa, N., & de Oliveira, J. E. (2009). The effectiveness of curriculum maps of alignment in higher education. In C. Nygaard, C. Holtham, & N. Courtney (Eds.), *Improving students' learning outcomes* (pp. 275-287) Copenhagen Business School Press.

Kong, S. (2015). 高中英语试题与课程标准的一致性研究[*A study on alignment between achievement tests in high schools and the curriculum standards*]. [Master's thesis, Qufu Normal University]. https://cdmd.cnki.com.cn/Article/CDMD-10446-1015408098.htm

Kurz, A. (2011) Access to what should be taught and will be tested: Students' opportunity to learn the intended curriculum. In S. Elliott, R. Kettler, P. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students* (pp. 99-129). Springer. https://doi.org/10.1007/978-1-4419-9356-4_6

Kurz, A., & Elliott, S. N. (2011). Overcoming barriers to access for students with disabilities: Testing accommodations and beyond. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques* (pp. 31–58). Information Age Publishing.

La Marca, P., Redfield, D., Winter, P., Bailey, A., & Despriet, L. (2000). State standards and state assessment systems: A guide to alignment. Council of Chief State School Officers.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). American Council on Education and Macmillan.

Ministry of Education of the People's Republic of China. (2001). *基础教育课程改革纲要(试行)* [The compendium for curriculum reform of basic education (Trial ed.)]. Retrieved March 17, 2020, from the China Education http://old.moe.gov.cn//publicfiles/business/htmlfiles/moe/s8001/201404/xxgk_167343.html

Ministry of Education of the People's Republic of China. (2011).*义务教育英语课程标准(2011年版)*[English curriculum standards for compulsory education (2011 version)]. Beijing Normal University Press.

Papageorgiou, S., Xu, X., Timpe-Laughlin, V., & Dugdale, D. M. (2020). *Exploring the alignment between a curriculum and a test for young learners of English as a foreign language* (Research Memorandum No. RM-20-08). Educational Testing Service.

Porter, A. C., & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. Fuhrman (Ed.), *From the Capitol to the classroom: Standards-based reform in the states. One Hundredth Yearbook of the National Society for the Study of Education* (pp. 60–80). University of Chicago Press.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*(2), 142–173. https://doi.org/10.1191/0265532205lt300oa

Rouffet, C., van Beuningen, C. & de Graaff, R. (2022). Constructive alignment in foreign language curricula: An exploration of teaching and assessment practices in Dutch secondary education. *The Language Learning Journal*, https://doi.org/10.1080/09571736.2022.2025542

Stecher, B., Chun, T., & Barron, S. (2004). The effects of assessment-driven reform on the teaching of writing in Washington state. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 53-71). Routledge.

Tekir, S., & Akar, H. (2018). The current state of instructional materials education: Aligning policy, standards, and teacher education curriculum. *Educational Sciences: Theory & Practice*, *19*(1), 22-40. https://www.doi.org/10.12738/estp.2019.1.043

Timpe-Laughlin, V. (2018). *A good fit? Examining the alignment between the TOEFL Junior® Standard test and the English as a foreign language curriculum in Berlin, Germany* (Research Memorandum No. RM-18-11). Educational Testing Service.

Umar, H. (2018). A study of English language teachers' reading skills activities and their alignment with the curriculum objectives. *Journal of Research in Social Sciences*, *6*(1), 20-40.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. National Institute for Science Education. https://files.eric.ed.gov/fulltext/ED414305.pdf

Webb, N. L. (1999). *Alig*nme*nt of science and mathematics standards and assessment in four states*. National Institute for Science Education. https://files.eric.ed.gov/fulltext/ED440852.pdf

Webb, N. L. (2002, April). *An analysis of the alignment between mathematics standards and assessments for three states* [Paper presentation]. American Educational Research Association Annual Meeting, New Orleans, LA, USA. https://www.researchgate.net/publication/252605969_An_Analysis_of_the_Alignment_Between_Mathematics_Standards_and_Assessments_for_Three_States

Wotring, A., Chen, H., & Fraser, M. (2021). Exploring curriculum alignment through syllabus document analysis: From national language policy to local ELT practice. *Iranian Journal of Language Teaching Research*, *9*(2), 57-72. https://ijltr.urmia.ac.ir/article_121045.html

Zou, S. & Xu, Q. (2017). A washback study of the Test for English Majors for Grade Eight (TEM8) in China: From the perspective of university program Administrators. *Language Assessment Quarterly*, *14*(2), 140-159. https://doi.org/10.1080/15434303.2016.1235170

APPENDIX

Appendices 1-5:

https://drive.google.com/drive/folders/1JD-e_D6ltFTzCmeAiIgnz9H3SAhz2xlZ