Language Testing in Asia

RESEARCH Open Access

Assessing the range of cognitive processes in the Hong Kong Diploma of Secondary Education Examination (HKDSE)'s English language reading literacy test

Pok Jing Ho*

*Correspondence:
hpj@alumni.stanford.edu

Stanford Graduate School of Education, 485 Lasuen Mall, Stanford, CA 94305, USA

Abstract

This study examines the range of cognitive processes assessed in the English language reading literacy test as part of the Hong Kong Diploma of Secondary Education Examination (HKDSE), the secondary school exit test in Hong Kong. Prior studies have suggested that higher order cognitive processes are often undermined in high-stakes tests and classrooms, due to what is called the "washback effect." This prompts the following hypothesis: despite its top-performing education system, Hong Kong has failed to provide a versatile collection of test items that involve complex reasoning. To investigate the cognitive demands of the test, as well as their relationship with student performance, this study maps the test items with reference to the cognitive levels assessed using in-depth document analysis. ANOVA is used to statistically determine differences in accuracy rates to supplement the analysis throughout. Results indicate that the test has placed an overwhelming emphasis on lower order cognitive processes over the past decade (2012–2019) and that items assessing higher order cognitive skills are, as expected, met with statistically significantly poorer performance in the test. Implications for future revision of the test and curriculum policy are discussed.

Keywords: Cognitive processes, High-stakes testing, HKDSE, Large-scale assessment, Reading literacy, Washback effect

Introduction

Students from Hong Kong have consistently given stellar performances in the Programme for International Student Assessment (PISA) in mother tongue reading literacy (Organisation for Economic Co-operation and Development, 2019a). In contrast, students' performance in second language reading literacy gives the city-state little cause to celebrate. Just over half of the test-takers (52.4% in 2019) who sit the local English reading literacy test, the Hong Kong Diploma of Secondary Education Examination (HKDSE) in the English language, attain the required level to qualify for local university admission (Hong Kong Examinations and Assessment Authority, 2019).

While students from Hong Kong have long held a local reputation as being rote learners and lacking in higher order cognitive skills, scholars have warned against such stereotypes in determining learner characteristics (Biggs, 1991; Ho, 2009).

This is not to dismiss the significance of lower order cognitive processes in second language reading literacy tests. Memory recall of specific, discrete information helps readers identify the fundamentals of the text (Bloom, 1956), such as characters, setting, and sequencing (Hess, 2004). Furthermore, lower order cognitive processes are inevitably involved in more complex psychological processes, rendering them indispensable. However, higher order cognitive processes are equally essential in assessment settings, in that they are "skills that support transferable learning" (Darling-Hammond et al., 2013, p. 3).

Limited emphasis on higher order cognitive processes could have detrimental implications on students' cognitive development due to what is called the "washback effect:" the effect of testing on the curriculum, pedagogy, and student learning taking place in classrooms and schools (Cheng, 1997). In the case of a traditional relationship between curriculum, instruction, and assessment, the curriculum serves to determine what the teacher should teach (hence *instruction*) and which learning outcomes students should manifest (hence *assessment*). However, the effect of washback could be robust enough to prompt the teacher to override the communicative values in language learning for the sake of test preparation. In this case, language learners are situated in negative washback, learning with a focus on the test and prompted to focus disproportionately on skills that are assessed by the test (Prodromou, 1995). In other words, if there is indeed a skew in the level of cognitive processes in the assessment, there is bound to be a similar skew in teaching and learning. The fact that students in Hong Kong have been found to share the Confucian heritage culture (Biggs, 1996; Ho, 2009), which is partially characterized by reliance on authoritarian sources of knowledge on the learners' part, might further exacerbate the consequences.

While some studies have addressed the consequences of washback in language testing (Cheng, 1997; Prodromou, 1995), few have addressed the issue by examining the actual test materials. Document analyses of the tests, in particular, would allow us to determine the range of cognitive processes being assessed and whether they are a possible source of negative washback. This study articulates the relationship between the range of cognitive processes assessed and the manipulation of test difficulty as a complicating factor, as illustrated by examining the test items and their accuracy rate with ANOVA. In this study, the accuracy rate is defined as the percentage of test-takers who answered the item correctly; in line with definitions put forth by Bloom (1956), cognitive processes are defined as processes of knowledge recall and the development of intellectual abilities and skills.

Background

Hong Kong boasts a primarily monoethnic population (92% ethnic Chinese) (Bacon-Shone et al., 2015). This ethnic composition, together with an absence of urban-rural differences, has been conducive to an effectively centralized education system. For three decades, Hong Kong's education system had closely followed Britain's 6-5-2-3 system:

6 years of primary education, 5 years of secondary education followed by the Certificate of Education Examination, 2 years of pre-university secondary education followed by the Advanced Level Examination, and 3 years of university education (Marsh & Lee, 2014). The British colonial government implemented the 9-year compulsory education policy in 1978, guaranteeing free primary education and junior secondary education. The Government of Hong Kong extended it to 12 years in 2007 to cover the entire period of secondary education and aimed for it to coincide with the revised 6-3-3-4 education structure. In 2009, the New Senior Secondary (NSS) curriculum was implemented with a significant overhaul of the test structure. The NSS curriculum is anchored by four core subjects (including the English language) and one major summative test, the HKDSE, which tests the corresponding subjects at the end of Form 6.

Hong Kong's education system is also characterized by its systematic approach, at least in rhetoric, to measuring students' learning outcomes in both summative and formative manners (see Curriculum Development Council, and Hong Kong Examinations and Assessment Authority, 2007). The Hong Kong Examinations and Assessment Authority (HKEAA) (hereafter referred to as the "testing authority"), a statutory body of the Government of Hong Kong, administers the tests. It is worth noting that the Territory-wide System Assessment (TSA) administered at the end of each key stage (Primary 3, Primary 6, and Secondary 6) is a formative, low-stakes test. This clarification is necessary to highlight the potentially substantial washback effect of the HKDSE, a major high-stakes and summative test. For the approximately 550,000 primary and secondary school students (Hong Kong Special Administrative Region Government, 2019) enrolled annually in the system, this university entrance test is the finale that concludes their years of compulsory education.

Assessment of English reading literacy in Hong Kong

Reading literacy, which the Organisation for Economic Co-operation and Development (2019b) affirms as a fundamental ability for learners to succeed in other subjects—and more importantly, in life—is a major subject on the PISA test. While PISA primarily captures reading literacy in the learners' first language, it does not mean that second language reading literacy is of lower importance. The development of reading literacy often stems from social and cultural needs (Grabe, 2008), and as previously explained, Hongkongers are inevitably subject to such needs. The same may be observed in other East Asian countries and regions, hence their respective standardization of the English language.

Hong Kong is one of the few regions in Asia that publish official documents on assessment frameworks. Still, documents from both the Education Bureau and HKEAA are vague, making informed instruction and test design difficult. This stands in contrast with the testing authority's enthusiasm for "Reading to Learn," a strategy that promotes in-depth reading for cognitive development since 2001 (see Curriculum Development Council, and Hong Kong Examinations and Assessment Authority, 2007). Part of the problem is that the construct *reading literacy* is arguably ill-defined, and the test designers have failed to include a matrix of descriptors in test specifications. It should be noted that the testing authority does not specify the distribution or weight given to each learning objective that is outlined in the

Assessment Guide (see Curriculum Development Council, and Hong Kong Examinations and Assessment Authority, 2007); the document analysis in this research paper, therefore, seeks to uncover this distribution.

Psycholinguistically oriented research on reading

It seems impossible to adequately define reading as a construct. Indeed, even Alderson (2010) struggles to give a clear-cut definition of the ability in his much-referenced book, except that it entails understanding of a text through both cognitive and metacognitive processes. Without any attempt to draw a firm dichotomy, one might acknowledge that it concerns both bottom-up and top-down cognitive processes in meaning making, the former entailing the use of smaller, linguistic units in understanding the bigger picture and the latter the activation of the reader's schemata in interpreting the text (Goodman, 1969, 1982, as cited in Alderson, 2010; Khalifa & Weir, 2009). While there is ample evidence supporting either the bottom-up or the top-down orientation, the interactive approaches to reading as underlined in Grabe (1991) encapsulate the intricate interaction of both orientations as one engages in reading.

Whether it is an indivisible whole or operates on dissectible components, however, has not achieved consensus. Goodman (1967, 1969, as cited in Koda, 2005), for example, contends that reading is naturally embedded in communication and therefore is unlikely to be understood by discrete measures of components. Similarly, Rosenshine (1980, as cited in Khalifa & Weir, 2009) concludes that there is inconsistent evidence concerning the distinctiveness of reading skills. On the other hand, the potential divisibility of the construct is probably more welcomed where testing is concerned, for the components can be measured post hoc with, for instance, correlational and factorial research (Khalifa & Weir, 2009) where reading is considered more a product than a process. But even these quantitative analyses are subject to sampling issues and variation in the method of analysis (Alderson, 2010; Khalifa & Weir, 2009).

It is therefore expected that second language (L2) reading can only be complicated in its operations. Researchers may have steered clear of the likelihood of it being a unitary construct given the dual-language involvement and taken on the componential view (Koda, 2007; Weir, 2005). Still, to measure L2 reading skills entails the assessment of both the reading ability and language proficiency, which are not entirely distinct from each other. Alderson (2010) contends based on research evidence that a test assessing L2 reading skills would require the reader to cross a "linguistic threshold" (p. 121), which varies according to task demands.

Assessment of reading literacy in national and international contexts

Two bodies of literature are helpful in situating the present study's aim: the first on cognitive processes in assessments of reading literacy, and the second on manipulation of test difficulty. This literature review aims to demonstrate that while researchers have long lauded the merits of and advocated for incorporating a comprehensive range of cognitive processes in language assessment, they have thus far made few attempts to

suggest corresponding measures to realign test difficulty, which serves to accommodate and edify varying learner attributes.

Levels of cognitive processes in reading literacy assessment

Bloom (1956) defines cognitive processes as those of recalling knowledge and developing intellectual abilities and skills. While the classification of cognitive processes in Bloom's (1956) taxonomy primarily serves the purpose of formulating educational objectives in curricula rather than assessments, it has since emphasized the inter-relation of the *range* of such objectives in the taxonomy and the importance of exploring educational activities for higher order problem solving. In other words, cognitive processes concern both breadth and depth of knowledge. Since assessment is integral to and should be aligned with curricular objectives, the development of scales and measurements of cognitive processes for analysis of assessment materials has been embraced by a host of disciplines. It is of particular interest to non-language subjects such as mathematics and science, both of which are "highly formal" and "practical" (Tamir, 1980, as cited in Edwards & Dall'Alba, 1981, p. 159) and therefore could be captured in the taxonomy.

Numerous studies have used variations of frameworks mostly based on Bloom's (1956) Taxonomy to analyze the cognitive processes in instructional materials and assessment tasks, though most have focused on mathematics and sciences (Berger et al., 2010; Momsen et al., 2017). There has been a lack of studies exploring the design of assessment tasks at different levels of cognitive processes in the context of language subjects or the humanities in general. This may be attributed to a correspondingly scarce amount of research on the development of scales and measurements for language assessment analysis. Hess (2004) first addresses this paucity of research by introducing a descriptive framework for assessing reading literacy based on Webb's (1999) Depth of Knowledge model. Johnson and Mehta (2011) later build on the Complexity-Resources-Abstractness and Strategy (CRAS) framework developed by Edwards and Dall'Alba (1981) and bring it to the attention of language assessors, taking advantage of the flexibility of the scale in terms of its applicability to an extensive range of subjects. It is therefore feasible to adopt one or more frameworks to analyze language assessment tasks, although prior analyses so far have been empirical and undertaken mostly at the classroom or school level, as opposed to the systemic level.

Unfortunately, at the systemic level, the significance of incorporating a range of cognitive processes in reading assessments has been overlooked in Hong Kong. As the city-state constantly claims one of the top spots in international testing, local authorities have downplayed the fact that students take the PISA in their native language of Chinese (Organisation for Economic Co-operation and Development, 2019a). As such, the PISA results provide little explanatory or predictive relevance to the English reading literacy test in the HKDSE. The transfer of abilities of word decoding from one's first language to a second language is far from automatic; it is especially true when the orthographic and phonological processes required of the first language and the second language are vastly different (Verhoeven, 2011), as is the case with English and Chinese. Therefore, the PISA results do not justify complacency; rather, it is a warning to the local testing authority that a critical missing piece in the HKDSE test must be filled: in contrast to PISA, the HKDSE does not currently guarantee the assessment of a full range of cognitive

processes, even though the curriculum documents present as such (see Curriculum Development Council, and Hong Kong Examinations and Assessment Authority, 2007). Researchers who wish to examine whether there is a skew in the range of cognitive processes in the test need to analyze each question with reference to a cognitive process model.

Manipulation of test difficulty

It is essential that the level of cognitive processes and test difficulty be saliently distinguished in the context of assessment, even though the two terms may be used interchangeably colloquially. Scholars have attempted to clarify the difference, and successfully so, by considering the level of cognitive processes as one of the many factors that might affect test difficulty (Bensoussan et al., 1984; Gan, 2011). Other factors, such as the number of steps required of the learner in each question, the sequencing of the tasks, and the number of parties involved (applicable to group work), can be manipulated to balance the difficulty brought by higher order cognitive processes (K. Hess, personal communication, January 27, 2020). It is therefore plausible to posit that with a well-established hierarchy of test difficulty, students with varying abilities would be able to complete tasks according to their ability without being restricted to the levels of cognitive processes required in the tasks. In other words, there can be difficult tests without higher order cognitive processes, and vice versa.

Researchers have also looked at the interplay among factors contributing to reading literacy (Olmez, 2016; Skehan, 2001). This body of research contributes to the literature of second language test difficulty; Skehan (2001), in particular, proposes five psycholinguistic categories that would affect test difficulty: *interactivity*, *familiar information*, *degree of structure*, *complex and numerous operations*, and *communicative stress*. These domains largely coincide with the *text* factor outlined in the assessment framework under the PISA reading literacy test. The *text* factor is one of the three factors identified by the OECD with respect to their PISA methodology, along with the other two factors, *reader* and *task* (Organisation for Economic Co-operation and Development, 2019b). While the inherent *reader* factor (e.g., motivation or prior knowledge of the learners) is independent from the test, meaning there is less of an intervention feasible on the *reader* factor side, the *text* factor and the *task* factor could be manipulated by assessors and used to interact with the *learner* factor by activating the learners' schema (Snow, 2002). The level of cognitive processes, while deemed highly relevant to test difficulty, essentially represents the *task* factor. Assessors may additionally turn to the *text* factor to manipulate test difficulty by determining four domains in the text, namely: source (whether the texts come from a single source or multiple sources), organization and navigation (whether the texts are dynamically or linearly organized), format (whether the texts are continuously presented in paragraphs or not), and text type (for example, literary or factual) (Snow, 2002). These could all be presented as possible solutions to appeal to policymakers to revise the assessment design of a test.

In conclusion, with test difficulty taken into consideration, a revision of assessment design that employs a full range of cognitive processes can be realized. While scales and measurements of cognitive processes for analysis of assessment have been primarily designed for non-language assessments, models have been further developed to fit

the language assessment context, though without evidence from document analysis. This research study attempts to fill the gap in the literature by shedding light on how the range of cognitive processes varies in the test over the past 8 years relative to the accuracy rate.

Conceptual framework

This study is guided by Bloom's (1956) taxonomy of educational objectives, which establishes the significance of the cognitive domain, and Stanovich's (1980) interactive compensatory model. The former is so prominent that it would be virtually impossible to write a paper on this topic without acknowledging it; the latter model posits that a deficit in one's cognitive process may be compensated for by greater reliance on another cognitive process, thus explaining the surprising empirical evidence of poor readers using higher order contextual processing to compensate for a deficit in word recognition. The aim is to determine whether a full range of cognitive processes is assessed in the HKDSE test, hence the emphasis on the *task* factor, which has been identified in the previous section as one of the key three factors involved in reading assessment methodology (the other two being the *reader* factor and the *text* factor). I also explore if items assessing higher order cognitive processes correspond to extremely low accuracy rates, also defined earlier as the percentages of test-takers who answered the items correctly. An affirmative answer to that analysis would identify a critical need for revision of the assessment design.

Taxonomy of educational objectives

The widely applied Bloom's (1956) taxonomy is a model for classifying and analyzing cognitive educational objectives devised primarily for teachers and curriculum makers. The cognitive objectives are categorized into six subdivisions according to the complexity of learner behavior in the outcomes: *knowledge*, *comprehension*, *application*, *analysis*, *synthesis*, and *evaluation*, with evaluation ranking as the highest in the taxonomy. While it is designed for formulating activities in curricula rather than objectives in assessments, the fact that assessment is integral to curricular objectives means that the taxonomy has significant implications on test design. It should be noted that the hierarchy is dynamic in a way that the highest order cognitive process in the taxonomy (i.e., evaluation) could herald the acquisition of new knowledge, which in turn leads to a new series of cognitive learning. Anderson et al. (2001) later revise the taxonomy with refined definitions of cognitive processes in each subdivision and with an increased emphasis on higher order cognitive processes, which enhance not only retention but also transfer of knowledge. Admittedly, the taxonomy is limited in its descriptive capacity as there have been disputes as to whether all educational objectives could be saliently classified; this potential limitation is mitigated by inter-rater reliability when coding the assessment items. Additionally, the use of hierarchy might induce the illusion that the attainment of skills and abilities in one level would become a prerequisite for the attainment of skills and abilities in the next level. This pitfall, nonetheless, could be addressed by Stanovich's (1980) interactive compensatory model.

Interactive compensatory model

The interactive compensatory model debunks the myth that higher order cognitive processes are only implicated in the reading process of good readers. Empirical evidence suggests poor readers might use more contextual reasoning, which is considered a higher order cognitive process, to aid their reading when compared to good readers. Instead of calling the evidence paradoxical, Stanovich (1980) explains it with the assumption of compensatory processing in the model. He posits that higher order cognitive processes are not exclusive to good readers, nor are they a prerequisite for entering a level in the cognitive process hierarchy. It is possible that poor readers "compensate" for their poor word recognition, which is at the lower end of cognitive processes, with a heavier reliance on higher order knowledge sources such as the context of the text to help with understanding. This contrasts poor readers with good readers who do not need to rely on contextual processing to aid in reading; it is posited that good readers could manage word recognition without a context provided. Therefore, higher order cognitive processes are not intended only for good readers, nor should they be. It should be noted, however, that while the theory focuses on reading fluency, it does not specify the context of second language learning. Furthermore, although it relies heavily on the interplay between the *reader* factor and the *text* factor, it mentions little about the cognitive processes implicated in the *task* factor.

My theoretical perspective combines Bloom's (1956) taxonomy of educational objectives and Stanovich's (1980) interactive compensatory model to analyze the range of cognitive processes assessed in the reading literacy test. This paper argues that a revision of the assessment would be warranted in the case of a consistent skew towards items assessing lower order cognitive processes. Bloom's (1956) taxonomy of educational objectives emphasizes the importance of exploring educational activities for higher order problem solving, which in turn influences assessment activities. Based on this theory, this study explores whether the testing authority places outsized emphasis on lower order cognitive processes. Stanovich's (1980) interactive compensatory model asserts that reading performance is a "synthesized" (p. 35) process in which readers draw on multiple knowledge sources, meaning even poor readers could tap high-level knowledge sources such as semantic knowledge and contextual knowledge. It hypothesizes that items assessing higher order cognitive processes could be incorporated in the test while maintaining reasonable test difficulty. Through a combined use of Bloom's (1956) taxonomy of educational objectives and Stanovich's (1980) interactive compensatory model, I explore a potential need to broaden the range of cognitive processes in Hong Kong's English language reading literacy assessment and present the following research questions:

1. How has the range of cognitive processes in the HKDSE English reading literacy test varied over the first 8 years of its administration? Is there a pattern to the cognitive processes assessed?

   Hypothesis: There has been a consistent emphasis on the assessment of lower order cognitive processes.

2. If there is a pattern, how does it correspond with the accuracy rate of the items?

   Hypothesis: *The higher order items are more error-prone.*

Methods and data

In conducting an analysis of test items in the HKDSE English reading literacy test administered between 2012 and 2019, I focused on the *task* factor to better understand how levels of cognitive complexity in the items might correspond to accuracy rates. Guided by this assumption, I examined and compared items in each test administration in terms of the level of cognitive processes assessed, as well as the corresponding percentage of test-takers who gave a correct answer. In order to determine the cognitive complexity of each item, a document analysis of raw data from official test documents was conducted.

An inevitable limitation to this research study is that the documents may lack data rich enough to adequately answer the research questions (Bowen, 2009; Gross, 2018), which could have been complemented by triangulating multiple types of data. This study has opted for the broadest representation in the sample of documents possible—using the test from all available years since the implementation of the new curriculum in 2012—to help mitigate the triangulation challenges.

I sampled the entire population of interest due to its small population (Maul, 2018) which amounts to eight sets of readily available tests. I purchased the tests in bulk in 2019 through the online bookstore of the testing authority. Each of the texts was published months after the test was administered. Each test comprises three sections: Part A (the compulsory section), Part B1 (easier section), and Part B2 (more difficult section).[1] Each part consists of about 20 questions and carries approximately 40 marks. Every year the test items are developed anew rather than anchored to an accumulated pool of items, which sets it apart from the practice of item sampling in international tests such as the OECD's PISA. In other words, even though the question types in the test may be comparable across the years, the passages and the questions are never recycled. It is therefore important to exercise caution when interpreting the accuracy rates across the years, given the cohort effects are uncontrolled (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), meaning there is no normalization of scores across cohorts.

Data collection and analysis are interwoven throughout in qualitative inquiry (Suter, 2012). Therefore, I analyzed the data using a combination of deductive and inductive coding (Creswell & Creswell, 2018); 13 out of 23 (57%) are data-driven codes while the rest (43%) are codes from an adapted matrix. The codebook was tested using a subset of data to improve appropriateness (Gross, 2018) and revised after checking with a second coder to enhance inter-rater reliability (Creswell & Creswell, 2018). I compared my codes with those from the second coder and found a 78.57% agreement rate, which is "substantial" according to Landis and Koch (1977). To further ensure reliability, I recoded a random sample of 10 items from each test year (80 items in total, which constitute 8% of the total items). Kappa statistics indicated an almost perfect level of agreement ($\kappa = 0.94$). In addition, even though the coding was performed by a single coder, potential imprecision of coding can be accommodated if the results are explicit (Tengberg, 2017). I also recognized my subjectivity when considering my position as the researcher, a former student who was educated in the Hong Kong education system, and a former teacher who taught

[1] HKEAA defines, in its words, Part B1 as the "easier section" and Part B2 the "more difficult section." Test-takers may attempt either Part B1 or Part B2, though the ceiling for grading is capped at level 4 for the former.

students to navigate the same system. Understanding that the analysis could therefore be susceptible to bias, I conducted inter-coder reliability tests with second coders and simultaneously clarified potential ambiguity in the coding scheme.

The adapted matrix takes into account two independent matrices. The first matrix is devised by the International Association for the Evaluation of Educational Achievement (IEA) for the Progress in International Reading Literacy Study (PIRLS), which comprises four comprehension processes as the framework for the development of the items (Mullis et al., 2013). The matrix might seem an unorthodox choice as a reference for this study, as PIRLS aims to assess fourth grade students' reading literacy as opposed to twelfth graders in the HKDSE. However, I reviewed the brief descriptors available under each process and found them adequately fitting for use as predetermined codes for data analysis. In a comparative qualitative study, Tengberg (2017) explores variations in national reading tests in three Nordic countries and corroborates the use of the matrix for the tests aimed at 13- to 16-year-olds. It is also worth noting that the matrix for PISA was not considered for this study because PISA does not anchor its test items to any curricula. PISA focuses on abilities applicable in adult life, such as digital literacy (Organisation for Economic Co-operation and Development, 2019b), which are not aligned with the focus of this study.

The second matrix, which originates from Norman Webb's Depth of Knowledge (DOK) framework for assessment in mathematics and science, has been previously adapted by Hess (2004) for state-wide, school-wide, and classroom-wide assessment in reading at the primary and secondary levels. Unlike those in Bloom's (1956) taxonomy, the descriptors from Hess (2004) focus on assessment rather than learning activities. The matrix resembles a rubric for test designers, indicating both the assessment tasks and the cognitive levels assessed. For standardized tests[2] at the senior secondary level, basic reasoning (level 2) as a lower order thinking skill and complex reasoning (level 3) as a higher order thinking skill should be valued equally, while recall of information (level 1) should be kept minimal (K. Hess, personal communication, January 27, 2020).

I primarily used Excel to apply the codes to the tests and to store the coded data. For the second part of the data analysis process, STATA was employed to run statistical *t*-tests to ascertain the differences in accuracy rates between the cognitive levels.

Results

The lack of detailed description for the construct *reading literacy* in the test specifications poses a challenge to those seeking to evaluate the instrument (see American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), specifically the validity of the interpretation of the HKDSE test scores for measuring attainment. Still, by coding and mapping each test item with reference to the cognitive level assessed, it is possible to delineate the focus of the test and explicate the relationship between cognitive demand and student performance.

[2] Level 4 is not applicable to such standardized tests as the HKDSE. Level 4 items assess extended reasoning and usually require the use of multiple texts over an extended period of time.

Consistent emphasis on assessment of lower order cognitive processes over time

It was important to establish the level and type of cognitive processes in order to understand the content of the test and the construct it assesses. Figure 1 shows the possible scores by section in the test by year. It shows that despite the peak in 2012 when the test was first administered and the trough in 2015, the total possible score has been consistently positioned in the range of 122 to 129 points.

This consistency in format and test length suggested by Fig. 1 corresponds to the consistency in the cognitive demand of the test, which is illustrated in Fig. 2. The results reveal that the testing authority has sustained an overwhelming emphasis on lower order cognitive processes across the 8 years. On average, 4% of all the test items ($n = 1015$) are coded as level 1, 87% as level 2, and 9% as level 3. In other words, items coded at level 2 on the matrix comprise the vast majority of the test. Slightly varying between 80 and 93% over the years, the high percentage of level 2 items crowd out the higher order cognitive processes that could be tested. The percentage of items coded at level 3 hovers as low as 4.8% in 2018, which is coupled with a minor jump in level 1 items for the same year, and as high as 17.5% in 2015, which surprisingly sees a number of items assessing higher order cognitive processes in the easier Part B1.

The results reported so far are consistent with the hypothesis associated with the first research question, which seeks to identify the range of cognitive processes in the HKDSE English reading literacy test over the 8 years. There is indeed a pattern of cognitive processes prescribed in the test that is heavily skewed toward basic reasoning. As noted previously, basic reasoning should be adequately assessed but not dominant in the test. This rationale has not been well established in the test specifications though, the ambiguity of which renders it difficult to determine whether the government has intended to produce such a consistent skew. This also seems to be departing

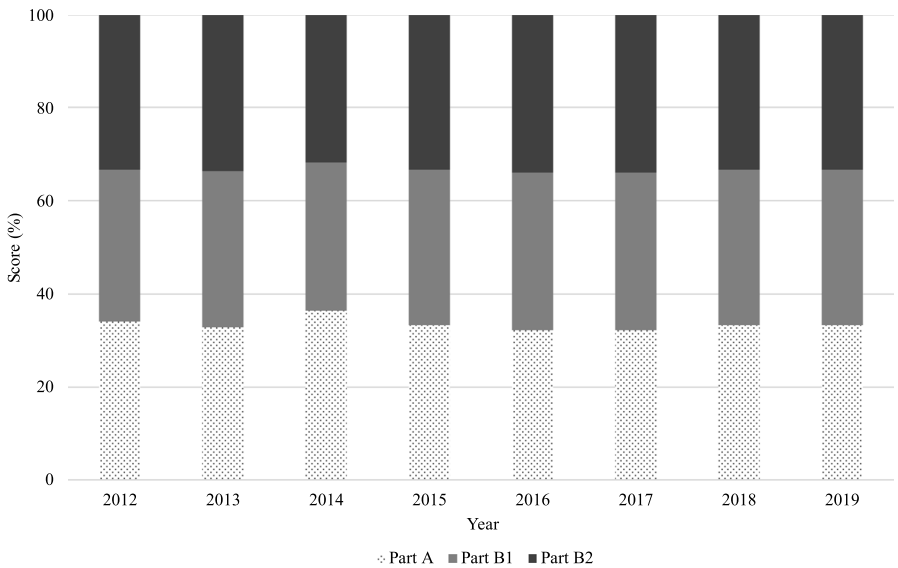

Fig. 1 Maximum possible scores by section by year. Part A = compulsory section. Part B1 = easier section. Part B2 = more difficult section

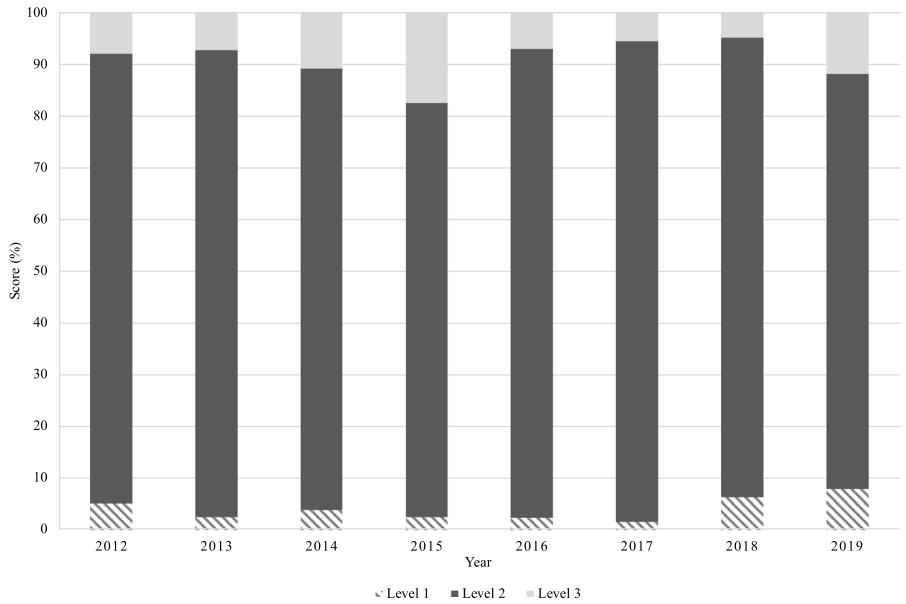

Fig. 2 Score distribution by cognitive level by year. *Level 4 items are not applicable to standardized tests such as the HKDSE

from the guidance in the assessment framework, which recommends the assessment of an array of cognitive processes, ranging from understanding language features to evaluating views in texts (see Curriculum Development Council, and Hong Kong Examinations and Assessment Authority, 2007). Based on the analyses conducted in this study, the test does not appear to have met these expectations.

Figure 3 shows the distribution of reading literacy processes in the test administrations across 8 years, which were coded according to the adapted matrix. A total of

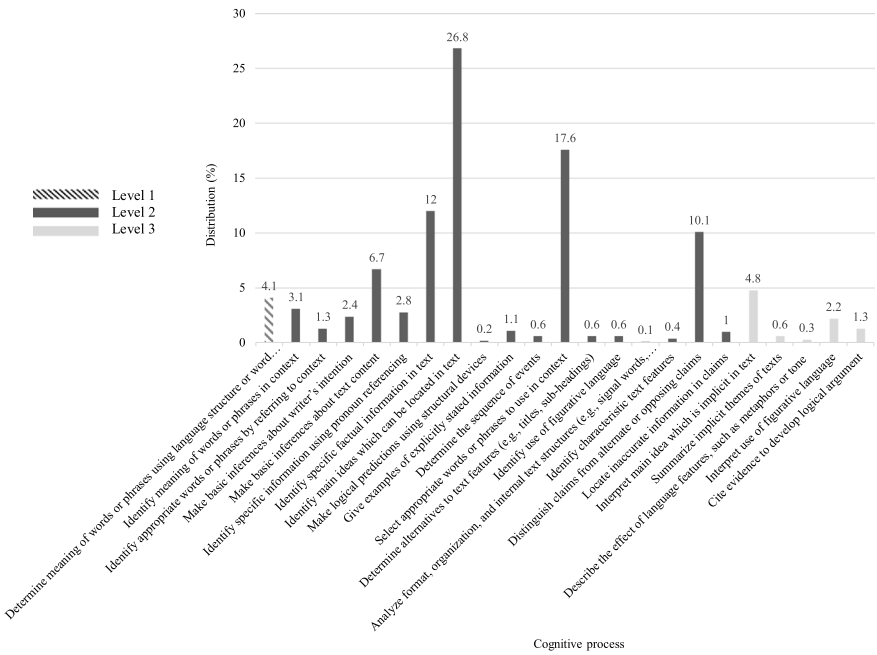

Fig. 3 Distribution of reading literacy processes across all years. *The percentage numbers add up to 100%

twenty-three reading literacy processes emerged as relevant codes in the test, including one from level 1, seventeen from level 2, and five from level 3. Three processes requiring basic reasoning alone take up over half of the items every year, namely *identify main ideas which can be located in the text* (26.8%), *select appropriate words or phrases to use in context* (17.6%), and *identify specific factual information in the text* (12%). These three processes have always dominated the test, even though the weight of each process varies across the years. This shows that the test does not only rely on lower order cognitive processes, but it is also composed of a limited range of such processes. The least represented ones, namely *analyze format, organization, and internal text structures (e.g., signal words, transitions, semantic cues)* (0.1%), *make logical predictions using structural devices* (0.2%), and *identify characteristic text features* (0.4%), contribute to less than 1% of all the items across the years.

Underperformance on higher order cognitive items

As previously noted, cognitive demands and item difficulty are not necessarily indicative of one another. As such, it has not been established whether test-takers perform statistically significantly worse in items assessing higher order cognitive processes. I used analysis of variance (ANOVA) to ascertain the differences in accuracy rates between the three cognitive levels of the test items. With cognitive levels as the independent variable and student performance the dependent variable, the ANOVA determines whether the hypothesis that items assessing higher order cognitive processes are met with poorer student performance holds. I computed the average accuracy rates for each of the three levels and obtained the values of 53.59 for level 1, 53.84 for level 2, and 30.44 for level 3. The paired significance is indicated in Table 1.

I checked the assumptions of ANOVA to establish statistical conclusion validity, even though it is a robust analytic tool. Student performance measured by the percentage of test-takers who gave a correct answer is a continuous dependent variable. Independence is ensured as each item is coded at only one cognitive level. With a large sample ($n = 1015$), student performance is expected to vary normally within each cognitive level. Also, to check for homogeneity of variance, I looked at the standard deviations of the samples (Bartlett's prob > chi square = 0.342).

The significant F-value of 56.43 indicates that the means are not all equal, rejecting the null hypothesis at the 5% significance level. The Bonferroni adjustment for post hoc t-test correction shows that the difference between the means of levels 1 and 2 is .25. This difference is not statistically significant at the .05 level. On the other hand, the difference in means between levels 2 and 3 is 23.40, which is statistically significantly different at the 5% significance level. Similarly, there is a statistically significant difference

Table 1 ANOVA results for average accuracy rates across the years

Cognitive level	*N*	*M/SD*	*F*	*p*	Pairwise comparisons (*p*)
Level 1 (L1)	41	53.59/17.03	56.43	.00	L1&L2 (1.00); L1&L3 (.00*); L2&L3 (.00*)
Level 2 (L2)	882	53.84/20.29			
Level 3 (L3)	92	30.45/19.86			

*The Bonferroni-adjusted difference is significant at the level of <0.1%

of 23.14 between the means of levels 1 and 3 at the 5% significance level. This provides evidence that, on average, student performance is indeed related to the cognitive level of the items. The negative values (−23.40 and −23.14) indicate that the more cognitively demanding level 3 items are indeed more error-prone. The hypothesis in the second research question is therefore confirmed.

Discussion

As previously discussed, Bloom's (1956) taxonomy is not meant to be considered as a linear hierarchy. An equally dynamic model is Stanovich's (1980) interactive compensatory model, which suggests that higher order cognitive processes are implicated in the reading process of not only good readers but also poor readers. This provides good grounds for test developers to measure higher order cognitive skills in such large-scale assessments as the HKDSE, debunking the myth that only a select group of good readers could manage items at higher levels of cognitive complexity. Recommendations from the *Assessment Guide* (see Curriculum Development Council, and Hong Kong Examinations and Assessment Authority, 2007) coincide with the model, as it endorses assessment of an array of cognitive processes ranging from understanding language features to evaluating views in texts for all learners. Given such guidance in the assessment framework, the cognitive processes prescribed in the test are expected to be represented in a balanced manner. However, based on the analysis, the test departs from these expectations.

The results presented previously in Fig. 2 confirm the hypothesis that there has been a consistent emphasis on assessing lower order cognitive processes. An explanation for this preference is that items aligned to these processes are inherently easier to develop, as they rely primarily on text-based information and therefore require minimal effort to develop the scoring rubrics. Items assessing higher order cognitive processes, on the other hand, involve contemplating not only the text but also the manipulation of language in the text. Item developers must make strenuous efforts to estimate the ways students make judgments in their responses, followed by construction of plausible distractors for selected-response items or specific scoring rubrics for constructed-response items. Only well-developed items could stand on statistical grounds and serve to make distinctions among test-takers of varying abilities.

In line with prior studies that indicate a lack of attention to higher order cognitive processes in Hong Kong's English language classroom, the substantial skew towards lower order cognitive processes revealed in the results carries troubling implications for the washback effect of the test. As previously stated, students under the influence of the Confucian heritage culture tend to be driven by achievement motivation in learning and often a misplaced emphasis on test examinations. As a result, the unwavering focus on lower order cognitive processes in the test is more likely to find its way to the classroom. This deprives the students of the opportunity to familiarize themselves with complex reasoning skills, which they will likely need to demonstrate later in their lives.

Another issue that arises is the so-called construct underrepresentation. According to the *Standards for Educational and Psychological Testing* (see American Educational

Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), a prescriptive guide for professional test developers, construct underrepresentation occurs when a test fails to capture comprehensively the constructs or concepts it was meant to measure in the first place. Without a sufficient variety of cognitive levels in the items, the HKDSE English reading literacy test risks underrepresenting *reading literacy* as a construct. This consequently compromises the validity of extant interpretations of the test scores. As noted earlier, the HKDSE stands as the single high-stakes test that determines admission to local, and in some cases, overseas universities. The "signaling" power of the test is therefore unmatched by any other tests in the city-state. Yet, these validity issues may compromise the public trust in the test as a primary source of evidence that substantiates students' cognitive ability and readiness for university, if it permits little useful inference for institutions and employers.

Evidence from this study suggests that there is also an underrepresentation of certain lower order cognitive processes in the test (see Fig. 3). In other words, the test not only relies on the assessment of basic reasoning, but is also composed of a limited range of lower cognitive processes (e.g., *identify main ideas which can be located in the text*, *select appropriate words or phrases to use in context*, and *identify specific factual information in the text*). There are two implications of this. First, it underlines a potentially intensified washback effect: not only can teachers focus on only teaching lower order cognitive processes, they can only teach the most popular lower order cognitive processes, which cumulatively account for 56.4% of all the items delivered in the test. Second, this finding challenges the assumption that the items would sufficiently tap the processes one would expect from a professional test. As a result of the lack of alignment between the content and cognitive demands of the items and those described in the *Assessment Guide*, the meaning of the test scores becomes even narrower. This further weakens the validity evidence in support of using the test scores to infer the English reading proficiency of individual test-takers or the overall level of high school students in Hong Kong.

To further illustrate the argument on limited validity, I reviewed items that assess retrieval of information (level 1). The results report only a small percentage of such items, as expected in standardized tests (K. Hess, personal communication, January 27, 2020). Still, the singular focus on definitions and synonyms of vocabulary in these level 1 items again raises concerns about the underrepresentation of other cognitive processes within the cognitive level. The fact that these items can be answered without referring to the text—a feature that differentiates level 1 items from level 2 items—trivializes the importance of understanding the contextual meaning of words and phrases. Unless vocabulary recall is defined as part of the test construct (i.e., *reading literacy*), which is not the consensus according to the *Assessment Guide*, these items illustrate the earlier point around construct-irrelevance and continue to complicate accurate interpretations of students' performance on reading tasks.

As the results have confirmed the expected pattern of the cognitive processes assessed in the test, I now discuss the second research question. The statistically significant difference in the means of test-takers' test scores between level 1 and level 3 items and

between level 2 and level 3 items confirms the hypothesis that the higher order items are more error-prone. Yet, such a difference may be mitigated if test difficulty is adjusted. As described in the literature on the manipulation of test difficulty, cognitive demands do not necessarily dictate test difficulty or accuracy rates. While items measuring higher order cognitive processes could be extremely difficult and consequently fail to contribute to the statistical reliability of the total possible score, test designers should strive to keep the items anyway. One way to do that would be to reduce the reading load or provide scaffolding to aid understanding. Failure to do so, as in the case of the HKDSE English reading literacy test, may compromise test content coverage and skew the test toward lower order cognitive processes.

The results of the current study provide empirical evidence that a comprehensive assessment framework must be established in response to the need to re-design the test. The most pressing need is to fill the gaps in the test specifications in which the constructs and psychometric specifications are not clearly defined. *Reading literacy*, in particular, is a construct that leaves ample room for interpretation given the interplay of the *text, task, and reader* dimensions, which were discussed at length previously. For example, when analyzing national reading tests in three Scandinavian countries, Tengberg (2017) finds that the tests are focused on different dimensions of the construct. The Danish test emphasizes the use of reading techniques such as skimming and identifying important information in a text, while the Norwegian and Swedish tests primarily require students to analyze, interpret, and reflect. The variation is expected and acceptable as long as the construct definitions are clearly defined by the assigned testing authority (which is often a part of the government) to allow validity evaluation. The HKEAA should consider operationalizing the construct of *reading literacy*, perhaps by looking comparatively at other national and international reading tests such as the National Assessment of Educational Progress (NAEP) in Reading in the USA and IEA's PIRLS as mentioned in the *Data Analysis*. The broad learning outcomes currently outlined in the *Assessment Guide* should be revised accordingly to include more precise parameters of the cognitive processes assessed, and decisions must be made about their relative weights on the test score. Specifically, the HKEAA should clarify the distribution of items across the spectrum of cognitive levels and include rationales about any adjustments for test difficulty.

Note that the cognitive level and reading literacy process assessed in individual items are determined by the reading passage as much as by the prompt (Rowe et al., 2006). It is beyond the scope of this paper to present a full discussion of the reading material around which the test questions are structured. However, future research would benefit from further analyzing the *text* factor, including the text genre (e.g., digital-based communication or scholarly essays), the text format (e.g., continuous, non-continuous, or mixed[3]), and the text length of each administered test, as well as examining the interaction between text characteristics and student performance. In addition, the nature of this study as a document analysis limits the validation of cognitive processes from the perspective of the readers. Given that cultural and social exposure may alter the cognitive demands of the items and hence lead to differences in reading achievement (Snow,

[3] Continuous texts are typically organized into paragraphs and found in articles and books; non-continuous texts are typically composed of lists, tables, and diagrams; mixed texts are a combination of both continuous and non-continuous texts.

2002), an empirical investigation of the *reader* factor would be necessary to generalize the results to a particular population of students. Evidence from think-aloud protocols, for example, helps to ascertain the level of cognitive processes used by the students as they take the test (Tengberg, 2017). Alternately, a continuation of this study could be enhanced by triangulating converging data sources (Creswell & Creswell, 2018), such as interviews with in-service teachers and representatives from the testing authority, which would help introduce multiple perspectives to the study. Also, since the HKEAA only publicly releases information on aggregate accuracy rates for each test item, this study is limited by the lack of individual response data—ideally, produced by a sample of examinees—that would otherwise enable the use of more sophisticated analytic approaches, such as cognitive diagnostic modelling.[4]

As suggested by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014), the test developer (in this case, the HKEAA) and the test user (in this case, the tertiary institutions and employers) are responsible for validating the interpretation of test scores. However, education policymakers can also make a positive contribution to this process in two ways: First, they can consider the implications of increased representation of items measuring higher order cognitive processes in such large-scale assessments as the HKDSE. Second, they can push for thorough re-design and/or continuous evaluation of the test. The next action of importance could be to improve students' skills in managing tasks that require higher order cognitive processes. This is a lofty goal, since actualizing it would imply making changes in the curriculum to match the test demands and additional training for teachers to help them overcome tendencies to teach to the test. Still, given the intensity of the negative washback effect in Hong Kong, we can be hopeful that the promise of positive washback could overcome the obstacles in policy fine-tuning and pedagogy.

Conclusion

In conclusion, the literature on levels of cognitive processes and manipulation of test difficulty shows that higher order cognitive processes and higher test difficulty do not necessarily go hand in hand. However, there exists a significant gap in the literature around the hypothesized skew towards lower order cognitive processes in standardized tests such as the HKDSE English reading literacy test. The research study fills this gap in the literature by conducting a document analysis of the test questions since the introduction of the new test in 2012, highlighting the implications of the underrepresentation of higher order cognitive processes on the washback effect. The results lend weight to the notion that, when executed well, assessment could go beyond an instrument with a bad reputation to become a motivational source to teach young people the cognitive skills required of them in workplace and in life. In a policy paper commissioned by the National Center for Education Statistics (2012), the researchers state that application or generalization essentially stems from what is taught. If there is a washback effect, let it be positive, and let it in turn guide what is taught and what is learned in classrooms.

[4] I thank the anonymous reviewers for this suggestion.

Abbreviations

AERA	American Educational Research Association
ANOVA	Analysis of variance
APA	American Psychological Association
CDC	Curriculum Development Council
CRAS	Complexity-Resources-Abstractness and Strategy
HKDSE	Hong Kong Diploma of Secondary Education Examination
HKEAA	Hong Kong Examinations and Assessment Authority
IEA	International Association for the Evaluation of Educational Achievement
NAEP	National Assessment of Educational Progress
NCME	National Council on Measurement in Education
NSS	New Senior Secondary
OECD	Organisation for Economic Co-operation and Development
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
TSA	Territory-wide System Assessment

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40468-022-00167-4.

Additional file 1: Appendix A. The adapted matrix, which serves as the coding scheme. **Appendix B**. The codes and their corresponding examples. **Appendix C**. Distribution of reading literacy processes in all the test administrations by year.

Acknowledgements
I am grateful to Gabriela Gavrila, Dr Christine Min Wotipka, Dr Martin Carnoy, and Dr Edward Haertel for their invaluable feedback on the manuscript.

Author's contributions
I am the sole contributor of all the content in this manuscript. The author(s) read and approved the final manuscript.

Funding
I am grateful for the financial support from **the Bei Shan Tang Foundation as well as** the ICE MA Fund and the Master's Student Travel Fellowship from the Stanford Graduate School of Education in the purchase of the assessment materials.

Availability of data and materials
See Additional file 1 attached.

Declarations

Competing interests
The author declares no competing interests.

Received: 22 November 2021 Accepted: 14 May 2022
Published online: 01 July 2022

References

Alderson, J. C. (2010). *Assessing reading*. Cambridge University Press.
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://doi.org/10.4135/9781506326139.n662.
Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., … Wittrock, M. C. (2001). *Taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
Bacon-Shone, J., Bolton, K., & Luke, K. K. (2015). *Language use, proficiency and attitudes in Hong Kong*. The University of Hong Kong.
Bensoussan, M., Goldenblatt, L., & Kreindler, I. (1984). Changing the difficulty level of multiple-choice EFL reading comprehension questions. *Language Testing*, *1*(1), 105–109. https://doi.org/10.1177/026553228400100110.
Berger, M., Bowie, L., & Nyaumwe, L. (2010). Taxonomy matters: Cognitive levels and types of mathematical activities in mathematics examinations. *Pythagoras*, *71*, 30–40. https://doi.org/10.4102/pythagoras.v0i71.4.
Biggs, J. B. (1991). Approaches to learning in secondary and tertiary students in Hong Kong: Some comparative studies. *Educational Research Journal*, *6*, 27–39.
Biggs, J. B. (1996). Western misconceptions of the Confucian-heritage learning culture. In D. A. Watkins, & J. B. Biggs (Eds.), *The Chinese learner: Cultural, psychological and contextual influences*, (pp. 45–67). CERC; ACER.
Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Longmans.
Bowen, G. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27–40. https://doi.org/10.3316/QRJ0902027.

Cheng, L. (1997). How does washback influence teaching implications for Hong Kong? *Language and Education*, *11*(1), 38–54. https://doi.org/10.1080/09500789708666717.
Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*, (5th ed.,). SAGE.
Curriculum Development Council, & Hong Kong Examinations and Assessment Authority. (2007). English language curriculum and assessment guide (Secondary 4-6). https://334.edb.hkedcity.net/new/doc/chi/curriculum2015/EngLang_CAGuide_2015.pdf
Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., … Steele, C. M. (2013). *Criteria for high-quality assessment*. Stanford Center for Opportunity Policy in Education https://edpolicy.stanford.edu/sites/default/files/publications/criteria-higher-quality-assessment_2.pdf.
Edwards, J., & Dall'Alba, G. (1981). Development of a scale of cognitive demand for analysisof printed secondary science materials. *Research in Science Education*, *11*, 158–170. https://doi.org/10.1007/BF02356779.
Gan, Z. (2011). Second language task difficulty: Reflections on the current psycholinguistic models. *Theory and Practice in Language Studies*, *1*(8), 921–927. https://doi.org/10.4304/tpls.1.8.921-927.
Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, *25*(3), 375–406.
Grabe, W. (2008). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.
Gross, J. (2018). Document analysis. In B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation*, (pp. 545–548). SAGE.
Hess, K. (2004). *Applying Webb's depth-of-knowledge levels in reading and writing*. Center for Assessment.
Ho, E. S. (2009). Characteristics of East Asian learners: What we learned from PISA. *Educational Research Journal*, *24*(2), 327–348.
Hong Kong Examinations and Assessment Authority. (2019). 2019 HKDSE Analysis of results of candidates in each subject. http://www.hkeaa.edu.hk/DocLibrary/HKDSE/Exam_Report/Examination_Statistics/dseexamstat19_5.pdf
Hong Kong Special Administrative Region Government. (2019). Hong Kong fact sheets: Education. https://www.gov.hk/en/about/abouthk/factsheets/docs/education.pdf
Johnson, M., & Mehta, S. (2011). Evaluating the CRAS framework: Development and recommendations. *Research Matters*, *12*, 27–33 https://www.cambridgeassessment.org.uk/Images/109988-research-matters-12-june-2011.pdf.
Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.
Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press.
Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning*, *57*(Suppl 1), 1–44. https://doi.org/10.1111/j.1467-9922.2007.00411.x.
Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310.
Marsh, C., & Lee, J. C. (2014). Asia's high performing education systems: The case of Hong Kong. In C. Marsh, & J. C. Lee (Eds.), *Asia's high performing education systems: The case of Hong Kong*, (pp. 1–13). Routledge.
Maul, A. (2018). Judgment sampling. In B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation*, (pp. 914–916). SAGE.
Momsen, J., Offerdahl, E., Kryjevskaia, M., Montplaisir, L., Anderson, E., & Grosz, N. (2017). Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE: Life Sciences*. *Education*, *12*, 239–249. https://doi.org/10.1187/cbe.12-08-0130.
Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2013). PIRLS 2016 Reading Framework. In I. V. S. Mullis, & M. O. Martin (Eds.), *PIRLS 2016 assessment framework*, (pp. 11–29). TIMSS & PIRLS International Study Center; IEA.
National Center for Education Statistics. (2012). NAEP: Looking ahead—Leading assessment into the future. https://nces.ed.gov/nationsreportcard/pdf/Future_of_NAEP_Panel_White_Paper.pdf
Olmez, F. (2016). Exploring the interaction of L2 reading comprehension with text- and learner-related factors. *Procedia Social and Behavioral Sciences*, *232*, 719–727. https://doi.org/10.1016/j.sbspro.2016.10.098.
Organisation for Economic Co-operation and Development (2019a). *PISA 2018 results (Volume I): What students know and can do*. https://doi.org/10.1787/5f07c754-en.
Organisation for Economic Co-operation and Development (2019b). *PISA 2018 assessment and analytical framework*. https://doi.org/10.1787/b25efab8-en.
Prodromou, L. (1995). The backwash effect: From testing to teaching. *ELT Journal*, *49*(1), 13–25. https://doi.org/10.1093/elt/49.1.13.
Rowe, M., Ozuru, Y., & McNamara, D. S. (2006). An analysis of a standardized reading ability test: What do questions actually measure? In S. A. Barab, K. E. Hay, & D. T. Hickey (Eds.), *Proceedings of the seventh international conference of the learning sciences*, (pp. 627–633). Lawrence Erlbaum.
Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing*, (pp. 167–185). Longman.
Snow, C. E. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. RAND.
Stanovich, K. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, *16*(1), 32–71. https://doi.org/10.2307/747348.
Suter, W. N. (2012). Qualitative data, analysis, and design. In W. N. Suter (Ed.), *Introduction to educational research: A critical thinking approach*, (pp. 342–386). SAGE.
Tengberg, M. (2017). National reading tests in Denmark, Norway, and Sweden: A comparison of construct definitions, cognitive targets, and response formats. *Language Testing*, *34*(*1*), 83–100. https://doi.org/10.1177/0265532215609392.
Verhoeven, L. (2011). Second language reading acquisition. In M. Kamil, P. D. Pearson, E. Moje, & P. Afflerbach (Eds.), *Handbook of reading research*, (vol. IV, pp. 661–683). Routledge.
Webb, N. (1999). *Research monograph no. 18: Alignment of science and mathematics standards and assessments in four states*. Council of Chief State School Officers.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan UK.

Publisher's Note