Decision making in large-scale language testing: Intersections of policy, practice and research

Yan Jin

Shanghai Jiao Tong University

In this article, I delve into the intricacies of strategic decision making when a large-scale language test is reformed. Adopting a case study approach, I examine the activities involved in the reform of the College English Test (CET), an English language test for university students administered nationwide in the Chinese mainland (Zhang, 2022; Zheng & Cheng, 2008). In its 30+-year history, the CET has undergone several revisions and reforms (see Jin, 2010; 2020 for an overview). To illustrate how the major policies on test revision and reform have been formulated and implemented in this particular context, I present a brief overview of some major decisions that have been made over the decades. Through these cases, I hope to demonstrate the intricate interactions among policy intentions, professional requirements, ethical considerations, practical constraints, as well as the role of language testing professionals when strategic decisions are made to ensure a sustainable and healthy development of large-scale and high-stakes language tests.

Key words: decision making, large-scale language testing, policy intention, practical constraint, professional responsibility

Introduction

The CET is an English language test on a vast scale. Even during the covid pandemic when test delivery was seriously affected, the number of CET test takers remained at 20 million

Email address for correspondence: yjin@sjtu.edu.cn

a year. The test also has very high stakes: it is used for making various decisions, including university graduation, admission to postgraduate programs, employment, or residential permits for top-tier cities (Jin, 2014; Yang, 2003). With such high stakes, the test developer has to take on the immense responsibility of ensuring its sustainable development (Fan & Frost, 2022; Jin, 2010, 2020). A comment made by Professor Huizhong Yang, former chair of the National College English Testing Committee (NCETC), two decades ago serves as a vivid reminder of the responsibility that has to be shouldered by those involved in large-scale testing:

> *If we used one second to count one test taker, 5 million [the annual test population of the CET in the early 2000s] seconds would be 58 days. Considering that each count could affect a student's life, what a huge responsibility on the professionals of a large-scale language test* (Yang, 2005, personal communication).

In one of his most frequently cited articles, Bachman (2000) also called attention to the responsibility of language testers. In his view, "validity and fairness are issues at the heart of how we define ourselves as professionals" (p. 25). To be responsible professionals, we should "link our most fundamental research questions to the ethical issues of how we practice our profession" (ibid.). One of the fundamental issues to be examined is how policies of test revisions and reforms are being made and executed. In this article, I first review the literature of decision making on language test creation and score interpretation, pointing out the lack of attention to strategic decision making from a broader social perspective. I then introduce how "strategic decision" is conceptualized, drawing on the management literature. This is followed by an analysis of some cases in which strategic decisions have been made about the reform of the CET. The case studies highlight the context-specific nature of decision making at a strategic level and emphasize the social responsibility of the test developer. A critical perspective of decision making is also recommended to empower the stakeholders most susceptible to the policies formed in a top-down approach. To conclude, I stress the need to take into account a broad spectrum of viewpoints so as to gain a more comprehensive and nuanced understanding of decision making.

Literature review

Decision making in language testing

A primary goal of language testing is to measure and report test takers' language proficiency based on evidence gathered from their test performances. To accomplish the aim, language testing professionals are required to make decisions at each phase of the test development cycle, including test design, administration, scoring, and score reporting. Fulcher and Davidson (2009) used an architectural metaphor to depict the multi-layered structure of decision making for language test development: a theoretical overview of language knowledge and language use (the level of model), test purposes and constructs (the level of framework), and test design and delivery (the level of specifications). Insightful as it is, the structure is mostly utilized to illustrate the creation and revision of a language test.

Language testing, as is now widely accepted, is not simply about making a test and giving a score, especially in high-stakes contexts. Decision making for test creation and score interpretation is only part of the mission of the profession. Over the past two decades, the field of language testing has been awakened to the social dimension of its professional services (McNamara & Roever, 2006; Shohamy, 2001a, 2001b; Yang & Gui, 2007, 2015). Consequences of test use have been incorporated into validity frameworks such as the argument-based approach (e.g., Kane, 1992, 2013; Chapelle, 2012), the socio-cognitive framework (Weir, 2005), Assessment Use Argument (Bachman & Palmer, 2010), and the theory of action framework (Chalhoub-Deville, 2016). Chapelle (2020) presented six empirical studies of test use for social functions such as accountability, signaling problems in educational programs, and social mobility. There are also accounts of language testing reforms where complex decisions are involved. Elder (2016) reports a number of instances in which decisions were made about language assessment programs at the macro-societal level, the meso-level of the program, and the micro-level where assessments are enacted by teachers and assessors. Shih (2023) presented the decision-making involved in the creation and termination of a reform-driven language test in Korea, holding the test provider accountable for its failure: "What if test providers had documented the claim of social impact, seriously addressed the validity arguments with both warrants and rebuttals, empirically explored counter-evidence to the negative impact disseminated in the media, and persuaded the public, colleagues, and policymakers of the NEAT's feasibility?" (p. 16).

Research of this kind, nonetheless, does not represent the mainstream of the field. Issues such as test misuse, malpractice, and bias are often discussed, but seldom systematically examined or documented. To better understand and evaluate the social functions of language tests and ensure their sustainable development, decision-making needs to be better theorized and more clearly located in the social context.

Conceptualizing and researching strategic decision making

To explore decision making from a broader socio-political perspective, I turn to the management literature for insights and inspiration on how strategic decisions are conceptualized. In project management, corporate and business management, or management of organizations of any types, a three-tiered hierarchy of decisions has been well established, that is, decisions can be made at the strategic, tactical, and operational levels (e.g., Ackoff, 1990; Demeulemeester, et al., 2007; Harrington & Ottenbacher, 2009; Khalifa, 2020, 2021; Nutt & Wilson, 2010). There is, however, no straightforward answer to the question "what makes a decision strategic". Nutt and Wilson (2010) admitted that, over the past 50 years, the term "strategic" has become "more confusing than enlightening"

(p. 4). In their historical overview of strategic decision-making studies, the focuses of research in different periods were summarized. An important feature of strategic decision making, as identified in the summary, is that policymakers "emphasize the social practice of decision making" and "have competing interests that prompt key players to use political pressure to ensure that a choice aligns with their preferences" (p. 3-4). It was also observed that in the new millennium, the "*strategy as practice*" approach is becoming more prevalent (p. 6). This activity-based approach highlights the importance of the situational contexts of decision making, in addition to the macro social, political, and economic contexts in which organizations are embedded.

Khalifa (2021) also conceded that the term "strategic" is "not only one of the most widely used adjectives in business but also one of the most overused and abused" (p. 381). Based on an extensive literature review, Khalifa (2021) weighed the strengths and limitations of various approaches to conceptualizing strategic decisions and proposed to draw on the military literature, in which "strategy is about winning wars, grand tactics are about winning campaigns, and tactics are executed to win battles" (p. 387). Based on his earlier work (Khalifa, 2020), strategy was defined as "an entity's evolving theory of winning high-stakes challenges through power creating use of resources and opportunities in uncertain environments" (p. 389).

It is disappointing to note that decision theorists have not come to an agreement on the conceptual framework of strategic decision making, probably because decision makers from vastly different domains may not share the core principles, values, or priorities embedded in the decisions to be made. Research of strategic decision making, therefore, is particularly challenging due to the lack of a coherent theory. Nutt and Wilson (2010) observed that the descriptive tradition dominates the field, whereas prescriptive work is less favored, because theoretical frameworks are not mandatory in a descriptive research paradigm. Case study, representing the descriptive tradition, is the most frequently used method in an empirical exploration of decision making because of its rich description of the specifics of what has happened.

Research of strategic decision making in language testing is also challenging. In the foreword to a special issue on negotiating tensions between language assessment policies and practices (Elder, 2021), McNamara (2021) commented that "(O)ur policy-centred field is slowly awakening to a self-consciousness of its character and articulating the dilemmas and challenges that this new awareness brings" (p. 1). He suggested that "(R)eflection on the collective experience of those engaging explicitly with policy contexts may suggest some useful ways forward, even if the theoretical issues remain for the moment intractable" (ibid.). The purpose of this article, therefore, is to delve into decision making at the strategic level by drawing on our experiences of developing and validating a language test of a super large scale and very high stakes.

In this article, to examine decision making at the macro-societal level in large-scale language testing, Khalifa's (2020, 2021) definition of strategy would be adopted, which sees strategy as a top-level decision to address the thorniest and most difficult challenge in corporate management by using the resources available. To be specific, strategic decision making in large-scale language testing refers to policymaking to resolve the most intricate and complex issues facing the development and reform of language tests by leveraging resources available to key stakeholders. Methodologically, from the *strategy as practice* perspective (Nutt & Wilson, 2010), a case study approach is adopted to examine the activities that stakeholders engage in when decisions on the major reforms of the CET were made. The analysis would take into consideration the macro-level socio-political context and the specific situational context of each case. As Chair of the NCETC, I paid special attention to the role of professionals in strategic decision making.

Strategic decision making in the College English Test

To streamline the case analysis, I first introduce the current management structure of the CET and its key stakeholders (see Figure 1). At the top of the structure is the Ministry of

Education (MOE), which makes educational policies at the national level. Before 2005, the Department of Higher Education (DHE) of the MOE was responsible for the CET program. In 2006, due to an adjustment of government functions, the managerial responsibility was transferred to the National Education Examinations Authority (NEEA). Over the past two decades, the NEEA, in its role as the test provider, has been responsible for making major decisions on the reform of the CET. The test developer is the National College English Testing Committee (NCETC), a professional organization consisting of about 30 professors from different universities across the country. Under the supervision of the NEEA, the NCETC works with tech providers on the design and delivery of the testing program, including task design, item writing, test delivery, rater or examiner selection and training, scoring, score equating, and score reporting. Provincial or municipal education examinations authorities oversee test administration and delivery. Each university acts as a test center, responsible for delivering the test to their students. Apart from students and teachers, CET results are also used by admissions officers and employers, for selecting talents from a large number of applicants.

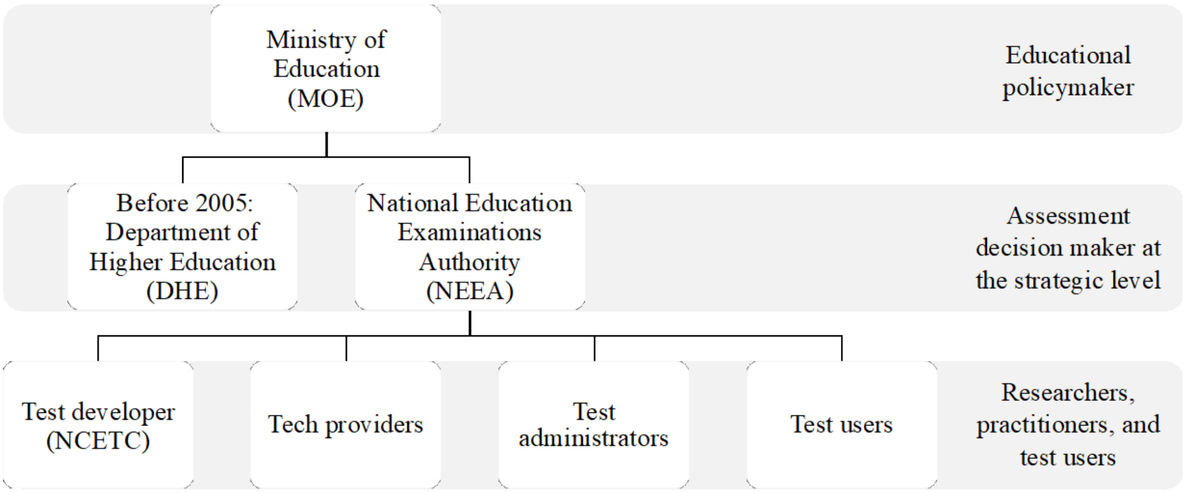
Ministry of Education (MOE)
Educational policymaker
Before 2005: Department of Higher Education (DHE)
National Education Examinations Authority (NEEA)
Assessment decision maker at the strategic level
Test developer (NCETC)
Tech providers
Test administrators
Test users
Researchers, practitioners, and test users

Figure 1. The CET management structure and key stakeholders

In the next section, I aim to unravel the complexities involved in decision making when issues concerning unintended test use, accessibility, malpractice, and automated scoring are addressed. In the analysis of each case, I seek to answer the following questions:

1) What was the problem and how was it identified?
2) How was the decision made and implemented?
3) What were the outcomes of the decision?

Case analysis

Case 1: Test use for unintended purposes

Problem identification

Like many of the products we develop and use in life, when a testing program has gained social recognition, it will have its own life, evolving with its users and adapting to their changing needs. The implementation of China's open-door policy in the late 1970s called for an urgent need to improve college students' level of English proficiency. Reading in English, for example, was seen as an essential skill for college graduates to keep abreast of the latest developments in science and technology. In the National College English Teaching Syllabus, the goal of the speed of reading was set at 70-100 words per minute (wpm) (State Education Commission, 1985, 1986). A national survey, however, showed that the average reading speed of college students at the time was only 17 wpm (Yang & Weir, 1998). The CET, therefore, was developed in the mid-1980s by a group of professors to check whether college students had met the curricular requirements and to promote the implementation of the national teaching syllabus.

Since the inception of the CET in 1987, there had been a steady rise in college students' level of English language proficiency, including their speed of reading in English, as demonstrated in their performances on the CET (Jin & Yang, 2006). The progress was attributed to teaching and learning, with the CET acting as a driving force for both teachers and students. With an increasing number of test takers each year, the CET program caught the attention of other users such as graduate program admissions officers and employers, and it began to be used for various gate-keeping functions, which were not envisaged or intended originally by the test developer. The high-stakes uses of the CET imposed huge pressure on College English teachers and students, leading to the negative washback of the test on teaching and learning (Wang, 2008).

A thornier issue that came with the high stakes was cheating or other forms of malpractice. The CET was originally targeted at registered college students and delivered on campus only. However, in the late 1990s, there was a surge in the demand for the CET certificate, particularly from college graduates who had not passed the test while studying at college, because the certificate could give their job application an edge and it was a steppingstone on their path to promotion. In the early 2000s, non-university-based test centers were set up to cater for the needs of college graduates. These centers, being financially independent and largely profit-driven, charged test takers a much higher registration fee. Even worse, there were frequent reports from inspectors on cheating through impersonation or copying neighbors' answers in these non-university-based test centers. Due to the lack of supervision from test administrators in colleges and universities, it was impossible to impose stricter regulations in these centers.

Decision making and implementation

The NCETC, the test developer, reported to the Department of Higher Education (DHE), the top-level assessment policymaker (see Figure 1 above), on the risks and consequences of test use for unintended purposes and the malpractice in non-university-based centers and discussed with the DHE the solutions to resolve the issue. As the National Education Examinations Authority (NEEA) was about to take over the managerial responsibility for the CET in 2006, it was also involved in the discussion. The DHE suggested that the test's score reporting system be reformed by ending the reporting of results as "Pass" or "Distinction". A new score reporting system was designed, providing a total score and several sub-scores. Through the implementation of the new system, colleges and universities were anticipated to formulate more reasonable requirements of the CET that were befitting to their specific contexts of teaching and learning. A press conference was called by the MOE to release the reform policies. At the conference, a Vice-Minister of the MOE introduced the reform initiative and the director of the DHE explained the reform policies. I, as Chair of the NCETC, took questions from news reporters and delineated the details of the measures to be taken. Among the package of decisions released at the press conference, the following three were intended to tackle the issues related to unintended test use and malpractice:

1) Close non-university-based test centers so that only registered students are eligible to take the CET.
2) Revise the CET score reporting system by replacing the certificate of “Pass” or “Distinction” with total and component scores.
3) Supervise and urge universities to re-evaluate their policy of using CET scores as a minimum requirement for graduation.

Outcomes of the decisions

In this case, the decisions were considered “strategic” because they were made by the top-level assessment policymaker to ensure the sustainable development of the CET. Throughout the process of decision making, the NCETC was actively involved in identifying and reporting the problems, explaining the reform policies to stakeholders, and implementing the decisions. Table 1 is a summary of the major changes brought about by the package of decisions.

Table 1. Outcomes of the reform to prevent unintended test use and malpractice

	Before the reform	After the reform
Closing non-university-based CET test centers	• Some college graduates took the test in non-university-based centers and cheated on the test.	• Cheating in non-university-based test centers was eliminated. • Only registered students are eligible to take the CET.
Revising the CET score reporting system	• Report 'Pass' or 'Distinction' in the CET certificate. • A pass certificate of the CET was required for graduation in many universities. • Universities pursued high pass rates regardless of their specific contexts of College English teaching and learning.	• Report total and sub-scores in the CET score report. • The requirement of a CET pass for graduation was removed by some universities. • Comparisons of pass rates cannot be made among universities. • Students tend to repeat the test to get higher scores.

The decision-making process was not without controversy, as the policies may infringe on the interests of some stakeholders. The closure of non-university-based test centers met opposition from students who were unable to pass the test prior to their graduation. The decision also provoked discontent among managers of non-university-based test centers, who were denied the chance to profit from the repeaters. However, the implementation of the decision has maintained the fairness of the CET and enabled the test to better fulfill its designated purpose. The revision of the score reporting system attained its objective of preventing simple comparisons of pass rates among universities (Jin & Yang, 2006). In some universities, a minimum CET score was removed from the requirements for graduation. However, with the change of scoring reporting system, the CET scores became difficult to understand for test users. Employers, for example, sought information on the cut-off scores of the CET when recruiting college graduates. A need arose for the test developer to re-educate stakeholders on score interpretations. Another unintended consequence of replacing a pass or distinction certificate with scores was that students were more inclined to repeat the test, often without further learning or preparation, in the hope of getting higher total or sub-scores.

Case 2: Accessibility of the speaking test

Problem identification

The second case concerns the reform of the CET–Spoken English Test (CET-SET), focusing on the conflict between the accessibility of the speaking test and its construct representation. In the mid-1990s, the CET-SET was developed in a face-to-face, group interview format, in which three test takers and two examiners formed a test group, performing a number of individual and group tasks. The format was considered to be most suitable for assessing interactional competence (Jin, 2010; 2020). By 2012, 58 CET-SET test centers had been established, and some 2000 examiners authorized. However, the maximum number of test takers each year remained about 100,000, due to the capacity limitation and a shortage of qualified examiners. Only those students who achieved a designated score in the CET written tests (e.g., 550 in the CET Band 4) were eligible to take the speaking test. Given the large number of CET test takers, the face-to-face format of the CET-SET raised serious concerns about its accessibility to the targeted test takers.

Decision making and implementation

Over a decade after its inception, the face-to-face test was abolished and replaced by the computer-based CET-SET in 2013. The decision on launching a computer-based speaking test was not straightforward because of the NCETC's concern with the lack of human interaction in a computerized test format. The NCETC argued that human interaction was essential to construct validity, whereas non-interactive, monologic tasks were typically employed in computerized speaking tests. The test provider, i.e., the NEEA, as well as some NCETC committee members, maintained that the CET-SET could only be a fair test if it was accessible to a wider range of college students. To resolve the conflict between construct representation and test accessibility, alternatives were explored. Research was conducted on a tape-mediated speaking test in the early 2000s, but the format was not put into operational use, due to logistical difficulties in test delivery and scoring (Jin & Guo, 2002). Surveys were conducted among teachers and students, who had mixed perceptions. While many liked the face-to-face format, some preferred a computerized format.

The final decision on replacing the face-to-face test with the computer-based CET-SET was made by the NEEA. To implement the policy, national and local examinations authorities set up new test centers. The NCETC collaborated with a tech company to develop the testing platform. With the support of the tech company, an online paired format was developed by the NCETC so that the construct of interactional competence could be adequately represented in the computer-based speaking test (Jin & Zhang, 2016; Zhang & Jin, 2021). The NCETC also developed a lower-level speaking test, the CET-SET Band 4 (CET-SET4), to accommodate the needs of students with a lower level of English-speaking proficiency (Zhang, 2022).

Outcomes of the decisions

Weighing the pros and cons and after careful deliberation, the NEEA made the strategic decision of moving away from the face-to-face testing of speaking towards computer-based testing. The NCETC, in collaboration with the tech company, developed the computer-based CET-SET to address the social need for university graduates with a higher level of oral English proficiency. At the operational level, the computerized speaking test has ensured construct representation, cost-effectiveness, and most importantly, the accessibility of the speaking test to the targeted test takers. In the process of decision making, the NCETC was actively engaged in the exploration of solutions through research and collaboration with the tech company so as to avoid construct underrepresentation (Jia, 2016). Table 2 is a summary of the changes taken place as a result of the reform.

Table 2. Outcomes of the CET-SET reform

	Before the reform	After the reform
Replacing the face-to-face CET-SET with the computer-based CET-SET4 and CET-SET6	• The administration of the face-to-face CET-SET was extremely resource-intensive. • The CET-SET was only accessible to a very small proportion of the targeted test takers. • The interlocutor (oral examiner) in the face-to-face CET-SET might introduce construct-irrelevant variance.	• The computer-based CET-SET is accessible to a wide range of test takers, and the test now has two levels: CET-SET4 and CET-SET6. • The construct of interactional competence is retained by the use of a paired discussion task. • The interlocutor effect is controlled by using a pre-recorded video for giving instructions. • Need for a large number of items and qualified raters.

Since the launch of the computer-based CET-SET, the speaking test has become more accessible to college students (Zhang, 2022). Before the pandemic, over 300 test centers were set up and the annual test population increased from 100,000 to over 1 million in 2018. A further strength of the computer-based CET-SET is its effective control of the interlocutor effect. Using a pre-recorded video to give instructions, the construct-irrelevant variance has been removed. There is nonetheless room for improvement: the test assesses audio-based, non-face-to-face interaction, which may differ, in important ways, from online interaction where speakers can see each other via video cameras (Zhang & Jin, 2021). Ockey and Neiriz (2021) argued that synchronous assessments with mediated visual presence most closely mirror the real-life speaking construct. The other challenge facing the computer-based test is the need for a large number of items and a large pool of well-trained raters. The scoring process is also labor-intensive and time-consuming. At this stage, except for the read-aloud task in the CET-SET4, which is automatically scored by computer, performances on the CET-SET are double scored by human raters.

Case 3: High-tech cheating and test security

Problem identification

Cheating is always a vexing problem for large-scale, high-stakes tests. The third case is about the prevention of cheating in both the traditional paper-based CET and the internet-based CET (IB-CET). Unlike cheating on an individual basis in non-university-based test centers as discussed in Case 1, in the paper-based CET, a so-called “business model of high-tech cheating” was developed: before the test, cheaters paid for “the service” and purchased cheating equipment; during the test, the cheaters took pictures of test items using mini cameras and sent pictures to “ghost test takers” hired by service providers to answer the questions; service providers sent the answers back to the cheaters, who received messages via mini-earphones. Jin (2014) described this kind of cheating as a crime jointly plotted by test takers and “service providers”. In the early 2010s, high-tech, mass cheating became a major concern of the NEEA.

Malpractice in the IB-CET, a new product of the CET testing series, was a different story. The project was initiated and funded by the Department of Higher Education (DHE) in the mid-2000s. The IB-CET was designed and developed by the NCETC in close collaboration with a high-tech company. In the IB-CET, speaking became a compulsory component, accounting for 15% of the total score; both audio and video materials were used as listening inputs; and a summary writing task was included, further broadening the test’s construct. During the field test, surveys were conducted to elicit views from students and teachers, who perceived the IB-CET positively (Jin & Wu, 2009, 2010). The NCETC, however, was concerned about cheating and the security of the item bank. Since universities were not involved in the administration of the IB-CET, the tech company was in charge of registration and oversaw test administration. Cases of impersonation and taking screenshots of items were reported to the NCETC, posing grave threats to the test’s fairness and security.

Decision making and implementation

To curb cheating in the paper-based CET, the NEEA made the decision to implement multiple forms and multiple versions (MFMV) (Jin & Wu, 2017). Here, "form" refers to different content (texts, input materials, questions), and "version" refers to the same content configured in different ways through re-arranging the order of texts and the options of a multiple-choice question. In each test administration, over a dozen test forms and versions are used. Test takers do not know which form or version they are taking, thus making it impossible for ghost test takers to send answers to a large number of test takers. To implement the decision, the NCETC is responsible for developing and equating multiple forms and versions and scoring multiple constructed-response tasks. The NEEA and local education examinations authorities are responsible for delivering the test and organizing scoring sessions.

As for the IB-CET, after receiving repeated reports of malpractice, the NCETC submitted reports to the DHE and the NEEA, requesting urgent measures to tackle the issues of cheating and item bank security. After consulting the project manager, i.e., the DHE, the NEEA made the final decision to suspend the IB-CET in 2015.

Outcomes of the decisions

In this case, using multiple forms and multiple versions was considered a decision at the strategic level, because high-tech, mass cheating, if left unchecked, would have a far-reaching impact on the testing program, undermining the integrity of the paper-based CET. While generally supportive of the decision, the NCETC had serious concerns over the practicality of the policy, due largely to the challenges such as expanding the item pool, ensuring the consistency of scoring multiple writing and translation tasks, and equating multiple test forms. Logistically, test paper production and distribution would be extremely complicated, given the large scale of the test. The decision to suspend the IB-CET was also a difficult compromise: putting a halt to an innovative project seven years after its inception is a great loss to the test provider, the test developer, and the tech company, considering the enormous effort that had been put into the development of the IB-CET. Table 3 is a summary of the changes that have taken place since the measures were taken.

Table 3. Outcomes of the measures to prevent malpractice in the CET and IB-CET

	Before the reform	After the reform
Adopting multiple forms and multiple versions in the paper-based CET	• There were widespread cases of high-tech cheating organized and provided by ill-intentioned "service providers".	• By using multiple forms and multiple versions, high-tech, mass cheating was effectively prevented. • The implementation of the MFMV policy is labor-intensive.
Suspending the IB-CET project	• Due to lax management, test takers resorted to impersonation to cheat on the IB-CET. • There were reports of the leakage of test items in the IB-CET.	• The IB-CET was suspended at the expense of an innovative project. • Security and fairness of the CET as a brand has been maintained.

During the process of decision making, the NCETC remained on high alert, identifying and reporting cheating and security risks to the test management department (i.e., the DHE) and the NEEA. Since the implementation of the MFMV policy, high-tech, mass cheating has been effectively prevented. This is evidenced by post-test statistical analyses, media reports, and local inspectors' reports. The decision to suspend the IB-CET has protected the brand of the CET as a fair and secure testing program and restored users' confidence in the CET, though the decision was taken at a high price.

Case 4: Automated scoring

Problem identification

Performance-based tasks could have a more positive washback on teaching and learning than selected-response items (Yu, 2013). Aside from a separate speaking test, the CET has two performance-based tasks: essay writing (30 minutes) and paragraph translation from Chinese into English (30 minutes). The workload of scoring is enormous. In the past decade, after each test administration, about 3000 trained raters in 12-13 scoring centers across the country work for over a week to complete the scoring of 10 million scripts of essay writing and 10 million scripts of paragraph translation. On top of the onerous workload, it is also demanding to recruit and train a large number of competent raters. And it is a costly endeavor to employ human raters to score millions of scripts.

Decision making and implementation

With the support of AI technologies in recent years, the use of automated scoring systems has been gaining traction in order to enhance scoring efficiency and consistency. In 2016, the NEEA signed a strategic contract with a high-tech company to develop CET automated scoring systems. Since the NEEA set the goal of using automated scoring to improve scoring efficiency and quality, the NCETC has been working with the tech company to develop and evaluate scoring models. There are mainly two types of automated scoring systems. One is the machine scoring model, in which a scoring machine is used as the sole rater. The other is a hybrid model, in which a scoring machine is used as either a check rater or a second rater. As a check rater, the machine scoring system is a quality control measure, and human scores are used as the final scores. As a second rater, the machine scoring system generates a score, which is factored into the final score.

In the process of developing the CET automated scoring systems, decisions need to be made as to which model should be adopted. The CET committee is responsible for setting the standards for using machine scoring as the sole rater or check rater. The tech company is expected to meet these standards through technological innovations. Take the translation task as an example. Currently, machine scoring is used as the sole rater for lower-level scripts (0 to 6 points out of 15). The standards for the scoring engine are 1) the accuracy rate of text recognition is above 97%; 2) the human-machine agreement rate on the performance level (Level 1 to 5) is higher than 95%; 3) the discrepancy between human and machine scores is less than 2 points (out of a total of 15); 4) the human-machine correlation is higher than 0.9.

Outcomes of the decision

Since the decision on developing CET automated scoring systems was made, the NCETC has been collaborating with the tech company to develop and evaluate the scoring engines. Table 4 summarizes the changes expected to take place as a result of the decision on CET automated scoring.

Table 4. Expected outcomes of CET automated scoring

	Human scoring	Automated scoring
Decision on developing CET automated scoring systems	• Scoring of CET Writing and Translation and the CET-SET is time-consuming and resource-intensive. • There is a need for a large number of qualified raters.	• A hybrid model is used in the CET-SET4 and Translation, improving the scoring efficiency. • Research is being conducted to improve CET automated scoring engines. There is also an urgent need to investigate score interpretability and the potential impact of automated scoring on teaching and learning.

In this case, the decision of developing CET automated scoring systems was made by the NEEA to improve scoring efficiency and validity. The decision was in fact met with mixed reaction from the NCETC. While the majority agreed that the technology could make the scoring process faster, possibly with a higher level of consistency, the NCETC was concerned about the validity of automated scoring and the interpretability of automated scores.

Human-machine correlations or agreement rates could provide necessary but insufficient validity evidence. Automated scoring is also confronted with fairness challenges due to potential biases in scoring algorithms. Regarding score interpretation, explainable AI has yet to be achieved by a scoring system based on deep learning technologies. Automated scoring may also have a negative impact on teaching and learning. Test takers may employ test taking strategies in order to outwit the machine and secure a better score. Empirical data have been collected to understand CET test takers' perceptions of automated scoring systems (Hong, 2022; Jin et al., 2017, 2020). Results show that the test takers lacked a sufficient knowledge of automated scoring, and their cognitive process of writing may be impacted by automated scoring. On-going research is directed towards improving score interpretability and the impact of automated scoring on teaching and learning.

Discussion

Strategic decision making as a "situated" socio-political activity

Strategic decisions, in Child et al.'s (2010) view, "involve a political problem of reconciling divergent interests as well as a technical problem of attempting to calculate the best decision given a number of parameters" (p. 105). In this article, stories of some major reforms of the CET were reconstructed to gain a deeper understanding of the contextualized nature of strategic decision making for high-stakes language testing and the intricate dynamics between different players. The analysis of the multiple cases shows that strategic decision making in language testing is a "situated" socio-political activity (Nutt & Wilson, 2010). From their *"strategy as practice"* perspective, "any particular action by managers must be seen and understood in the context of the situation in which that action occurs" (p. 7). Reform policies, therefore, should be made and understood by taking into consideration the socio-political context and case-specific situational contexts.

In China, large-scale language testing is operated through a centralized management system with a hierarchy of decision-making structure. Final decisions on reform strategies typically come from the top-level assessment policymaker such as the DHE and NEEA in the cases presented in this article. The hierarchical system has empowered the policymaker to make difficult decisions, even in cases where the interests of some stakeholders may be compromised. For example, the closure of non-university-based test centers could be quite difficult in a different political system. The decision, in fact, was met with opposition from university students, in particular those who were unable to achieve a satisfactory score during their studies at the university. Test center managers who had been profiting from these test takers also strongly opposed the policy. In the views of the test provider (i.e., the NEEA) and the NCETC, however, it was considered essential to ensure the test's fairness and reinstate its educational function.

The social responsibility of language testing professionals

Language testing professionals have long embraced an ethical perspective of the profession, which underscores the individual responsibility of language testers (Bachman, 2000). A traditional approach to ethical language testing practice, according to McNamara (2000), "limit(s) the social responsibility of language testers to questions of the professional ethics of their practice" (p. 75). It was argued that language testing as a socio-political endeavor necessitates an expanded sense of responsibility, which sees ethical practices as "involving test developers in taking responsibility for the effects of tests" (p.72). From this broadened viewpoint, the developer of a language test needs to be accountable to the people immediately affected by the test, mainly the test takers and the test users, as well as its washback on teaching and learning and impact on the community as a whole.

In the case of the CET-SET, the strategy to move away from the face-to-face interview test towards a computer-based speaking test was driven by the government's intention to expand the test's accessibility. The NCETC was initially divided on this transition: some agreed that test accessibility should be the primary concern whereas others contended that construct validity should be equally, if not more, important than test accessibility. Nonetheless, once the policy was formulated, the NCETC, in close collaboration with the tech company, made strenuous efforts to implement the strategy by adopting an innovative design, i.e., computer-based speaking test in a paired format, which is capable of tapping into test takers' interactional competence (Zhang & Jin, 2021). As for the IB-CET, a pioneering project spearheaded by the NCETC, a decision was made by the government departments (i.e., DHE and NEEA) to suspend the internet-based test when its fairness and security were at risk. The whistle blower was actually the NCETC, who informed the policymakers of the malpractice and urged for immediate action. Although the issue was addressed at the cost of a project with great potential, from an ethical stance, the decision was deemed a responsible one, for the CET test takers and the test users.

A critical perspective of strategic decision making

Compared to an ethical view of language testing, critical language testing is "a much more radical view of the social and political role of tests" (McNamara, 2000, p. 76). From a critical perspective, language testing gets redefined "in socio-political terms" and is seen as an "exercise of power" (ibid.). The underlying belief of this view is that "the principles and practices that have become established as common sense or common knowledge are actually ideologically loaded to favour those in power, and so need to be exposed as an imposition on the powerless" (ibid.). Critical language testing requires that professionals move beyond ethical considerations and strive to empower the stakeholders most vulnerable to the decisions of those in power.

One of stakeholder groups to be empowered are learners or test takers. The analysis of the cases presented in this article reveals that there is a discrepancy between the belief that learners should be the driving force behind all of our endeavors and the reality that decisions, particularly those at a strategic level, are often made in a top-down approach without hearing the voice of learners. When automated scoring systems were to be developed, for instance, the decision was made with the best of intentions, yet without much thought towards how it might impact learners' motivation to write or speak and the cognitive processes involved in task completion. Survey results have shown that college students had insufficient knowledge about automated scoring, and as a result, when writing or speaking to the machine, they may be engaged in different cognitive processes with the expectation of boosting automated scores (Hong, 2022; Jin, et al., 2017). Jin and Fan (in press) addressed the oversight of the impact of technological innovation on test takers and provided some practical guidance on test taker engagement in AI technology-mediated language assessment.

Shohamy (2001) pointed out in her seminal work *The Power of Tests* that "test takers have no say about the content of tests and about the decisions made based on their results; worse, they are forced to comply with the demands of tests by changing their behaviour in order to succeed on them" (p. 375). Echoing her view, Jin (2023) noted in an editorial of a virtual special issue on test takers' insights for language assessment that "in large-scale language testing, test takers are mainly viewed as the target of measurement, rather than active participants whose insights are welcomed and voices heard" (p. 193-194). Based on her review of studies on test takers' experiences of using English testing for immigration purposes in Australia, Frost (2021) also called for a renewed criticality in language testing by focusing on the expectations of test takers as the stakeholder group to whom we must be primarily accountable.

Conclusion

Decision making in large-scale testing, as documented in this article, involves policy discussion, professional requirements, ethical considerations, and practical constraints. At the strategic level, decisions are typically made by policymakers (e.g., governmental departments) to re-set the agenda. To bring about desirable effects of strategic decisions, professionals of large-scale language testing have a critical role to play. This article has showcased the role of the NCETC in strategic decision making and implementation. Over the decades, the NCETC has stayed alert to the threats to the test's validity and fairness and made every effort to communicate with policymakers. After decisions have been taken, it strives to explain the policies to stakeholders, collaborate with tech companies, carry out validation research, and monitor the test's washback and social impact. However, the views presented in the article by an insider could be one-sided. To gain a more comprehensive and nuanced understanding of decision making, it is beneficial to take into account a diverse range of perspectives, including those of test takers, language teachers, tech providers, and test users.

Acknowledgements

This article was based on the Cambridge/ILTA Distinguished Achievement Award lecture presented at the 43rd Language Testing Research Colloquium (online) in March 2022. I would like to thank the nominator(s) and the award committee for selecting me as the recipient of this prestigious award. I would also like to thank Professor Antony Green, Professor Guoxing Yu and the two anonymous reviewers for their insightful and constructive comments on the manuscript.

References

Ackoff, R. L. (1990), Strategy. *Systems Practice, 3*(6), 521-524. https://doi.org/10.1007/BF01059636

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42. https://doi.org/10.1177/026553220001700101

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice.* Oxford University Press.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing, 33*(4), 453-472. https://doi.org/10.1177/0265532215593312

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple… *Language Testing, 29*(1), 19-27. https://doi.org/10.1177/0265532211417211

Chapelle, C. A. (2020). An introduction to Language Testing's first Virtual Special Issue: Investigating consequences of language test use. *Language Testing, 37*(4), 638-645. https://doi.org/10.1177/0265532220928533

Child, J., Elbanna, S., & Rodrigues, S. (2010). The political aspects of strategic decision making. In P. C. Nutt & D. Wilson (Eds.), *The Handbook of Decision Making* (pp. 105-137). Wiley.

Demeulemeester, E., Deblaere, F., Herbots, J., Lambrechts, O., Van de Vonder, S. (2007). A multi-level approach to project management under uncertainty. *Tijdschrift voor Economie en Management, LII*(3): 391-409.

Elder, C. (2016). Introduction to Special Issue (Evaluating language assessment programs and systems in use). *Papers in Language Testing and Assessment, 5*(1), iii-viii.

Elder, C. (Ed.) (2021). Negotiating tensions between language assessment policies and practices: The role of the language testing professional. A celebration of 30 years of work at the University of Melbourne's Language Testing Research Centre (1990-2020). *Papers in Language Testing and Assessment, 10*(1), *Special Issue*.

Fan, J. & Frost, K. (2022). At the intersection of language testing policy, practice, and research: An interview with Yan Jin, *Language Assessment Quarterly*, 19(1), 76-89. https://doi.org/10.1080/15434303.2021.1938570

Frost, K. (2021). Negotiating the boundaries of responsibility: Rethinking test takers and the ethics of testing. *Papers in Language Testing and Assessment, 10*(1): 70-83.

Fulcher, G. & Davidson, F. (2009). *Test architecture, test retrofit. Language Testing, 26*(1): 123-144. https://doi.org/10.1177/0265532208097339

Harrington, R.J. & Ottenbacher, M. C. (2009). Decision making tactics and contextual features: Strategic, tactical and operational implications. *International Journal of Hospitality & Tourism Administration, 10*(1), 25-43.

https://doi.org/10.1080/15256480802557259

Hong, W. (2022). *An investigation of test taker perceptions of automated essay scoring and test-taking strategy use*. [MA thesis, Shanghai Jiao Tong University]. China.

Jia, G. (2016). On the intended washback of English speaking tests: The case of College English Test – Spoken English Test (CET-SET). *Foreign Language Testing and Teaching, 4*: 1-9+51.

Jin, Y. (2010). The National College English Testing Committee. In L. Cheng & A. Curtis (Eds.), *English Language Assessment and the Chinese Learner* (pp. 44-59). Routledge, Taylor & Francis Group.

Jin, Y. (2014). The limits of language tests and language testing: Challenges and opportunities facing the College English Test. In D. Coniam (Ed.), *English Language Education and Assessment: Recent Developments in Hong Kong and the Chinese Mainland* (pp. 155-169). Springer Singapore.

Jin, Y. (2020). Testing tertiary-level English language learners: The College English Test in China. In L. I-W. Su, C. J. Weir, & J. R. W. Wu (Eds.), *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts* (pp. 101-130). Routledge.

Jin, Y. (2023). Test-taker insights for language assessment policies and practices. *Language Testing, 40*(1), 193-203. https://doi.org/10.1177/02655322221117136

Jin, Y. & Fan, J. (in press). Test taker engagement in AI technology-mediated language assessment. *Language Assessment Quarterly*.

Jin, Y. & Guo, J. (2002). Research on the validity of the CET Semi-direct Oral Proficiency Test. *Foreign Language World, 5*: 72-79.

Jin, Y., Wang, W., Zhang, X., & Zhao, Y. (2020). A Preliminary Investigation of the Scoring Validity of the CET-SET Automated Scoring System. *China Examinations, 7*: 25-33.

Jin, Y. & Wu, E. (2017). An argument-based approach to test fairness: The case of multiple form equating in the College English Test. *International Journal of Computer-Assisted Language Learning and Teaching, 7*(3): 58-72.

Jin, Y. & Wu, J. (2009). Design principles of the internet-based College English Test.

Foreign Language World, 4: 61-68.

Jin, Y. & Wu, J. (2010). A preliminary study of the validity of the internet-based CET-4: Factors affecting test-takers' perception of and performance on the test. *Computer-Assisted Foreign Language Education, 2*: 3-10.

Jin, Y. & Yang. H. (2006). The English proficiency of college and university students in China: As reflected in the CET. *Language, Culture and Curriculum, 19*(1): 21-36. https://doi.org/10.1080/07908310608668752

Jin, Y. & Zhang, L. (2016). The impact of test mode on the use of communication strategies in paired discussion. In G. Yu & Y. Jin (Eds.), *Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums* (pp. 61-84). Palgrave Macmillan.

Jin, Y., Zhu, B., & Wang, W. (2017). Writing to the machine: Challenges facing automated scoring in the College English Test in China. The 39th Language Testing Research Colloquium 2017, Bogota, Colombia.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2013) The argument-based approach to validation. *School Psychology Review, 42*(4), 448-457, https://doi.org/10.1080/02796015.2013.12087465

Khalifa, A.S. (2020). Strategy: Restoring the lost meaning. *Journal of Strategy and Management, 13*(1), 128-143. https://doi.org/10.1108/JSMA-09-2019-0158

Khalifa, A.S. (2021). Strategy and what it means to be strategic: redefining strategic, operational, and tactical decisions. *Journal of Strategy and Management, 14*(4), 381-396. https://doi.org/10.1108/JSMA-12-2020-0357

McNamara, T. (2000). *Language Testing*, Oxford University Press.

McNamara, T. (2021). Language testing within policy contexts: Conceptual and instrumental challenges, *Papers in Language Testing and Assessment, 10*(1): 1-3.

McNamara, T. & Roever, C. (2006). *Language testing: The social dimension.* Blackwell Publishing Limited.

Nutt, P. C. & Wilson, D. C. (2010). Crucial trends and issues in strategic decision making. In P. C. Nutt & D. C. Wilson (Eds.), *Handbook of Decision Making* (pp

3-29). John Wiley & Sons, Ltd.

Ockey, J. G. & Neiriz, R. (2021). Evaluating technology-mediated second language oral communication assessment delivery models, *Assessment in Education: Principles, Policy & Practice, 28*(4), 350-368, https://doi.org/10.1080/0969594X.2021.1976106

Shih, D. (2023). Critical discursive approaches to evaluating policy-driven testing: Social impact as a target for validation. *Language Testing, 0*(0). https://doi.org/10.1177/02655322231163863

Shohamy, E. (2001a). Democratic assessment as an alternative. *Language Testing, 18*(4), 373-391. https://doi.org/10.1177/026553220101800404

Shohamy, E. (2001b). *The power of tests: A critical perspective on the uses of language tests*. Pearson Education Limited.

State Education Commission. (1985). *College English Teaching Syllabus (Science and Technology)*. Shanghai Foreign Language Education Press.

State Education Commission. (1986). *College English Teaching Syllabus (Arts and Science)*. Shanghai Foreign Language Education Press.

Wang, H. (2008). A systems approach to the reform of College English Testing: Report on the "Survey of College English Testing Reform". *Foreign Languages in China, 4*: 4-12.

Weir, C. (2005). *Language testing and validation: An evidence-based approach.* Palgrave Macmillan.

Yang, H. (2003). The 15 years of the CET and its impact on teaching. *Journal of Foreign Languages, 3*: 21-29.

Yang, H. & Gui, S. (2007). The sociology of language testing. *Modern Foreign Languages (Quarterly), 4*: 368-374.

Yang, H. & Gui, S. (Eds.) (2015). *The Sociology of Language Testing*. Shanghai Foreign Language Education Press.

Yang, H. & Weir, C. (1998). *The validation study of the College English Test (CET).* Shanghai Foreign Language Education Press.

Yu, G. (2013). Performance assessment in the classroom. In A. Kunnan

(Ed.), *Companion to Language Assessment.* https://doi.org/10.1002/9781118411360.wbcla133

Zhang, L. (2022). Test review: College English Test – Spoken English Test (CET-SET). *Studies in Language Assessment, 11*(2): 164-180.

Zhang, L. & Jin, Y. (2021). Assessing interactional competence in the computer-based CET-SET: An investigation of the use of communication strategies, *Assessment in Education: Principles, Policy & Practice, 28*(4): 389-410. https://doi.org/10.1080/0969594X.2021.1976107

Zheng, Y. & Cheng, L. (2008). The College English Test (CET) in China. *Language Testing, 25*(3): 408-417. https://doi.org/10.1177/0265532208092433