O'Sullivan and Cheng
Language Testing in Asia (2022) 12:52
https://doi.org/10.1186/s40468-022-00201-5

REVIEW

Open Access

Lessons from the Chinese imperial examination system

Barry O'Sullivan[1*] and Liying Cheng[2]

*Correspondence: barry.osullivan@britishcouncil.org

[1] British Council, Newmount House, 22/24 Lower Mount Street, Dublin 2, Ireland
[2] Faculty of Education, Queen's University, Duncan McArthur Hall, 511 Union Street, Kingston, ON K7M 5R7, Canada

Abstract

In this paper, we set out to explore the world's first major standardised examination system. In the field of language testing and assessment, works such as measured words (Spolsky, 1995), measured constructs (Weir, Vidakovic & Galaczi, 2013), and Cambridge English exams — the first hundred years (Hawkey & Milanovic, 2013) all point to the fact that contemporary tests reflect many years of accumulated knowledge and practice. Perhaps more importantly, they also remind us of the social and educational impact of the tests we develop. With this in mind, we explore the very first example of a standardised examination system — the Chinese imperial examination system (the Kējǔ — in Chinese Hanyu Pinyin 科举).

Keywords: Chinese imperial examination system, Validity, Test development and administration, Test consequences, Integrated arguments approach, Historical interest

Introduction

In dynastic China, the Kējǔ served as a mechanism to select the empire's highest officials. The Kējǔ was the world's first merit-based examination system (Hu, 1984; Lai, 1970), the origins of which can be traced back nearly 2000 years to the Han dynasty (206 BCE to 220 CE). The examination system was first administered in 605 CE during the Sui dynasty and continued almost uninterrupted until it was finally abandoned in 1905 CE. Before we proceed with a brief overview of the Kējǔ, we should point out that the imperial examination system also had tests on military knowledge and skills, with successful candidates in the examinations becoming military officers. While this paper will not focus on these tests, they are also worthy of further exploration.

The essential function of the system was to identify those exceptionally talented male citizens who had mastered what was considered at the time to be the knowledge and skills to ensure the continuation of the empire. The most successful of these became the Mandarin (senior civil servants). As is the case with any long-running testing system, the contents and focus were to change over time, as were specific details relating to scoring. Nevertheless, the concept and function of the Kējǔ as well as much of the social and political value and ritual associated with the examinations would remain essentially the same until the system was abandoned shortly before the end of the imperial system in the early twentieth century.

As well as identifying appropriate individuals for the imperial civil service, the Kējǔ also came to be recognised as an instrument through which subjugated *barbarians or foreigners* were obliged to adopt Chinese culture and recognise imperial power, as can be seen in Wade's (2005) translation script from 1425:

> *If prefectural, sub prefectural and county instructors are appointed to guide the training of the native people, to provide correct models and to demonstrate the ways of education, then day by day the students will be thoroughly imbued with what they see and hear. Thereby, their 'man' and 'yi' [i.e., foreign or barbarian] ways will be changed and again they will take on the ways of the Chinese. This would certainly be a wonderful thing! "The Emperor said: "The ancients said: 'There should be no differentiation in employing the worthy.' They also noted: 'In teaching there should not be any distinction between classes.' Teachers act as models. Using Chinese ways to change barbarian ways--there is nothing more important than this. Order the Ministry of Personnel to select suitable personnel and make the appointments.*

This *civilising* practice spreads as the influence of the Chinese empire grew to encompass what Fairbank (1968) referred to as the Sinic zone, which was comprised of nearby territories sharing a similar culture, Korea, Vietnam, the Ryūkyū Islands (part of Japan since 1879), and for a relatively short time, Japan itself (Ko, 2017). Interestingly, the most obvious symbol of China's influence, the examination system, lasted longest in Korea (958–1894 CE) and then Vietnam (1075–1913 CE). However, the system was never fully adopted in Japan, most likely due to the long-established practice there of linking power to the status and social standing of a person's family or clan.

In the late sixteenth century, the Kējǔ came to the attention of European scholars, primarily through the work of Jesuit missionaries, and in time, it spreads to France, Germany, and the UK (Higgins & Zheng, 2002; Miyazaki, 1976) and later to the USA. However, despite its global influence, particularly in the areas of competitive examinations and educational testing, relatively little has been written in English about the examinations to date (Cheng, 2010). This paper does not attempt to offer a detailed history of the Kējǔ; such histories can be found elsewhere (e.g. Elman, 2000; Miyazaki, 1976). Instead, we will look back on the examination system from a modern perspective, not to evaluate the system as that would be unwise but to demonstrate how so many of the issues we currently deal with in the test development and administration of tests were understood and approached under the Kējǔ.

Figure 1 represents the general structure of the system with its four levels of examinations. Once a prospective candidate was accepted for the apprentice or county examination, they were awarded the title Tongsheng. Those who succeeded at this level received the title Shengyuan which allowed them to progress to the provincial examination. A pass at this level came with the title of Juren and brought with it the right to sit for the metropolitan examination. A pass here brought the title Gongshi and the automatic right to sit the palace examination and the possibility of attaining the ultimate title, Jinshi. The roles for which successful candidates were considered qualified to fill were strictly governed by the level of examination passed. Those not finding employment in the imperial civil service typically went on to become local administrators or tutors in local test preparation academies.

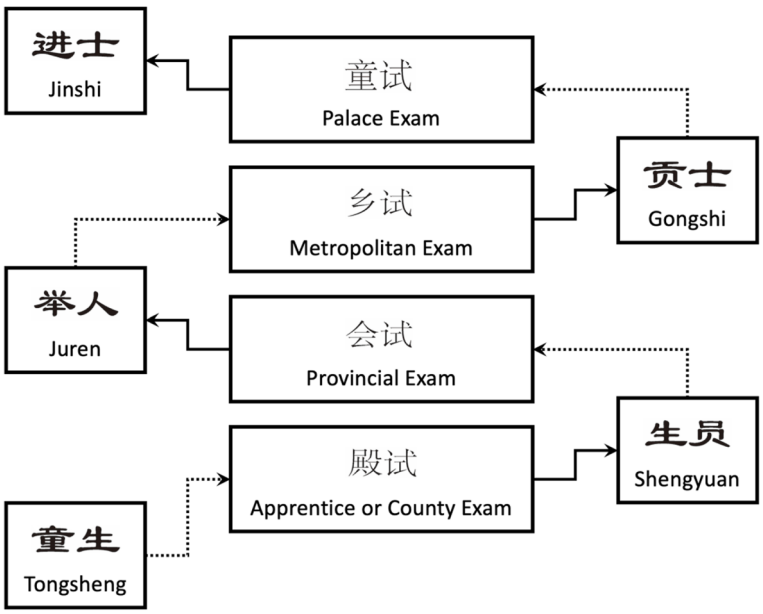

Fig. 1 The *Kējǔ* structure

The Chinese imperial examinations system: test use analysis

Setting the context

In the remainder of this paper, we will focus on the final iterations of the Kējǔ system that evolved during the Ming (1368–1644 CE) and Qing (1644–1912 CE) dynasties. To do this, we will draw on the most recent interpretation of the socio-cognitive model of test development and validation (Chalhoub-Deville & O'Sullivan, 2020), the integrated arguments (IA) approach. This entails looking at the examination system from four different perspectives, two related directly to the test (measurement and development) and two related to the consequences of the test (theory of action and communication). In the next section, we offer a brief overview of Chalhoub-Deville & O'Sullivan's (2020) approach.

However, before moving on the integrated arguments, we should point out that it is important to recognise the distinction between the Kējǔ as an examination and the Kējǔ as a system. In terms of the latter, it should be clear that the first two elements of the integrated argument (development and measurement) refer to the examination itself — where the examination is seen as a singular event. The second two elements (theory of action and communication) are less focused on the test itself than on the consequences or impact of the whole system on Chinese society.

The integrated arguments socio-cognitive model

The socio-cognitive approach to test development and validation was first called for by O'Sullivan (2000) and then outlined by O'Sullivan & Weir (2002 — a summary of which was republished in 2020). Weir's seminal publication in 2005 then sets out a series of frameworks, each focusing on a single language skill (e.g. speaking), which he saw as supporting practitioners in language test development and validation. The basic premise

of the socio-cognitive approach was that any test development project should be driven primarily by a clearly described language model while also taking into consideration the social context of the language use situation (e.g. the relative age, gender, acquaintanceship, or language level of interlocutors). In addition, the aim of such an assessment should be that tasks should strive to reflect real-world cognitive demands while maintaining an appropriate psychometric profile.

O'Sullivan (2011, 2016) began the process of defining the underlying model that drove the Weir frameworks, taking into consideration the context in which a test is developed and delivered as well as the critical area of test consequence, which he argued had never been adequately addressed in the literature. This work led to a collaboration with Chalhoub-Deville (Chalhoub-Deville & O'Sullivan, 2020) in which they proposed moving from a single structured validation argument (e.g. Messick, 1979; Bachman & Palmer, 1996; Kane, 1992, 2006, 2012, 2013) to a more complex series of arguments built around four models. The four models can be seen to focus on two main areas of interest:

- *The test*: This will be driven by a *development model* (to systematise the development process while ensuring that the test is supported by a clearly described language model) and a *measurement model* (to ensure that the resulting test has appropriate measurement characteristics (i.e. reliability, consistency, accuracy)
- *Consequence*: O'Sullivan (2016) argued that the context of development and use is defined by the stakeholders who populate the context. These stakeholders will represent a range of backgrounds, from test-takers, parents, teachers to education board members, policymakers, and influencers. The integrated arguments approach proposes two additional models to address the area of consequence to stakeholders. The first of these is the *theory of action* model, which aims to clearly outline to stakeholders the rationale for and approach to the development of a new (or revised) assessment. The idea here is to achieve buy-in from the beginning of the development process and to identify, and respond to, stakeholder concerns: in other words to attempt to predict the impact of the intervention on stakeholders. In order to ensure that communication with stakeholder groups is appropriate and targeted (in terms of language, channel, and content), we also need a clearly formulated *communication model*. This model is expected to identify the nature and channel of communication between the test developer and the stakeholders.

Figure 2 offers a graphical overview of the Chalhoub-Deville & O'Sullivan (2020) approach.

As mentioned above, the integrated arguments approach, as operationalised within the socio-cognitive model, expands the sources and types of evidence expected in a validation argument. It does this by integrating the evaluation of test-focused qualities (through the development and measurement models) with the evaluation of educational/social consequences. In practice, the approach allows for the developer to attend to both intended and unintended test consequences. In doing this, the IA approach offers us a unique framework through which any test might be evaluated.

Table 1 presents details on how we operationalise the IA approach in our analysis. We start with a brief overview of the theories of change and action since these represent the

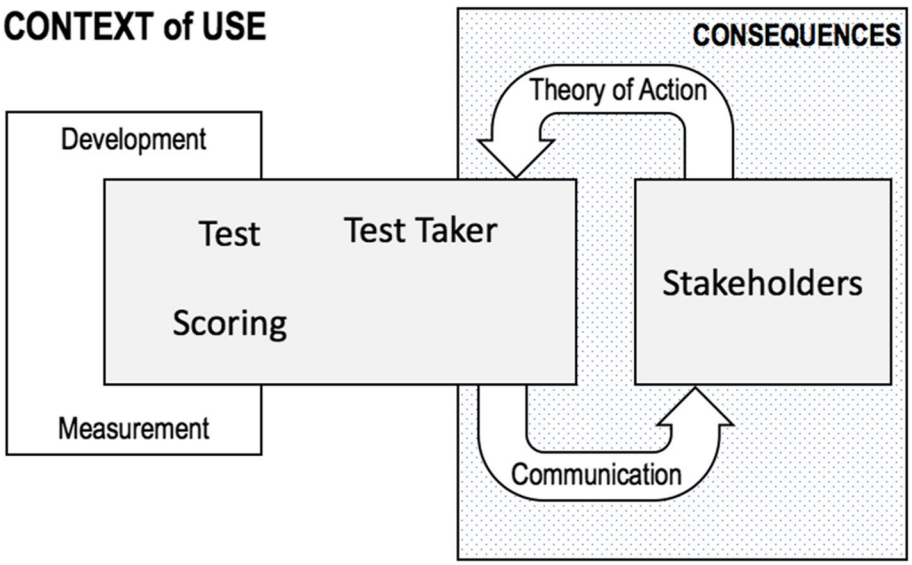

Fig. 2 Socio-cognitive model (based on Chalhoub-Deville & O'Sullivan, 2020)

Table 1 Operationalising the socio-cognitive model (Chalhoub-Deville & O'Sullivan, 2020)

Focus	Model	Detail
Focus on the test	Development	Test content, administration, and security
	Measurement	Scoring system innovation
Focus on test consequences	Communication	Reporting results and highlighting success
	Theories of change and action	Rationalising the need for a competitive system and outlining its approach

rationale offered by contemporary officials or implied by historians for the introduction of the Kējǔ. We then focus on the test itself before moving on to the communication model where we consider how the empire communicated the importance of the system to its people. We also take some time to explore the broader historical and ongoing impact of the system within and outside of China.

Theories of change and of action

From the earliest point in its history, the rulers of the empire were keen to identify the best talent to serve as its administration (Wood, 2020). However, the earliest examinations, during the Han dynasty (2006 BCE–221 CE), were limited in that they were open only to those recommended by aristocratic families. All that changed during the short-lived Sui dynasty (589–618 CE) when the authorities introduced the earliest standardised tests which were open to all male candidates in an effort to "assert their authority in the face of the old aristocratic families" (Miyazaki, 1976, 9). In asserting this authority, the empire attempted to identify the most talented men to act as its civil service. From the beginning and at regular intervals throughout its history, governments reiterated the underlying principle of absolute fairness (Liu, 2018, 38), stressing the fact that the system was open to all qualifying (male) citizens. So, essentially, the theory of change

Table 2 Test content

Exam	Ming dynasty	Qing dynasty
One	Answer three questions from the *Four Books*[a] and four questions from the *Five Classics*[b]	Write three essays about the *Four Books* and a five-character poem
Two	One essay and five comments on specific issues	Write an essay related to each of the *Five Classics*
Three	Five questions related to historical and current affairs	Five questions related to comments on specific issues derived from literary classics

[a] The Great Learning, the Doctrine of the Mean, the Confucian Analects, and the Works of Mencius

[b] The Book of Songs, The Book of History, the Book of Changes, the Book of Rites, and the Spring and Autumn Annals

appears to have been to strengthen the empire by the identification and selection of the highest quality, politically agnostic (in terms of any familial power base) officials. This would be achieved (the theory of action) through the use of a competitive and demonstrably fair competitive examination system.

Development

Many see the concept of standardisation as a relatively modern concept which began with the introduction of competitive examinations, first for the military, then the civil service, and later for broader education in Europe and the USA. Large-scale application of standardisation to assessment systems first emerged in the USA in the early decades of the twentieth century (Spolsky, 1995). In fact, the earliest example of a large-scale standardised test battery from this period was the Courtis standardised tests which were launched in 1914 and had sold over fifteen million copies within little over a decade (Richardson & Johanningmeier, 2008, 235). However, as we will see in this section, standardisation was a major feature of the Kējǔ from its inception and was seen in test content and format, test administration, and test security.

Test content: input and output

The content of the examination was strictly controlled by the government from the start, while the content input was relatively transparent, in that the underlying curriculum tended to change little, so was at all times available to candidates and their teachers. There could be no test-day surprises. However, test content did change over time. For example, Table 2 shows how the three examinations that made up the provincial examination (2nd-level examination) changed from the Ming to the Qing dynasties. The changes outlined here relate less to the type of questions asked than to the input upon which the questions were asked. Clearly, candidates were expected to be extremely familiar with the source texts. For a more complete overview of the contents of the provincial examination in early Qing, see Lui (Lui & Y-C., 1974).

Although the Kējǔ was subject to quite a lot of changes over the years in terms of the type of questions included (e.g. poetry, rhyme-prose, essay), the examinations consistently included a policy focus (Elman, 1989, 1991, 2000). Policy-focused essay questions were intended to elicit extended responses related to opinions about government policies, either based on current or historical events (Jin, 1990; Liu & Li, 2004; Lu, 2005; Zhang, 1993). The topics covered a wide range of study, including classical studies, local

Table 3 Eight-legged essay structure

Leg	Title	English	Focus	Approx. length
1	破**题** (poti)	Break open the topic	To interpret the assigned topic	2 sentences
2	承**题** (chengti)	Carry forward the topic	To further elaborate the topic	3 to 4 sentences
3	**起**讲 (qijiang)	Kick-start the discussion	To shape the topic using specific 'sacred words' from the classics to reflect the words of the sages	10 sentences
4	**入题** (ruti)	Lead the topic	To introduce the main idea of the essay	2 to 4 sentences
5	**起**股 (qigu)	Initial leg	To show the general stance and the main purpose of writing (use of antithetical parallelism — AP)	4 to 9 sentences
6	**中**股 (zhonggu)	Middle leg	To fully develop the argument: this is the heart of the essay (AP)	4–10+ sentences
7	后股 (hougu)	Later leg	To review and conclude the discussion (AP)	10–20+ sentences
8	束股 (shugu)	Concluding leg	To summarise the argument and reiterate the main topic (AP). Author's own opinion allowed here	2–4 sentences

Antithetical parallelism is where two points are made, with the second contrasting or negating the first

governance, economy, national defence, history, philosophy, and world politics (Cheng, 2010; Elman, 2000).

The following is an example of a policy-focused essay question from the metropolitan exam of 1571 from Zhang (1993):

王者**与**民信守者**法**耳, 古今宜**有一定之法**。**而**孟轲, 荀卿皆**大**儒也一谓**法**先**王**一谓**法**后**王**, 何**相**左欤

(This can be translated as follows: the emperors and people all believe that we should observe established laws and practices; there must be established laws to which we can refer based on thousands of years of practice. Mencius and Master Xun are both Confucian gurus, yet they observe the laws and practices differently. What is your take?)

Note that Mencius and Master Xun were well-known and respected theorists who held very different views on this matter. Mencius argued that governments and their advisors should follow the teachings of ancient emperors such as Yao and Shun dating back over 3000 years, while Master Xun believed that they should follow the teachings of more recent emperors. In this case, the candidate was asked to write an essay to explain their opinion on this policy. This kind of policy question remains in use in Gaokao (高考 — the national university entrance examinations) and the civil service examinations in China. For example, understanding and analysing current events and government policies are part of the Gaokao syllabi for social science tests (e.g. Ministry of Education, 2018).

Test output also shifted over time from poetry to prose to essays. However, a particularly rigid format was expected of the essay response. This became known as the *baguwen* (八股文 — the eight-legged essay). The format was initially devised during the Song dynasty, and by the Ming dynasty, it had become the standard required response format. The response was expected to include the eight formatted elements or legs outlined in Table 3, which is based on the descriptions contained in Elman (2000), Liu & Li (2004), and Zhang (1993).

In addition to these structural requirements, the eight-legged essays also set many rhetorical constraints, such as the tasks of providing pairs of complementary

propositions, balanced pairs of characters and clauses, and imitation of the perspectives proffered by the likes of Mencius and Master Xun, as seen in the example above. The argument behind this strict structure was that it helped candidates avoid wandering, unfocused narratives, while also facilitating the examiners' evaluation and ranking of essays, since they focused primarily on how well candidates followed the prescribed genre (Elman, 2000; Zhang, 1993). However, the use of this response format which appears quite inflexible to the modern reader remains contentious. Some critics note that it led to a focus on appearance and structure (e.g. well-presented essays would nowadays be recognised as works of calligraphic art). The obligation to conform to appearance and structure, and the expectation that test-takers should reflect accepted political thought, had the unintended consequence of limiting candidates to meaningless uncritical responses. Thus, the eight-legged essay has been long criticised as a tool that constrained candidates' thinking (Zhang, 1993). However, more recently, scholars have argued that the eight-legged essay was unfairly criticised, asserting that the standardisation of the response actually helped candidates by replacing the pressures of conceiving a structure with a set of clearly defined guidelines (e.g. Qi, 2009; Zhang et al., 2015). Elman (2000) suggests that further academic research is needed to explore the legacy of the eight-legged essay.

Test administration procedures

Several elements of the Kējǔ procedures which emerged historically continue to influence testing administration procedures today. Specifically, these procedures determine restricted registration, fixed test dates, and centralised test venues.

Restricted registration

Unlike the earlier system in which prospective candidates were recommended by aristocratic families, any qualified individual could, in theory, apply for a place in the Kējǔ (Liu & Li, 2004). While there was no age limit for candidacy, there were many constraints on candidates' eligibility. Registration was restricted to males from a family of good standing and engaged in a respectable profession. While these respectable professions were not actually listed, it was made clear that candidates should not be descended from people associated with such *inferior occupations* as prostitution, acting, jailors, sedan bearers, barbers, or those running for public office. It was also required that candidates should not have been reported for any misconduct or malpractice. Neither were candidates allowed to register to take the examinations if their parents had recently died or if they displayed any physical disability. This was because according to Confucian philosophy, it is an important virtue that a person be seen to observe a mourning period of 27 months after the death of a parent as an expression of filial piety. Having satisfied all of the above requirements, candidates were only allowed to register in their area or region of residence (Liu, 2011; Liu & Li, 2004). While most of these constraints are no longer applicable to today's test candidates, it is interesting to note that China's current Gaokao continues to adopt this concept of registering in their area or region of residence (Zheng, 2010).

Fixed test dates

The various levels of the Kējǔ were administered regularly on fixed dates. For example, provincial examinations were administered every 3rd year on August 9, 12, and 15[1]; metropolitan examinations were administered the following year on March 9, 12, and 15; and palace examinations were usually held on April 21 right after the metropolitan examinations (Miyazaki, 1976; Zhang, 1993). At the provincial level, different provinces held the tests on the same dates, which made it impossible for candidates to take the tests twice (Zhang, 1993). In today's China, many large-scale tests follow a similar practice. For example, the annual Gaokao usually takes place between June 7 and 9, while the College English tests (CET) often fall on the third Saturday of June and December. In addition, both the Gaokao and the CET are held at the same time across the country.

Centralised test venues

The design of test sites across the country was generally standardised, though there were differences in scale. For example, the provincial test centre in Nanjing held over 20,000 candidates, while others were a third of that size. Irrespective of size, all enclosures were entered through a three-gate portal. Various activities such as purification and candidate identification were associated with the three buildings connected to the gates. From the three-gate portal, the candidate would make his way to his allocated cell (a space of about 1 m by 2 m) where he would stay for the duration of each of the three examinations.

Each test centre also contained a main tower building from which the chief invigilators oversaw the examinations and managed the routine invigilator patrols. During the Ming and Qing dynasties, this building was known as the *Hall of Supreme Fairness* with reference to the system's guiding principle (Liu, 2018, 39). In addition to the main test centre, more rooms housed the administration, transcription, and examination management functions.

Test delivery

The discovery and rapid availability of printing were integral to the entire examination system as it ensured standardisation of test item questioning and administration. For example, there were standard identification forms which captured details of the physical characteristics of the candidates as well as details of their lineage and good standing — critical to gaining approval to enter for the examinations in the first place. The introduction of timed papers was another aspect of standardisation, which have direct influence on current international language testing practices globally. The three imperial examinations described above would have taken place over 9 days, with candidates working from sunrise to sunset. Of the eight nights, six were spent in the small (approx. 1 m by 2 m) cell allocated to the candidate.

Test security: measures to prevent fraud, leak, and corruption

In order to ensure that the test experience for all candidates was standardised and to tackle the issue of cheating, a rigid system of invigilation was set in place. Below are

[1] All test dates are based on Chinese lunar calendar.

some of the many procedures used by the authorities to ensure a fair and consistent examination experience for all candidates.

Pretest

The standardised identification papers described above allowed officials to record a range of physical characteristics for each candidate, while the character references allowing for registration were also standardised.

In addition to candidate-focused procedures, the officials responsible for creating the test itself were kept isolated in a secure compound throughout the process right up until the announcement of test results (Zhang, 1993). This measure was taken to prevent item writers and test developers being contacted and bribed by others to share the information — a practice which continues in China for major examinations such as the Gaokao.

Test-day entry

On arrival at the test centre, each candidate was expected to produce his identification sheet and to present proof of his good character through affidavits included on the form as described above. Any perceived deviation from the details contained on the identification sheet was dealt with by immediate expulsion from the examinations. The repercussions for expulsion could be severe, not just for the candidate but also for his descendants given the rules around the good character requirements referred to above which extended to the candidate's family.

Candidates were instructed to only bring items that were required to complete the examination tasks, including ink, ink stones, and brushes as well as basic living needs: clothing, a chamber pot, food, a light source (typically an oil lamp), and a bamboo wife (a sleeping mat made of loosely woven bamboo). All of these items would have been subjected to a thorough search on entry to the test centre. In addition to searching the goods brought into the centre, candidates themselves were thoroughly searched, and it was not unusual for candidates to be asked to strip down for a full body search (Wells Williams, 1876). Despite this, candidates are known to have attempted to smuggle extracts from crucial texts or preprepared responses into the test centre through food and writing implements.

The allocation of specific cells for each candidate was typically based on prior test performance and could not be changed (Zhang, 1993). This was intended to limit the possibility of collusion between candidates since the chances of being placed next to a friend or colleague were slim. It also allowed examiners to identify cheating post-test during the scoring process since test-takers presenting similar responses could be identified by their location.

Test-day delivery

The candidate cells consisted of a three-sided space with an open front towards which the candidate faced while working on his responses. Therefore, passing invigilators could easily see into the cell and observe all the goods the individual candidate had brought into the centre as well as observing the candidate himself. While there were some reports of collusion between candidates and invigilators (typically to allow for

communication with other candidates), this practice was itself rigorously monitored. Punishment to candidates and invigilators was severe; for example, one could suffer banishment from the Kējǔ system and spelling disaster for the perpetrators and their descendants.

Invigilation

The main invigilator tower was the most recognisable feature of the entire security system. From the tower, the chief examiner and his team were allowed a clear overview of the entire centre — though it should be stressed that a centre such as the size of that in Nanjing (it accommodated over 20,000 candidates at one time) would have been far too large even given its multi-storied tower. For this reason, the use of regular ground patrols was essential to maintaining test security. In addition to the measures employed to ensure that candidates complied with the examination regulations, invigilators and scorers were also monitored to ensure that they were doing their job.

Subverting the system: approaches and punishment

Despite the best efforts of the authorities, and the seriousness of the punishments meted out to those found guilty of cheating, over the centuries, candidates and their tutors demonstrated incredible ingenuity and creativity in finding ways to subvert the system. Similarly, the authorities constantly monitored these efforts and put in place many measures designed to minimise any risk to the integrity of the system.

Over the centuries, a number of candidates are known to have worn undergarments upon which entire texts were written, using very small script (typically using rat hair as a writing implement) (see Tung (1959)). There are excellent examples of these garments to be found in Kējǔ museums across China. These garments were difficult and expensive to produce, suggesting that only relatively wealthy families were in a position to prepare and use them. We can also assume that these were not created to be used once and probably enabled a number of generations of a family or village to succeed at various levels of the system.

Johnston (1900, p.17) refers to a number of other means of cheating the system including "tunnels dug beneath the examination halls, through which surreptitious knowledge is passed up to the candidates" and "offices where needy and brilliant essayists are hired to personate dull, wealthy scholars". Johnson suggests that bribes increased with the level of the examinations.

In addition to the above efforts, Miyazaki (1976) described a range of measures devised by candidates and their tutors to undermine the scoring system and the corresponding precautions taken by the examination authorities to combat these. Some examples include the following:

- Candidates taught to use the characters from their names in the first section of their response. This was done to allow an examiner to identify his own (or a close friend or colleague's) student in order to ensure they were awarded a high/passing grade. This was countered by banning the use of characters used in names in the essay — particularly in the early stages of the response. This was possible as the range of characters allowed for naming was strictly limited.

- Candidates were taught to write using a particular orthographic style, again to allow for easy identification. To counter this, all original scripts were written in black ink. These were then transcribed onsite using red ink by a team of scribes trained to use a single prescribed style. The transcribed scripts were then presented to examiners for scoring. While this transcription may have prevented the targeted cheating method, it may also have introduced human error, though this point has not been made in the literature and does not appear to have been recognised as an issue at the time.

Measurement

Success at the various levels of the examination system was strictly controlled. In his excellent overview of the final 500 years of the Kējǔ, Elman (1991) highlighted this with an example of the passing rate from 1850, where just 1.5% of the candidates sitting for apprentice (also referred to as county) examinations succeeded. Of these, only 5% passed the provincial examination (2nd level of the Kējǔ), and of those, only 20% passed the metropolitan examination (3rd level — see Fig. 1). This meant that for Wang et al., (2021), 18) estimated 2.5 million Tonsheng (candidates accepted for the apprentice/county exam) in the late Qing dynasty, an average of about 240 was likely to reach the Jinshi level.

According to Chu (2015, p.169), "[N]umerical assessments were not introduced in China until the late nineteenth century and were never adopted in the Kējǔ". Instead, the examiner was expected to mark the text with a series of symbols that represented the level of quality (e.g. O represented the highest quality, while X represented the lowest). Associate examiners would first review papers and propose them to senior and chief examiners. A pass was awarded only when both of these agreed. In addition, the chief examiner would read over those papers not recommended by the assistant examiners to ensure fairness. So, a multiple marking system was in place, though there was no rating scale. The few candidates who succeeded were entered into the records and generally went on to attain a prestigious rank within the establishment, while for the rest, there was nothing. Chu (2015, 171) discusses the practice of returning examination papers to unsuccessful candidates to allow them to "understand how their papers had been graded". The fact that examiners were expected to indicate their evaluation of the work through a series of agreed markings on each answer script meant that candidates could see that the paper had been thoroughly read and appropriately judged. Examiners were judged on the level of effort they were assumed to have put into the scoring of scripts based on the number of markups they had made on the scripts they examined.

Communication

Major national examinations globally tend to be so firmly established in the educational and social systems that formal examination communication is essentially limited to confirmation of test dates, rare announcements of changes to the system, and the systematic publication of test results. In many ways, the same can be said of the Kējǔ. Test regulations rarely changed (except to deal with a perceived or an actual security risk), while changes to the content or focus of the tests typically remained unchanged for generations.

In terms of reporting and announcement of test results, Kējǔ results at all levels were publicly proclaimed. The test results were usually posted on a screen wall facing the gate of the examination compound, showing the names of candidates who passed the examinations in a ranking order based on their test performance. At the same time, a messenger was sent to the homes of those candidates who had passed to deliver the message in person. The newly presented scholars, upon announcement of the palace examination results, would join in a ritual parade led by the top scorer, known as Zhuangyuan (状元) to the Confucian temple, as well as attending other social events. These candidates also gained the right to wear status-specific clothing and to display outside their homes large carved and painted plaques declaring their new status. The plaques marked a public recognition of the distinction attached not only to the individual candidate but also to his family and even his village. Often, villages would gather to support a local school or academy in order to help aspiring local candidates to achieve as high an examination result as possible — thus bringing pride and, in time, financial and political gain to the village.

The rewards afforded to successful candidates in terms of prestige, honour, and respect acted to communicate to the masses the importance of the examination system and heightened its impact on society. However, the level of public recognition associated with success also served to highlight a sense of failure in the majority who had not managed to pass the examinations. This typically led to years of additional study (and often penury) and even to suicide to alleviate the shame such failure would bring to the candidate's family and village or town.

Consequences

Historical impact on Chinese education and society

The development of schools in dynastic China was closely connected with the Kējǔ. Between the Sui (581–618) and Yuan (1279–1368) dynasties, taking Kējǔ and attending government-run schools were the two paths to becoming a government official (Hu, 1984; Zhang et al., 2015). During this period, the curriculum for the Kējǔ and for the government-run schools was similar. Both focused on the *Four Book* (四书) and the *Five Classics* (五经). The school system therefore became an elaborate test preparation academy, a situation which continues to flourish in contemporary China (Cheng & Qi, 2006; Ma & Cheng, 2016).

In the late Tang (618–906) dynasty, new components were introduced to the Kējǔ curriculum. However, the curriculum in government-run schools remained unchanged, thus putting students in these schools at a significant disadvantage. This resulted in a decline in the popularity of government schools while signalling the significant growth of private schools where the focus was purely on training students for the Kējǔ (Zhang et al., 2015).

However, this situation changed during the Ming (1368–1644) and Qing (1644–1912) dynasties, when it was decided that all Kējǔ candidates were required to have graduated from government-run schools (Zhang, 1993). This had the effect of turning these schools into test preparation academies (Zhang et al., 2015) where students studied the Confucian classics and had to write monthly and quarterly examinations, the content and format of which were simulations of the Kējǔ examinations.

By the late nineteenth century, the dynastic system was under severe pressure on many fronts, from political to economic to societal. The government recognised the need for change and allowed new schools to open in the late nineteenth century. These schools focused on subjects such as mathematics, science, and engineering and turned their back on the traditional curriculum (Zhang et al., 2015). However, the limiting of reform to education when the empire faced such a range of serious issues was to prove costly.

Local positive impact

Within the context of the empire, the Kējǔ was undoubtedly seen by the many dynasties as a major success for almost all of its time in existence. The examination system was widely recognised within the empire as allowing for a male child of even the lowliest family the possibility (however remote) of attaining great distinction and reward through advancement within the civil service of the imperial system. The system brought communities together to form schools and academies to further the learning of their sons, who, when they attained positions of power and prestige, were expected to remember their origins and offer significant financial and political support to their place of birth. In fact, such reciprocity was not uncommon. Du Halde[2] (op. cit.: Du Halde, 1739, p. 4) discusses how this worked in practice:

> *In order to examine if the Children improve the following Method is practis'd in many places: Twenty or Thirty Families, who are all of the fame Name, and of confequence have one common Hall of their Anceftors, agree to fend their Children together twice a Month into this Hall to compofe: Every Head of a Family, by turns, gives the Thefis, and provides at his own Expcnce the Dinner for that Day, and takes care it be brought into the Hall; Iikewife it is he who judges of the Compofitions, and who determines which has compofed the beft, and if any one of this little Society is abfent on the Day of compofing, without a fufficient Caufe, his Parents are obliged to pay about Twenty-pence, which is a fure means to prevent their being abfent.*

Perhaps the most significant evidence of the positive impact of the Kējǔ is the fact that it lasted for so long! Given the extremely limited lifecycle of current examination systems (the oldest of our modern educational testing systems has only been in existence for about 130 years), the fact that the Kējǔ served the empire so well for ten times longer is remarkable and a testament to its power. For much of this time, the system proved robust in providing the empire with the calibre of civil servants required to successfully manage its needs. However, by the late nineteenth century, this was no longer the case, and the test came under increasing scrutiny and criticism.

Local negative impact

While the Kējǔ clearly had a significant and positive impact for much of its lifecycle, by the mid-to-late nineteenth century, progress made in other major countries (particularly the UK, the USA, and Japan) contrasted unfavourably with China's industrial and technological stagnation. Public disaffection with the direction taken by the Qing dynasty in

[2] This quotation comes from the original Du Halde text which was published at a time when the letter "s" within a word was represented by the letter "f".

maintaining a clearly outdated system saw the rapid disintegration of the whole system in the space of less than half a century. Opponents of the system saw industrialisation as the key to future prosperity, arguing that the contents of the examinations were outdated. Slowly at first, key individuals were sent overseas for training and education in the newly emerging sciences and technologies. By the early years of the twentieth century, it became clear that the examination system and the empire that spawned it were doomed. In 1905, the final imperial examinations were held, and just 6 years later, a series of revolutions saw the end of the empire.[3] Nevertheless, the connection between schools and high-stake tests remains true in today's China.

Ongoing impact on Chinese education and society

That the Kējǔ continues to influence China's education system is reflected in the role of examinations within the system but is also visible in the continued focus on an essentially Confucian approach to learning. We focus below on three areas in which this influence can be most clearly observed.

First, before Kējǔ, selection and education were two separate practices. Many contemporary Chinese scholars believe that with the advent of the Kējǔ, these two practices overlapped, thus promoting the development of an examination-focused educational system (Cheng, 2008, 2010; Sun, 2000). The belief in an education system that integrates selection and education is still evident in contemporary China, with schools, both public and private, training students to improve test performance in high-stake tests (e.g. Cheng, 2005, 2008; Qi, 2005; Xie & Andrews, 2013; Yu et al., 2017).

A second aspect of the impact of the imperial examination system can be seen in current Chinese testing procedures related to such diverse aspects as selection/passing quotas and secrecy in the development and scoring system. For example, access to tertiary education for candidates taking the national matriculation examinations varies by geographic area (provinces) and by ethnic group, which is clearly a continuation of the policy described above in which the Kējǔ success rates were systematically adjusted to reflect the makeup of Chinese society at that time. Additionally, to ensure test security, test designers for each subject area of the national matriculation examinations are locked into a location unknown to the public for several weeks every year until after the actual examination in that subject area has been held, a procedure which started with the Kējǔ.

Finally, the long history of the imperial examinations continue to influence Chinese society in terms of people's trust in the value and fairness of examinations. For example, the tradition originating with the imperial examinations of using examinations for selection purposes is still evident in the current education system in China. A student starts to take examinations as early as the age of 4 with the entrance test for kindergarten. Throughout their education, from kindergarten through to university, students take numerous examinations at the school, municipal, provincial, and national levels (see Cheng, 2008).

[3] See Suen and Yu's (2006) excellent discussion of the negative consequences of the system, which saw an overemphasis on rote learning of exemplar responses and a focus on test-taking skills, both seen by Suen and Yu as representing examples of construct irrelevant variance.

Broader historical and ongoing impact beyond China

As early as the late sixteenth century, the Chinese imperial examination system started to receive attention and consideration in Western publications on China. It is believed that the Jesuits brought the Kējǔ examination system to their more than 600 schools across Europe at around that time (Madaus, 1990). An indication of the growing recognition of the value of competitive examinations can be found in Burton's philosophical treatise *The Anatomy of Melancholy*. Burton (1631, 630) argues for a utopian world with "rectors of benefices to be chosen out of the Universities, examined and approved, as the literati in China".

Within a century, a number of leading members of the French enlightenment, for example Voltaire, Montesquieu, Diderot, and Rousseau, became convinced of the value of the Chinese imperial examination system of selection on merit. François Quesnay, a leading French economist of the period, also advocated for the introduction of the system to Europe. France introduced a civil service examination system in 1791 soon after the revolution that established the republic in 1789. While the system was abandoned due to corruption after a short time, its value was recognised, and a fully functioning alternative was finally in place by 1875.

In the middle of the nineteenth century, the East India Company followed the Chinese in implementing a competitive examination-based selection system. In 1855, based on the selection system of East India Company, the British government introduced a competitive examination system within its civil service. The experience in Britain influenced other European countries such as France and Germany and had a significant impact on civil service selection systems globally (Higgins & Sun, 2002; Higgins & Zheng, 2002; Teng, 1943). This influence was reflected in the USA, where Martin (1870, 64) declared that "[T]he mandarins of China are almost without exception the choicest specimens of the educated classes". Martin went on to argue for the introduction of a similar approach in the USA, declaring that "it is humiliating to reflect that our 'mandarins' are so far from being the most intellectual class of the community" (Op. Cit., 77). His position held sway, and in 1883, Congress enacted a law stipulating that government officials should be selected through a public examination.

Miyazaki (1976, 124) argued that the "case for Chinese influence upon the development of civil service examinations in Europe is strong". Interestingly enough, at the time when Miyazaki was writing, the academic study of the Kējǔ was out of favour in China, where it was viewed as a relic of the imperial past. However, in more recent times, it has come to be seen as one of the key examples of the Chinese cultural impact on the world (Cheng, 2010; Liu, 2011; Liu & Li, 2004).

Conclusions

In this paper, we set out to give the reader a broad overview of some of the most influential features of the Chinese imperial examination system (the Kējǔ), from a *contemporary validity* perspective. To do this, we looked to the socio-cognitive model of test development and validation, as updated by Chalhoub-Deville & O'Sullivan (2020). This approach allowed us to explore two key aspects of the Kējǔ: its structure and consequences. While we were not engaged in building a validation argument in support of the

system, the approach we adopted here allows the reader to interpret the details of the system presented here in a systematic way.

That the Kējǔ was both innovative and consequential is beyond doubt. In terms of the test itself, this was evidenced by details relating to test production (e.g. using printed question and response sheets while basing the contents on specified Confucian texts); delivery (e.g. test centres and security); scoring (e.g. having a monitored approach while reacting to attempts to undermine test security); and reporting (e.g. public display of results and recognition for successful candidates). When considering the consequences, we reflected on the rationale behind the introduction of the system through its theories of change and action while focusing on the historical and contemporary impact of the system both within China and globally. In fact, it can be argued that the essential infrastructure of test development, delivery, scoring, and reporting used globally today can still be traced back to the Kējǔ. In addition, the global recognition of the democratisation of learning and mobility afforded by a socially accepted and culturally appropriate examination system eventually saw the spread of the competitive examination around the world. Writing about the impact of the Kējǔ in cultural relations, O'Sullivan & Patel (2019) remind us that this examination is now recognised by many in China (including the government) as being the fifth of China's significant contributions to human civilisation along with papermaking, the compass, gunpowder, and printing. Indeed, Cressey (1929, 252) argues that the "system was one of the most important factors in the preservation of the ancient Chinese culture".

Just like all examination systems, the Kējǔ was far from perfect. Despite the stated aim to identify only the most qualified individuals, the exclusion of women and the disabled clearly meant that the talent pool was significantly limited. In addition, the significant costs associated with preparing for the different level examinations were likely to be prohibitive for the majority of the population. However, Ho (1962) estimated that in the Qing dynasty up to 45% of Juren and 38% of Jinshi came from families from outside the elites.

While there was some support available to candidates, this was generally limited to one's family or local community. Official, centrally run school education reached its peak during the Tang dynasty (618–907). However, shortly after this period, school-based education was slowly abandoned, and by the Qing dynasty (1644-1911), it existed in name only. Instead, education was fully focused on the Kējǔ in what must be the world's most egregious example of negative washback. The Kējǔ remained focused on the maintenance of Empire into the late nineteenth century when it had become clear to many in China that a modern education system was urgently required to encourage industrial growth and economic and political stability. The system's context of use and the underlying construct had become outdated, and change was needed.

Lessons learnt

We see the flowing as representing the main lessons we can learn from the Kējǔ:

- The underlying construct being tested cannot be unchanged when the context of use is unlikely to remain the same over time. Changes in society; teaching, learning, and assessment (TLA); and to educational, industrial, and political needs can

all impact on the context of use, so a test's underlying construct must be continuously revisited in a systematic way over time.

- It is not enough to just focus on major systemic changes such as construct definition and operationalisation. High-stakes testing will always attract efforts to subvert the system, and as tests become more and more digitised, this issue is not going away. Like the Kējǔ authorities, test developers and users should proactively review their security systems on a regular basis in order to eliminate as much as possible the potential for cheating.
- The separation of testing from the education system can be contentious. While the two main approaches (full separation and partial/no separation) can be seen to share a single underlying construct, the way it is manifested can be radically different, compared, for example, to the systems in the USA (full separation) and Europe (partial or no separation). The existence of washback (extreme in the case of the Kējǔ) suggests that total separation or dependence can be problematic, particularly where institutions are forced to aim for higher test scores as opposed to improved education. Where there is an effort made to integrate all elements of the system, negative washback can be limited or even avoided; see O'Sullivan (O'Sullivan & Patel, 2019).

Although the Kējǔ served the Empire well for many centuries, in the end, its negative impact was highlighted by its inability to develop the type of person needed to build a modern China. The examination system was finally abandoned in 1905, and the empire itself fell just 6 years later following repeated revolutions. However, change is often slow and is not always welcomed. Speaking to an audience in Shanghai shortly after the Kējǔ was abandoned, Ferguson (1906, 82) expressed the fear that paying "more attention to the new than to the old learning in disobedience of the regulations … would result in the production of men who had no moral standards and would therefore be useless to their country". Given the negativity around the Kējǔ, particularly in its later years, it is likely that many Chinese in his audience would have argued that this was already the case.

In this paper, we have identified how the Kējǔ has taught us much about test development, measurement, communication, and consequence. It is particularly relevant today as it highlights the dangers associated with examination systems that focus on outdated constructs which fail to deliver meaningful outcomes for society.

Abbreviations

CET	College English test
Kējǔ	Chinese imperial examinations
TLA	Teaching, learning, and assessment

Acknowledgements

Professor Eddie Williams and Dr. Maura O'Regan for reading earlier drafts of the manuscript, though all inaccuracies or inconsistencies remain ours.

Authors' contributions

O'S 60% and C 40%. LC undertook Chinese language research and also drafted the early background section. BO'S contributed much of the English language literature review, particularly the historical works. BO'S also drafted the main body of the work as regards the socio-cognitive approach and its interpretation as used here. Both BO'S and LC contributed to the discussion and conclusions. The final draft of the paper was composed by BO'S. The authors read and approved the final manuscript.

Funding

The paper was supported financially by the British Council's Assessment Research Group and further supported by the authors' institutions in terms of time.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 29 June 2022 Accepted: 24 October 2022
Published online: 17 November 2022

References

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical Development and Integrated Arguments*. Equinox.
Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.
Burton, R. (1631). *The Anatomy of Melancholy*. In *Project Gutenberg e-book* Available at: http://www.gutenberg.org/etext/10800.
Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge University Press.
Cheng, L. (2008). The key to success: English language testing in China. *Language Testing*, *25*(1), 15–38.
Cheng, L. (2010). The history of examinations: Why, how, what and whom to select? In L. Cheng, & A. Curtis (Eds.), *English language assessment and the Chinese learner*, (pp. 13–26). Routledge.
Cheng, L., & Qi, L. (2006). Description and examination of the National Matriculation English Test in China. *Language Assessment Quarterly: An International Journal*, *3*(1), 53–70.
Chu, S. (2015). Failure stories: Interpretations of rejected papers in the late imperial civil service examinations. *T'oung Pao*, *101*(1-3), 168–207.
Cressey, P. F. (1929). The influence of the literacy examination system on the development of Chinese civilization. *The American Journal of Sociology*, *35*(2), 250–262.
Du Halde, J. B. (1739). *The General History of China, Volume the Third*. John Watts.
Elman, B. A. (1989). Imperial politics and Confucian societies in late imperial China: The Hanlin and Donglin Academics. *Modern China*, *15*(4), 379–418.
Elman, B. A. (1991). Political, social, and cultural reproduction via civil service examinations in late imperial China. *The Journal of Asian Studies*, *50*(1), 7–28.
Elman, B. A. (2000). *A cultural history of civil examinations in Late Imperial China*. University of California Press.
Fairbank, J. K. (1968). *The Chinese World Order: Traditional China's Foreign Relations*. Harvard University Press.
Ferguson, J. C. (1906). The abolition of the competitive examinations in China. *Journal of the American Oriental Society*, *27*, 79–87.
Hawkey, R. A. and Milanovic, M. (2013). Cambridge English Exams – The first hundred years: A history of English Language Assessment from the University of Cambridge, 1913-2013. Studies in Language Testing series 38, Cambridge: UCLES/Cambridge University Press.
Higgins, L., & Sun, C. H. (2002). The development of psychological testing in China. *International Journal of Psychology*, *37*(4), 246–254.
Higgins, L. T., & Zheng, M. (2002). An introduction to Chinese psychology — Its historical roots until the present day. *The Journal of Psychology*, *136*(2), 225–239.
Ho, P.-t. (1962). *The Ladder of Success in Imperial China: Aspects of Social Mobility, 1368–1911*. Columbia University Press.
Hu, C. T. (1984). The historical background: Examinations and controls in pre-modern China. *Comparative Education*, *20*(1), 7–26.
Jin, F. Z. (1990). A comprehensive discussion of China's ancient civil service system. *Social sciences in China*, *XI*(2), 35–59.
Johnston, C. (1900). The struggle for reform in China. *The North American Review*, *171*(524), 13–25.
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527.
Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement*, (4th ed., pp. 17–64). Westport: American Council on Education and Praeger.
Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing*, *29*, 3–17.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.
Ko, K. H. (2017). A brief history of imperial examination and its influences. *Culture and Society*, *54*(3), 272–278.
Lai, C. T. (1970). *A scholar in imperial China*. Kelly & Walsh.
Liu, H. (2018). *The examination culture on imperial China*, (translated by Yu Weihua). Paths International.
Liu, H., & Li, B. (2004). 中国科举史 *[History of Keju in China]*. Orient Publishing Center.
Liu, X. (2011). 高考户籍**制的历**史镜像、**现实**困境**与反**思[History, contemporary dilemmas and reflections on Gaokao Hukou system]. *Journal of National Academy of Education Administration*, *11*, 57–61.
Lu, L. Z. (2005). The policy evolution and characteristics of the candidate qualifications of the imperial examinations. *Higher Education*, *12*, 100–109.
Lui, A., & Y-C. (1974). Syllabus of the provincial examination (hsiang-shih) under the early Ch'ing (1644-1795). *Modern Asian Studies*, *8*(3), 391–396.
Ma, J., & Cheng, L. (2016). Chinese student' perceptions of the value of test preparation courses for the TOEFL iBT: Merit, worth and significance. *TESL Canada Journal*, *33*(1), 58–79 http://www.teslcanadajournal.ca/index.php/tesl/article/view/1227.

Madaus, G. F. (1990). *Testing as a social technology. The inaugural annual Boise lecture on education and public policy*. Boston: Boston College.
Martin, W. A. P. (1870). Competitive examinations in China. *The North American Review*, *111*(228), 62–77.
Messick, S. (1979). Potential uses of noncognitive measurement in education. *Journal of Educational Psychology*, 71(3), 281–292. https://doi.org/10.1037/0022-0663.71.3.281.
Ministry of Educationm (2018). 2018高考大纲 [2018 Gaokao test syllabus (political studies)]. Retrieved from http://gaokao.eol.cn/gkdg/zz/201712/t20171215_1573586_2.shtml
Miyazaki, I. (1976). *China's examination hell: The civil service examinations of imperial China.* (C. Schirokauer, Trans.). Yale University Press.
O'Sullivan, B. (2000). *Towards a Model of Performance in Oral Language Tests*. UK: PhD Thesis, The University of Reading.
O'Sullivan, B. (2011). Theories and Practices in Language Testing. In Philip Powell-Davies (Ed.), *New Directions: Assessment and Evaluation - A collection of papers*. (pp.15–24). London/East Asia: British Council. Available at: https://www.teachingenglish.org.uk/sites/teacheng/files/download-accessenglish-publications-ebe-proceedings-2012.pdf.
O'Sullivan, B. (2016). Validity: What is it and who is it for? In Yiu-nam Leung (ed.). *Epoch Making in English Teaching and Learning: Evolution, Innovation, and Revolution*. Taipei: Crane Publishing Company Ltd.
O'Sullivan, B., & Patel, M. (2019). English language assessment as cultural relations. In J. P. Singh (Ed.), *The British Council Cultural Relations Collection. British Council* Accessed from: https://www.britishcouncil.org/sites/default/files/english-language-assessment-cultural-relations.pdf.
Qi, G. (2009). 说八股 [On eight legged essays]. In H. Liu (Ed.), 二十世纪科举研究论文选编 *[Selected papers on Keju studies in the twentieth century]*, (pp. 422–451). Wuhan University Press.
Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*(2), 142–173. https://doi.org/10.1191/0265532205lt300oa.
Richardson, T. R., & Johanningmeier, E. V. (2008). *Educational Research, The National Agenda, and Educational Reform: A History*. Charlotte: Information Age Publishing.
Spolsky, B. (1995). *Measured Words: The development of objective language testing*. Oxford: Oxford University Press.
Suen, H. K., & Yu, L. (2006). Chronic consequences of high-stakes testing? Lessons from the Chinese civil service exam. *Comparative Education Review*, *50*(1), 46–65.
Sun, P. Q. (Ed.) (2000). *The history of Education in China. [*中国教育史*]*. East China Normal University Press.
Teng, S. (1943). Chinese influence on the Western examination system: I. *Introduction. Harvard Journal of Asiatic Studies*, *7*(4), 268–312.
Tung, S.-K. (1959). A Chinese cribbing garment. *The Princeton University Library Chronicle*, *20*(4), 175–181.
Wade, G. (2005). *Southeast Asia in the Ming Shi-lu: An open access resource*. Singapore: Asia Research Institute and the Singapore E-Press, National University of Singapore http://epress.nus.edu.sg/msl/reign/hong-xi/year-1-month-7-day-12, Accessed 22 Jan 2019.
Wang, M. B., van Leeuwen, B., & Li, J. (2021). *Education in China, ca. 1840-present*. Brill.
Weir, C. J., Vidakovic, I. and Galaczi, E. D. (2013). Measured constructs: a history of Cambridge English Examinations, 1913-2012. *Studies in Language Testing series 37*. Cambridge: UCLES/Cambridge University Press.
Wells Williams, S. (1876). China: The country and people. *Journal of the American Geographical Society of New York*, *8*, 269–284.
Wood, M. (2020). *The story of China: A portrait of a civilisation and its people*. Simon & Schuster.
Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with structural equation modeling. *Language Testing*, *30*(1), 49–70. https://doi.org/10.1177/0265532212442634.
Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., … Fang, L. (2017). *Preparing for the speaking tasks of the TOEFL iBT® test: An investigation of the journeys of chinese test takers (ETS RR-17-19)*. *ETS Research Report Series*: Princeton Retrieved from http://doi.wiley.com/10.1002/ets2.1214.
Zhang, X. (1993). 中国科举考试制度 *[The imperial examination system in China]*. Xinhua Publishing House.
Zhang, X., Mao, P., & Li, S. (Eds.) (2015). 中国科举制度通史 *[A comprehensive history of Keju in China]*. Shanghai People's Publishing House.
Zheng, R. (2010). The National College Entrance Examination reform: Concerns and practice. *Peking University Education Review*, *8*(2), 14–29.

Publisher's Note