Rzetelność i trafność egzaminów Cambridge English

Gillian Horton-Krueger

Egzaminy i testy językowe tworzone przez Cambridge English mogą być różnie wykorzystywane: „lokalnie"przez nauczycieli, np. dla zbadania postępów uczniów, w celach diagnostycznych (umożliwiających np. planowanie lekcji), lub też na szeroką skalę, jako egzaminy o charakterze doniosłym, których wyniki mogą być uwzględniane w rekrutacji na uczelnie czy przez pracodawców. Zespół Cambridge English Language Assessment, zajmujący się konstruowaniem tego typu egzaminów, ma na celu tworzenie testów odpowiadających powszechnemu zapotrzebowaniu i stanowiących trafne i rzetelne narzędzie ewaluacyjne. Tekst w języku angielskim jest poprzedzony krótką syntezą w języku polskim, opracowaną przez dr Agnieszkę Dryjańską, redaktor JOwS.

Tworzenie przez Cambridge English wysokiej jakości egzaminów, które w wiarygodny sposób sprawdzają umiejętności językowe kandydatów, nie jest związane ze stosowaniem jakiejś „tajemniczej receptury", ale stanowi skomplikowany i wieloetapowy proces. Wymaga on starannych badań, testowania próbnego (wstępnego) oraz analizy rezultatów. Niniejszy artykuł prezentuje wybrane procedury tego procesu ze szczególnym uwzględnieniem etapów opracowywania zadań egzaminacyjnych i systemu zapewniającego wysoką jakość oceniania kompetencji produkcyjnych – mówienia i pisania.

Etap opracowywania zadań egzaminacyjnych

Zanim egzamin trafi do rąk kandydata, musi przebyć dwuletnią drogę przygotowania, począwszy od etapu zamówienia wstępnej wersji zadań egzaminacyjnych, aż do etapu redagowania ich ostatecznej wersji. Ten proces ma na celu przede wszystkim zapewnienie porównywalności różnych wersji określonych komponentów tego samego testu. W tym celu wybierane są osoby odpowiedzialne za każdy z komponentów testu (ang. *assessment manager*), współpracujące z ekspertami zewnętrznymi (ang. *chair*), kontrolujące pracę autorów i konsultantów poszczególnych jednostek testu, a także zajmujące się etapem ponownej weryfikacji opracowanych zadań.

Opracowanie jednostek egzaminacyjnych odbywa się z wykorzystaniem poradnika (*Item-Writer Guidelines*) zawierającego wskazania dla autorów: informacje o zasadach konstruowania zadań testowych, przykłady jednostek testowych oraz wybór tematyki tekstów. Ten poradnik zawiera także wskazówki

Egzaminy Cambridge English osiągają wysoki poziom rzetelności i trafności dzięki wieloetapowej procedurze tworzenia zadań testowych. Niezwykle ważne jest zaangażowanie ekspertów mających nie tylko wiedzę teoretyczną (...), ale także bogate doświadczenie w nauczaniu języka angielskiego.

usprawniające poszukiwanie tekstów źródłowych oraz niezbędne informacje z *Europejskiego systemu opisu kształcenia językowego* (ESOKJ) odnośnie do konkretnego poziomu w zakresie wymagań gramatycznych czy leksykalnych.

Zadania, które zostaną przyjęte, po przejściu procesu edytorskiego są poddawane kluczowej procedurze – testowaniu próbnemu na reprezentatywnej dla danego egzaminu populacji uczniów. Ten etap dostarcza informacji zwrotnych dotyczących poziomu trudności pytań oraz tego, w jaki sposób badane zadania pozwalają rozróżnić zdających o wyższym i niższym poziomie. Te dane statystyczne, w połączeniu z oceną ekspertów, są brane pod uwagę przy dalszym opracowywaniu testu. Materiały, które finalnie odpowiadają wszystkim wymaganiom (odpowiednio opisane, z uwzględnieniem typu zadania, tematyki i długości tekstów czy wieku egzaminowanych) trafiają do „banku jednostek testowych" – bazy danych, z której bezpośrednio korzystają twórcy konkretnych egzaminów.

Zapewnienie wysokiej jakości oceniania

Zadania umożliwiające obiektywną ocenę to te niewymagające oceny eksperta: można je oceniać albo automatycznie, albo z wykorzystaniem klucza zawierającego wszystkie dopuszczone odpowiedzi. Ten typ oceniania jest wykorzystywany w zadaniach sprawdzających kompetencje receptywne: czytanie, słuchanie oraz test gramatyczno-leksykalny (ang. *Use of English*).

Najbardziej złożonym zagadnieniem w zakresie oceniania jest zapewnienie rzetelnej oceny kompetencji produkcyjnych. Głównie z myślą o tym właśnie rodzaju oceniania opracowano specjalny system zapewniający jakość oceniania – *Quality assurance for marking* (*QA*).

Skale ocen wykorzystywane w ocenianiu kompetencji produkcyjnych: pisania i mówienia, są tworzone w odniesieniu do zaleceń ESOKJ przez ekspertów mających bogate doświadczenie w nauczaniu języka angielskiego. Egzaminatorzy, którzy zajmują się ocenianiem kompetencji produkcyjnych kandydatów, są poddawani wieloetapowemu procesowi szkoleń i ewaluacji umiejętności egzaminacyjnych. Jest to szczególnie trudne w przypadku oceniania wypowiedzi ustnych, gdyż wymaga wyszkolenia wielkiej liczby egzaminatorów na całym świecie (20 tys.). W celu zapewnienia odpowiedniej jakości szkolenie tych egzaminatorów ma charakter kaskadowy – lider regionalny nadzoruje pewną liczbę egzaminatorów, z których każdy nadzoruje kolejną grupę dobraną już w mniejszych okręgach itd.

W przypadku wypowiedzi pisemnych liczba oceniających jest mniejsza, gdyż mniej jest ośrodków centralnych, w których oceniane są przesłane prace, stąd proces szkolenia i kontroli jest prostszy.

Szkolenie egzaminatorów polega na dostarczeniu im odpowiednich materiałów o charakterze zarówno merytorycznym, jak i organizacyjnym. W przypadku wypowiedzi ustnych i pisemnych udostępniane są dwa zestawy materiałów audiowizualnych: pierwszy szkoleniowy, drugi sprawdzający. Certyfikat dopuszczający do egzaminowania otrzymują tylko te osoby, których oceny prezentowanych wypowiedzi mieszczą się w zakresie ustalonym przez ekspertów.

Podsumowując, egzaminy Cambridge English osiągają wysoki poziom rzetelności i trafności dzięki czasochłonnej i wieloetapowej procedurze tworzenia zadań testowych. Niezwykle ważne jest zaangażowanie w ten proces ekspertów mających nie tylko wiedzę teoretyczną na temat konstruowania testów, ale także bogate doświadczenie w nauczaniu języka angielskiego. Ostatnim omawianym elementem istotnie wpływającym na zapewnienie jakości procesu egzaminowania proponowanego przez Cambridge Assessment jest staranne szkolenie egzaminatorów, szczególnie tych, którzy są odpowiedzialni za ocenianie wypowiedzi pisemnych i ustnych.

***

Ensuring validity and reliability in large-scale examinations.
Aspects of the Cambridge English approach

Tests are produced to meet particular needs and need to be fit-for-purpose in the contexts they serve. They may be set locally, for example by a teacher to measure progress and provide diagnostic information for lesson planning. Or they may be part of large-scale assessments, which will be used for high-stakes decision-making such as selection for further study, employment or migration. An exam board like Cambridge English Language Assessment, producing large-scale assessments, will strive to develop tests which stand up to public scrutiny and fulfil their purpose as a broad-based measurement tool in a valid and reliable manner, over time.

Delivering a language exam which is a thorough and fair test of the language skills and abilities of the candidates is a complicated and multi-faceted process, requiring careful research, trial and analysis, with ongoing commitment through all phases of implementation. The quality of Cambridge English exams is less a "secret recipe" than a painstaking series of processes which reflect a deep engagement with the research field and a rigorous approach to practical detail. This article selectively describes some of the practical procedures which are followed to ensure that a Cambridge English exam tests what it is supposed to test and provides reliable results. These descriptions focus on the question paper production process and the quality assurance systems for marking the productive skills, Writing and Speaking.

The question paper production (QPP) process

When students sit a Cambridge English exam, the papers presented to them will have taken up to two years to produce, from the commissioning of draft materials to the production of the final question papers (whether paper – or computer-based). The QPP process works to ensure comparability across test versions, one of the essential features of reliable certification. For each of the test components (e.g. Cambridge English: First Listening), QPP is managed by an Assessment Manager working with an external expert, called a Chair, to lead a team of item writers and consultants within a yearly planning and review cycle.

Item writing draws on a set of Item-Writer Guidelines, which are test specification documents produced for each exam and used by the item writers who are commissioned to produce tasks. (The slimmer Handbooks published for teachers are also test specification documents, in this case intended to support teaching and learning, rather than generate test items.) The Guidelines take a practical approach and include information about the test construct and task requirements, as well as example items and guidelines about topic choice. There is also information, based on past experience, which can help item writers in the writing process; for example advice on what to look for – and avoid – when searching for a source text. The Guidelines contain information relating to the relevant Common European Framework of Reference (CEFR) level, including appropriate grammar, vocabulary and functions or the amount of support or scaffolding to be included. For example, tests at A1 and A2 tend to offer more visuals to provide support and there are empirically sourced wordlists at these levels to guide learners' lexical development. At all levels, minimum and maximum word lengths are specified for any input reading for each task.

Questions that do not meet the quality criteria from the outset are rejected or rewritten. Questions that are accepted are taken through a thorough editing process. A key component

The quality of Cambridge English exams is less a "secret recipe" than a painstaking series of processes which reflect a deep engagement with the research field and a rigorous approach to practical detail.

is pretesting, where material is tested on student populations who are as similar as possible to the future candidate populations. This provides performance data for each task, including how difficult the sample of candidates found the questions and how well the questions discriminated between stronger and weaker candidates. These statistics, along with the expert judgement of a pretest-review panel, enable further adaptations to be made.

Materials which finally meet all requirements go into an item bank, the database for the management of test content, ready for test construction. Along with the item-writing process, the item bank supports test comparability and thus plays a pivotal role in the test construction process. The quality of an item bank depends not only on the number of items it contains, but also on the quantity and quality of the data it holds *about* the items: more information stored about an item allows for more automated selection processes. Item features such as task type, topic, word count, testing focus and target age-group are logged. Other details, such as accent for Listening tasks, can also be recorded. Statistical information, produced after a test administration from candidate responses, is uploaded to the relevant item. This data includes the difficulty of the item, the facility, which is the proportion of candidates who answered the item correctly, and the discrimination index, which indicates how well the item discriminates between the strong and weak candidates. Items are also classified according to their calibration status. An uncalibrated item has no item statistics; an item that is part-calibrated has been pretested, so with statistics indicating how the item is expected to perform in a live test; fully calibrated items have been included in a live test session and taken by a sufficiently high number of candidates with different first language backgrounds. The information stored in the item bank is then utilized for test construction, when a test is compiled for live use.

Quality assurance for marking

Objectively scored items are those which do not require expert judgement for marking, in other words which can be reliably marked using automated processes or where a key containing all possible answers can be supplied to a human marker. This type of marking is used for the task types currently found in the Reading, Listening and Use of English components of Cambridge English examinations.

For the Writing and (face-to-face) Speaking components, assessment scales tied to the Common European Framework of Reference are used, applied by expert examiners, in this case experienced language teaching professionals. It is self-evident that wherever marks are given by human raters using open scales, a comprehensive quality assurance (QA) system needs to be in place to ensure a standardized application. For both Writing and Speaking in Cambridge English examinations, examiners must fulfil Minimum Professional Requirements in order to be considered. They commit to a process of induction, training and ongoing (annual) certification, as well as monitoring during live marking sessions and regular statistical reliability checks afterwards. Examiners receive feedback about their performance and may only continue to mark if the checks indicate they are on track.

To support these QA processes, Cambridge English Language Assessment employs a highly structured Team Leader System which works on a cascade principle. Speaking Examiners are grouped under Team Leaders. Team Leaders are grouped under Regional Team Leaders, who in turn are overseen by Professional Support Leaders. This system is critically important, as Speaking examiners are recruited locally by exam centres. In other words, quality and consistency of marking have to be maintained across a worldwide network of 20,000+ Speaking examiners. The system is similar for Writing, but less elaborate in terms of the Team Leader System, due to the smaller number of examiners involved, in fewer places.

For the examiners' yearly certification requirements, a series of exemplar videos for Speaking, and a series of scripts for

Delivering a language exam which is a thorough and fair test of the language skills and abilities of the candidates is a complicated and multi-faceted process, requiring careful research, trial and analysis.

Writing, are marked by groups of senior examiners. The submissions are analyzed and the outcome is a set of robust, standardized marks. All examiners are given a selection of the exemplar scripts/videos, first a set to analyse, then sets to mark. The marks they award are compared to the standardized marks, and examiners must meet a specified level of accuracy before they are certified to conduct live sessions. All examiners up to and including Professional Support Leaders, must undergo the standardization process including marks collection, much of which today takes place online.

In the case of the Speaking tests, alongside reliability with respect to marking, the aspect of procedure is an additional requirement of the QA system. Here again, the Team Leader System around the world ensures that expertise cascades through the system. Team Leaders guide new examiners through a practical training process, familiarizing them with the procedures as well as the assessment of Cambridge English Speaking tests. This includes examiner roles, test security, test format, materials handling, and the function of the standardized interlocutor scripts ("frames"). Training is followed by certification, which from then on, as for assessment, is an annual standardization process. The monitoring which takes place during live tests every two years, either face-to-face or through audio recordings, covers procedure as well as marking. Thus candidates can rely not only on standardized marking but also on standardized administration of the Speaking tests.

Further reading

This article has outlined processes routinely implemented to produce each Cambridge English test paper version, as well as some of the quality assurance processes governing the use of examiners. If you would like to read more about these and other aspects involved in the production and processing of a Cambridge English examination, including grading and score reporting, use of the Cambridge English Scale, and malpractice detection, you will find longer and more detailed articles in Research Notes 59, February 2015. Research Notes is a quarterly published by Cambridge English Language Assessment, reporting on learning, teaching and assessment. Issue 59 centres on the life cycle of Cambridge English language tests, explaining which analyses and processes are used to ensure delivery of accurate and meaningful results. It is downloadable free of charge from http://www.cambridgeenglish.org/research-notes/.

The following five volumes in of the Studies in Language Testing (SiLT) series set out the theories of communicative language ability underpinning Cambridge English examinations and how they feed into test development and design (all published by Cambridge University Press):

- Geranpayeh, A. and Taylor, L. (eds) (2013*) Examining Listening: Research and Practice in Assessing Second Language Listening*. SiLT, vol. 35.
- Khalifa, H. and Weir, C. J. (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*. SiLT, vol. 29.
- Shaw, S. D. and Weir, C. J. (2007) *Examining Writing: Research and Practice in Assessing Second Language Writing*. SiLT, vol. 26.
- Taylor, L. (ed.) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. SiLT, vol. 30.
- Weir, C.J., Vidakovic, I., Galaczi, E.D. (2013) *Measured Constructs. A history of Cambridge English Language Examinations* 1913-2012. SiLT, vol. 37.

Gillian Horton-Krueger

Head of Assessment Services Northern Europe for Cambridge English Language Assessment (Berlin). Before joining the staff of Cambridge English in 2014, Gillian had over twenty years of experience in English language training, assessment and consultancy in tertiary and continuing education as well as in-service teacher development. She was Professional Support Leader for Cambridge English in Germany from 2009 to 2014.