AUNQUE LAS EVALUACIONES NO ENGORDAN A LA CABRA...

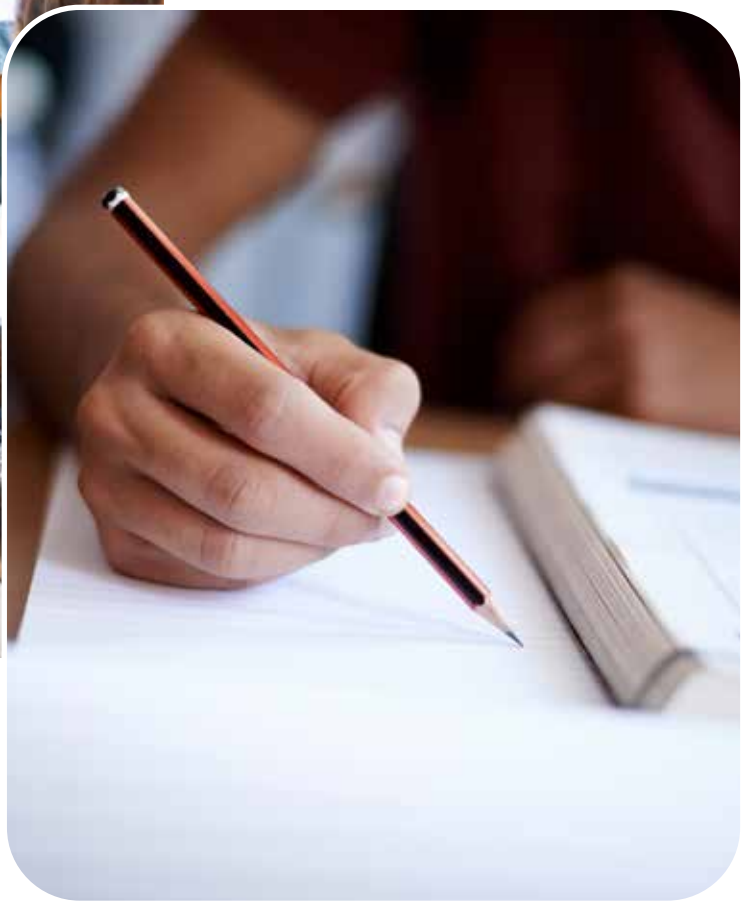
ALTHOUGH TESTING DOESN'T FATTEN THE GOAT...

Russell Whitehead

Director de LT123, compañía que creó en 2011 para ofrecer servicios de evaluación y materiales de evaluación para los departamentos de evaluación, editoriales y universidades mundialmente.

Felicity O'Dell

Consultora en LT123. Después de dedicarse a la enseñanza por muchos años, ahora dedica su tiempo a la escritura. Sus intereses principales son la enseñanza y aprendizaje y la evaluación.



En este artículo analizaremos primero algunas actitudes negativas encontradas en las pruebas, antes de mencionar las razones positivas para evaluar. Nos referimos a las características de las "buenas" pruebas e indicamos que la de entrada/salida de Richmond las evidencia.

Actitudes hacia las pruebas

Vivimos en una era de datos y medición. Es difícil no generar datos en nuestra vida diaria. Todo lo que hacemos en Internet es grabado, rastreado, interpretado. Nuestras preferencias, nuestras compras, nuestras conexiones —todo se nota constantemente y se pone en uso. Tenemos la oportunidad de limitar o ajustar algo de esto, pero no mucho. Es utilizado por diferentes organizaciones para diferentes propósitos en diferentes momentos. Se negocia entre organizaciones.

Se ha observado que las personas que hacen fila en la oficina de correos en las grandes ciudades siempre están muy impacientes y ansiosas por ser atendidos, mientras que la gente en pequeños pueblos es paciente. Quienes viven en la ciudad conocen y resienten el hecho de que siempre están afanados, con poco tiempo, debido a que siempre están viajando etc., dejándolos exasperados por la espera adicional. En las áreas rurales, por su parte, disfrutan de la oportunidad de pasar un poco de tiempo charlando con los vecinos y disfrutar del ritmo lento.

No es irrazonable conjeturar que, la gente sabe que se les está midiendo constantemente en tantas formas, en su mayoría impulsadas económicamente, en secreto, su sensación de resentimiento hierve cuando se encuentran con cualquier evaluación o medición. Hemos escuchado maestros lamentarse sobre las pruebas y no siempre parece totalmente razonable hacerlo.

Se ha aseverado que las pruebas reducen el interés de los alumnos en el aprendizaje, que los maestros están obligados a centrarse en la preparación tediosa de exámenes en lugar de en actividades más creativas y, lo más grave, que los resultados de las pruebas no dan una imagen justa y precisa de la capacidad del candidato que las toma. Si las pruebas son deficientes en ciertos aspectos, entonces la primera y la tercera aseveración pueden ser correctas. Si las pruebas son deficientes y además los maestros no están bien entrenados, entonces la segunda aseveración puede ser correcta también.

La evaluación es un negocio complejo y muchos factores están involucrados en ella. Afecta a muchas personas de muchas maneras y se ve afectada por muchos elementos. La política de evaluación de LT123 intenta enumerar los componentes más importantes para tener en cuenta al considerar las pruebas, ya sea al diseñarlas o al usarlas. Puede verlo aquí. https://lt123.co.uk/testing-manifesto/

Razones para evaluar

¿Se subiría a un avión si creyera que las habilidades de vuelo y liderazgo del piloto no se habían probado con rigor?

¿Se imagina a un actor que no pudiera decir sus líneas antes de que empezara la obra?

¿Entraría a un maratón sin haberse preparado, entrenado y cronometrado una variedad de carreras y pruebas?

Por supuesto, es cierto que no se puede engordar a una cabra pesándola repetidamente. Pesamos a la cabra para ver si nuestras supuestas mejoras en su dieta están teniendo el efecto deseado. (Así como probamos nuevas vacunas, métodos de control de enfermedades, etc.)

De la misma manera procedemos con las pruebas. Debemos observar muchos factores. Una prueba hace observaciones sobre el candidato, pero a su vez el candidato genera información sobre la prueba.

> Testing is a complex business and many factors are involved in it.

In this article, we first discuss some negative attitudes encountered to testing, before mentioning some positive reasons to test. We refer to some characteristics of 'good' tests, and indicate that the Richmond Entry/Exit test exhibits these.

Attitudes to testing

We live in an age of data and measurement. It is difficult not to generate data in our daily lives. Everything we do on the internet is recorded, tracked, interpreted. Our preferences, our purchases, our connections – all is constantly noted and put to some use. We do have the opportunity to limit or adjust some of this, but not much of it. It is used by different organisations for different purposes at different times. It is traded between organisations.

It has been observed that people queuing in the post office in big cities are always very impatient and anxious to get on, while people in villages stand patiently. Urbanites know and resent the fact that they are always behind, always time-poor, because of the time spent commuting and so on, leaving them exasperated by additional waiting. Country-dwellers, meanwhile, enjoy the chance to spend a bit of time chatting to the neighbours and relish the slow pace.

It's not unreasonable to conjecture that since people know they are constantly being measured in so many, mostly economically-driven, ways behind the scenes, their sense of resentment boils over when they encounter any assessment or measuring. We have heard teachers really moaning about testing and it doesn't always seem wholly reasonable to do so.

Charges are levelled that testing damages learners' interest in learning, that teachers are obliged to focus on tedious exam preparation rather than more creative activities, and, most seriously in a way, that the test results do not give a fair and accurate picture of the test-taker's ability.

If tests are poor in certain respects, then the first and third charges may be right. If the tests are poor and the teachers not well trained, then the second charge may be right.

Testing is a complex business and many factors are involved in it. It affects many people in many ways and it is affected by many things. The LT123 Testing Manifesto tries to list the most important things to bear in mind when considering tests, either when designing them or using them. You can see it here. https://lt123.co.uk/testing-manifesto/

Reasons to test

Would you want to get on a plane if you didn't believe the pilot's flying and leadership abilities hadn't been very rigorously tested?

Can you imagine an actor not seeing if they could say their lines before the play started?

Pero las principales razones para probar antes, durante y después del proceso de aprendizaje son apoyar, orientar y mejorar el aprendizaje de los estudiantes.

Esto es importante porque los estudiantes, en particular en edad escolar, necesitan estar protegidos. Son los usuarios de la educación y la evaluación, pero no son el cliente, son los clientes los que suelen tener el poder, no los usuarios. Una prueba proporcionada fuera del aula, en efecto, sin depender del profesor, del curso, los materiales del curso, permite al usuario saber si se está progresando, si se están alcanzando metas. La evaluación facilita las relaciones entre los alumnos y las sociedades en las que viven. Supervisa el éxito o no de la educación. Indica dónde pueden ser necesarias las correcciones y qué se ha hecho bien. Proporciona una imagen precisa de lo que alguien puede hacer y genera credibilidad en una amplia gama de personas y organizaciones.

Una prueba nos permite saber no solo si un alumno recuerda algo, sino que realmente lo ha aprendido, al ver si puede aplicar el aprendizaje en escenarios que no son simplemente repeticiones del escenario original en el que se encontró.

A medida que toda la actividad humana se mueve más y más en línea, es obvio que pensemos en mirar lo que se puede hacer allí. Si hay amplio uso del lenguaje en general y gran parte del estudio se lleva a cabo en línea, entonces es allí donde deberíamos estar evaluando.

Por lo tanto, las pruebas nos ayudan a ver lo que realmente

> La evaluación facilita las relaciones entre los alumnos y las sociedades en las que viven.

está sucediendo, ayudan a orientar y definir el aprendizaje. Hay una serie de propósitos dentro de ese objetivo general. Con los cursos en línea, un papel importante para las pruebas es ubicar a los alumnos en el nivel correcto de curso o acceso a materiales, para que no afecten su aprendizaje al involucrarse con cursos inapropiados.

Características de las pruebas 'buenas'

Es crucial que una prueba comience éticamente desde un buen punto. Las pruebas no deben utilizarse injustamente como dispositivos para excluir a los migrantes, por ejemplo, estableciendo tareas deliberadamente demasiado difíciles.

Suponiendo, entonces, que estamos probando con esencialmente buenas intenciones, necesitamos establecer algún tipo de asociación entre la prueba y la toma de pruebas. Esto significa que la prueba debe ser transparente. El candidato debe saber lo que el examen está destinado a evaluar y debe entender lo que se le pide que haga en el examen. Se necesita documentación e instrucciones claras, con ejemplos proporcionados si es necesario. Debe quedar claro para el candidato cómo se otorgan las puntuaciones, para que sepa en qué preguntas, tal vez, dedicar más tiempo, y así sucesivamente.

Would you enter for a marathon without timing yourself on a range of preparation and training runs and workouts?

It is of course true that you cannot fatten a goat by weighing it repeatedly. We weigh the goat to see if our supposed improvements to its diet are having the desired effect. (Just as we test new vaccines, methods of disease control, and so on.)

We must proceed with care and we must observe many things. A test makes observations about the test-taker, but in turn the test-taker generates information about the test. But the main reasons to test before, during and after the learning process are to support, shape and enhance the learning and the learners.

This matters because learners, particularly school-age learners, need to be safe-guarded. They are the users of education and assessment, but they are not the customer, and it is customers who usually hold the power, not the users. A test provided from outside the classroom, provided in effect independently of the teacher, of the course, the course materials, enables the user to know if progress is being made, if goals are being attained. Assessment facilitates the relationships between learners and the societies in which they live. It monitors the success or otherwise of education.

It indicates where repairs may be necessary and what has gone well. It provides an accurate picture of what somebody can do and it can be believed by a wide range of people and organisations.

A test enables us to know if a learner has not just remembered something but actually learnt it, by seeing if the leaner can apply the learning in scenarios that are not simply repetitions of the original scenario in which it was encountered.

As more and more of all human activity moves online, of course we would think of seeing what can be done online. And if much language use in general, and much studying, takes place online, then that's where we should be testing.

Testing therefore helps us to see what is really happening and it helps to shape and define learning. There's a range of purposes within that overall objective. With online courses one important role for testing is to place learners in the right level of course or materials access, so that they don't impair their learning by engaging with inappropriate courseware.

Characteristics of 'good' tests

It's crucial that a test starts from a good place ethically. Tests should not be used as devices unfairly to exclude migrants, for example, by setting tasks deliberately too hard.

Assuming, then, that we are testing with essentially good intentions, we need to establish some form of partnership between test and test-taker. This means that the test should be transparent. The test-taker should know what the test is intended to do and they should understand what they are requested to do by the test. This means clear documentation and clear instructions, with examples provided if necessary. It should be clear to the test-taker how the scores are awarded, so that they know which questions, perhaps, to spend longer on, and so forth.

Ideally, the test-taker should find the test a learning experience in itself, though that may not be fully possible in every situation. As a minimum, the test should be pleasant to engage with. This applies both to the content and to the user experience. The test should be in line with the test-takers and their experience. Young learners should get a test appropriate to what they know and they have studied and will go on to study. Nothing in the test's content should be offensive or alarming or upsetting – and this can be critical when designing and developing globally used tests which will travel across different cultures.

Idealmente, el candidato debe encontrar en la prueba una experiencia de aprendizaje en sí misma, aunque eso puede no ser totalmente posible en cada situación. Como mínimo, debe ser agradable de realizar. Esto se aplica tanto al contenido como a la experiencia del usuario. Debe estar alineada con los candidatos y su experiencia. Los jóvenes estudiantes deben tener pruebas apropiadas para lo que saben, lo que hayan estudiado y lo que vayan a estudiar. Nada en el contenido de la prueba debe ser ofensivo, alarmante o molesto, y esto puede ser crítico al diseñar y desarrollar pruebas usadas globalmente que viajarán a través de diferentes culturas.

En el caso de muchas pruebas, es inevitable que las personas se preparen para estas. Esto tiene mucho sentido. Probablemente querría prepararse para el examen de conducción, así que ¿por qué no para el examen de inglés? Si la gente va a pasar tiempo preparándose, entonces el diseño de la prueba y su contenido, deben fomentar actividades de aprendizaje pedagógicamente útiles como preparación.

Sin embargo, el equilibrio con esa ambición es la cuestión de la practicidad. Si queremos, y normalmente lo hacemos, pasar un tiempo relativamente corto siendo evaluados, entonces los diseños de prueba basados en la eficiencia de los informes pueden no estar siempre llenos de material realmente interactivo, divertido y dinámico.

Las pruebas deben estar en el nivel adecuado para los que toman el examen. Esto es más fácil de lograr ahora, existe la tecnología para habilitar pruebas adaptativas. La prueba es impulsada por el rendimiento del candidato, moviéndose hacia arriba y hacia abajo en el nivel de acuerdo con las respuestas dadas. Es útil si los niveles reportados por una prueba están en concordancia con los de otros. Si todas las pruebas se asignan en un marco como el MCER, entonces podemos saber que el resultado del examen X es equivalente al resultado del examen Y, incluso si esas pruebas están en diferentes idiomas y países.

Todos estos elementos cobran relevancia en el contexto de la formación en inglés y en la búsqueda de mejorar los sistemas educativos en Colombia, abriendo oportunidades para la innovación en tiempos de cambio. Richmond en alianza con LT 123, propone un ecosistema digital para promover la evaluación sistemática y lograr que a través de la analítica de aprendizaje se pueda evidenciar el progreso de los estudiantes a lo largo de su proceso de aprendizaje. A continuación, presentamos la prueba de ingreso/salida del sistema evaluativo de Richmond.

Prueba de entrada/ salida de Richmond

Para el desarrollo de la prueba de entrada/salida de Richmond, el equipo de LT123, con colegas de Richmond, trabajo con dedicación para aplicar los puntos antes mencionados. El siguiente es el resumen de nuestro informe de recomendaciones iniciales:

LT123 utilizará su experiencia para diseñar y validar sus pruebas de entrada y salida de jóvenes. Esto incluye:

- Amplio conocimiento de jóvenes estudiantes como participantes en el examen (Papp y Rixon 2018)
- Experiencia con la medición del desarrollo del idioma inglés de los estudiantes de escuela
- Prueba de procesos de desarrollo y validación que tienen en cuenta los últimos estudios en campos relevantes, entre ellos:
  - Evolución reciente del MCER para estudiantes en primaria
  - Las últimas investigaciones acerca del English Vocabulary Profile y el English Grammar Profile
  - Configuración estándar en el desarrollo y validación de las pruebas de estudiantes jóvenes.

La prueba fue diseñada siguiendo los pasos estándar de garantía de calidad. El proyecto tenía cierta complejidad y debía tener en cuenta diferentes niveles y diferentes edades.

Era importante que la prueba fuera totalmente apropiada para los jóvenes que toman el

In the case of many tests, it is inevitable that people will prepare for the test. This makes a lot of sense. You probably would want to prepare for your driving test, so why not your English test? If people are going to spend time preparing, then the test's design, its content, should encourage pedagogically useful learning activities as preparation.

Balanced with that ambition, however, is the issue of practicality. If we want, and we usually do, to spend a relatively short time being tested, then test designs based on efficiency of reporting may not always be full of really interactive, fun and dynamic material.

Tests must be at the right kind of level for the test-takers. This is easier to achieve these days now there is the technology to enable adaptive testing. The test is driven by the test-taker's performance, moving up and down in level according to the answers given. It is useful if the levels reported by one test are in concordance with those from others. If all tests are mapped against a framework such as CEFR then we can know that exam result X is equivalent to exam result Y, even if those tests are in different languages and countries.

All this becomes relevant in the context of English teaching and learning and in the pursue of improvement for Colombian educational programs, thus opening the door for innovation in times of change. Richmond in alliance with LT123, proposes a digital ecosystem to promote systematic evaluation so that through learning analytics students' progress throughout their learning process can be evidenced. Next, we present the entry and exit test that belongs to the Richmond evaluating system.

Richmond Entry/Exit test

Turning now to the Richmond Entry/exit test, we at LT123, working with colleagues in Richmond, worked hard to apply these points to this test. This summary is from our initial recommendations report:

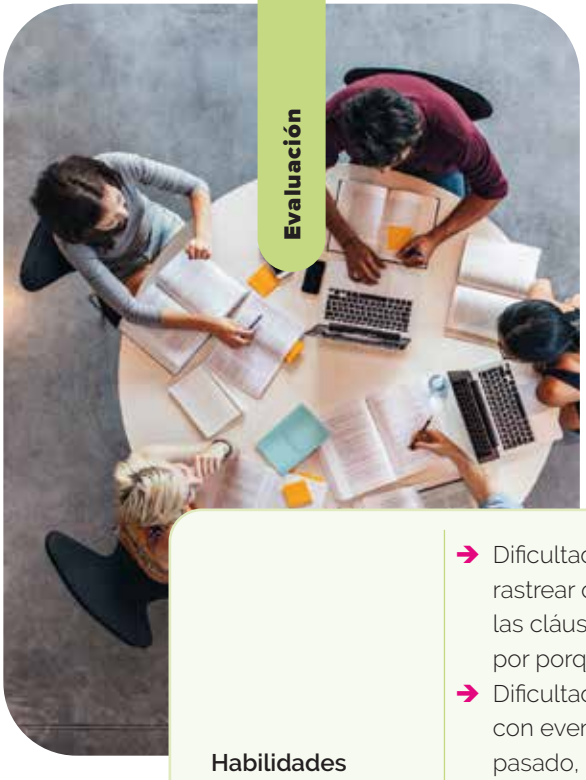
'LT123 will use its expertise to design and validate its Young Learners entrance and exit tests. This includes:

- Extensive knowledge of young learners as test takers (Papp and Rixon 2018)
- Experience with measuring school learners English language development
- Test development and validation processes that take the latest knowledge in relevant fields into account, including:
  - Recent developments in the CEFR for young learners
  - Latest research informing English Vocabulary Profile and English Grammar Profile
  - Standard setting in developing and validating young learners' tests.'

The test was designed following standard steps of quality assurance. The project had some complexity in that we needed to account for different levels and different ages.

It was important to make the test fully appropriate for the young test-takers. The consultant team at LT123 includes Dr. Szilvia Papp, a widely recognised expert in this area, as well as Dr Felicity O'Dell and Frances Treloar, both highly experienced in young learners tests within assessment organisations and as authors of preparation and course materials for young learners.

Here are just a few examples of the kinds of considerations made, taken from Szilvia's own publications, showing in each case the characteristic of this age of learner, together with the problem and the solution.



examen. El equipo de consultores de LT123 incluye a la Dra. Szilvia Papp, una experta ampliamente reconocida en esta área, así como a la Dra. Felicity O'Dell y Frances Treloar, ambas con experiencia en pruebas de estudiantes jóvenes dentro de organizaciones de evaluación y autoras de diseño de cursos y materiales de aprendizaje para estudiantes jóvenes.

Estos son solo algunos ejemplos de los tipos de consideraciones realizadas, tomadas de las publicaciones de Szilvia, que muestran, en cada caso, la característica del alumno de estas edades, junto con el problema y la solución.

Habilidades limitadas de razonamiento	➔ Dificultad como oyentes y lectores para rastrear desde los efectos a las causas en las cláusulas de procesamiento conectados por porque, hasta los 11 años ➔ Dificultad con los contrafactuales relacionados con eventos que no tienen lugar en el presente o pasado, es decir, procesar una situación contraria a la realidad fáctica, difícil porque implica negar una declaración positiva en la cláusula if, hasta los 10 años *(Estamos visitando un país muy seco). Si lloviera allí, tomaríamos un paraguas.* ➔ *(Ayer no llovió). Si hubiera llovido,* habríamos llevado un paraguas. ➔ Dificultad con la resolución de problemas y las pruebas de hipótesis	➔ Utilizar eventos con cronología básica ➔ Utilizar causa y efecto común, 'Sucedió porque no fuimos cuidadosos' ➔ No se refieren a eventos que no tienen lugar en el presente o pasado ➔ Crear relaciones claras entre los personajes
Problemas en la construcción de una representación de discurso más amplia	➔ Problemas para incorporar una información en un contexto más amplio ➔ Capacidad limitada para obtener la esencia y ver los vínculos entre las partes ➔ Dificultad para recopilar información, integrar múltiples piezas de evidencia ➔ No sensible a los conectivos que mantienen unido un texto	➔ Contextualizar la tarea en la rúbrica (¿quién, dónde, por qué?) ➔ Utilizar conectivos simples ➔ Utilizar tipos de discursos familiares y secuenciales (narración, instrucción, descripción) ➔ Utilizar conexiones transparentes entre expresiones o giros ➔ Evitar saltos engañosos de tema
Conocimiento mundial limitado	Puede tratar el medio ambiente inmediato y temas familiares	➔ Hacer coincidir el contenido/tema con el conocimiento de fondo esperado de los candidatos ➔ Necesidad de material de prueba para ser culturalmente neutral
Comprensión limitada de las convenciones del discurso escrito	Dificultad para interpretar los objetivos / intenciones del escritor	A los niños se les puede pedir que empaticen y reconozcan cómo se sienten los demás, pero es posible que no respondan como se espera.

Limited reasoning skills	➔ Difficulty as listeners and readers to trace back from effects to causes in processing clauses connected by *because* until age 11 ➔ Difficulty with counterfactuals related to events that do/did not take place in the present or past, i.E. Processing a situation that is counter to factual reality, difficult because they entail negating a positive statement in the if clause, until age 10 *(We're visiting a very dry country). If it rained there, we'd take an umbrella.* *(It didn't rain yesterday). If it had rained, we'd have taken an umbrella.* ➔ Difficulty with problem solving and hypothesis testing	➔ Use events with basic chronology ➔ Use common cause and effect, 'it happened because we were not careful' ➔ Do not refer to events that do/did not take place in the present or past ➔ Create clear relationships between characters
Problems in constructing a wider discourse representation	➔ Problems in embedding a piece of information in a wider context ➔ Limited ability to get the gist and to see the links between parts ➔ Difficulty collating information, integrating multiple pieces of evidence ➔ Not sensitive to the connectives which hold together a text	➔ Contextualize the task in the rubric (who, where, why?) ➔ Use simple connectives ➔ Use familiar, sequential discourse types (narration, instruction, description) ➔ Use transparent connections between utterances or turns ➔ Avoid misleading leaps of topic
Limited world knowledge	➔ Can deal with immediate environment and familiar topics	➔ Match content/topic to expected background knowledge of candidates ➔ Need for test material to be culturally neutral
Limited understanding of the conventions of written discourse	➔ Difficulty interpreting writer's goals / intentions	➔ Children can be asked to empathise and recognize how others feel, but they may not respond as expected

La prueba fue diseñada para tener en cuenta muchos factores como estos. También tenía que ser fiable como herramienta de medición, además de ser agradable como una experiencia para el candidato. La prueba es relativamente corta, clara y accesible para interactuar con ella.

Creamos un equipo muy bien gestionado de escritores y editores muy experimentados, así como estudios de audio e ilustradores, para crear el material inicial. Se creó un panel de expertos investigadores para establecer el nivel del material en relación con el MCER. Esto se utilizó para impulsar el diseño de las pruebas previas y el contenido fue probado en varias escuelas en América del Sur. Los resultados se analizaron estadísticamente y se realizaron ajustes cuando fue necesario. El resultado es una herramienta de evaluación diseñada para mejorar la experiencia de aprendizaje ayudando a ubicar a los alumnos en los cursos correctos y a proporcionar una medida de su progreso como resultado de los cursos.

Por supuesto, solo podemos pesar la cabra y no el campo en el que vive. Diferentes estudiantes hacen varias cosas para mejorar su inglés más allá de tomar un curso en particular, así la prueba nos habla del alumno, no de los cursos como tal, pero entonces ese es el camino correcto, ¿no?. **RM**

BIBLIOGRAFÍA Y REFERENCIAS

http://www.santillana.com.co/rutamaestra/edicion-28/referencias

The test was designed to take account of many factors such as these. It also had to be reliable as a measuring tool as well as being pleasant as an experience for the test-taker. The test is relatively short and clear and accessible to engage with.

We ran a tightly managed team of very experienced writers and editors, as well as audio studio and illustrators, to create the test material. A level-vetting panel of experts was set up to establish the level of the material in relation to CEFR. This was used to drive the design of the pre-tests and the content was trialled in a number of schools in South America. The results were analysed statistically and adjustments were made where necessary. The outcome is a testing tool designed to improve the learning experience by helping to place learners into the right courses and to provide an indication of their measurable progress as a result of those courses.

Of course, we can only weigh the goat and not the field it lives in. Different learners do various things to improve their English above and beyond taking one particular course so the test tells us about the learner, not the courses as such – but then that's the right way round, isn't it? **RM**

> Different learners do various things to improve their English above