EXÁMENES ESCRITOS
DE DESEMPEÑO
EN L2:
INCIDENCIA DE LAS DIVERGENCIAS
EN LA CORRECCIÓN

María Laura Lanzoni

Claudia López Camelo

Junio 2007

Introducción

¿Por qué en una prueba escrita de lengua segunda dos candidatos pueden obtener diferentes puntajes? La primera respuesta que surge es que el puntaje refleja las diferencias en su habilidad para producir un texto en esa lengua. Pero las investigaciones en el área de evaluación en L2 han hallado que, en las pruebas de desempeño, además de esa posibilidad, existen otras fuentes de variabilidad que explican la divergencia. Entre ellas las más importantes, según McNamara (1996), son:

1) La divergencia relacionada con la tarea encomendada (en los casos en que el candidato tiene la posibilidad de elegir entre dos o más tareas propuestas).

2) La divergencia asociada al corrector: el texto de un candidato puede recibir un puntaje diferente, por ejemplo, según la severidad o indulgencia de la persona que lo evaluó. La interacción entre las características del corrector y las cualidades de las escalas de evaluación que se utilizan tiene una influencia crucial en los puntajes otorgados, más allá de la calidad del desempeño del candidato.

McNamara menciona, además, que las variaciones que introducen estos dos factores son grandes, sobre todo la asociada con el corrector. Este es un hecho que merece un tratamiento adecuado, ya que muchas veces se toman importantes decisiones para el futuro de un candidato a partir de la información suministrada por un test. Las evaluaciones de L2 son utilizadas, por ejemplo, para determinar si una persona está en condiciones de comunicarse en esa lengua en un futuro trabajo o en un contexto académico. Por eso es necesario tomar en cuenta estos factores de variación, si un examen se propone lograr resultados estables y justos.

En este trabajo nos vamos a centrar específicamente en la variación introducida por la función del corrector en la evaluación de los textos escritos producidos en el examen CELU, que se propone evaluar "la competencia que tiene el hablante de español como lengua extranjera para comunicarse oralmente y por escrito en lengua española, de manera efectiva en situaciones cotidianas, laborales o académicas"[1]. De esta manera, el candidato CELU debe dar respuesta al planteo de actividades similares a las realizadas en la vida cotidiana. Las mismas son situaciones de uso de la lengua con un propósito social que exigen la ejecución de más de una habilidad lingüística para poder cumplirlas. Esto significa que las habilidades del candidato son evaluadas en forma integrada.

Si analizamos los resultados obtenidos en las distintas tomas del examen CELU, hallamos que hay un porcentaje de divergencia entre los correctores, como es de esperar en un examen de desempeño. Nuestro propósito es describir el proceso de corrección de las actividades escritas del

[1] *Manual del candidato CELU* (2005)

examen CELU y analizar de manera cuantitativa y cualitativa las discrepancias que surgieron entre los correctores del CELU 206 (noviembre del 2006), en el momento de asignar un nivel al candidato. Para ello tomaremos en cuenta la bibliografía especializada en el tema y una encuesta realizada para tal fin entre los correctores de la mencionada toma.

Evaluación del desempeño en L2

Hay diferentes formas de evaluar la proficiencia de un candidato en lengua extranjera. En las pruebas de desempeño como el examen CELU, el candidato tiene que demostrar el dominio de alguna habilidad. De acuerdo con McNamara (1996), en este tipo de pruebas el instrumento de evaluación elicita una performance o conducta que se denomina "respuesta construida"; esto significa que no hay una sola respuesta posible, lo que sí sucede en una prueba objetiva; en el examen de desempeño el candidato es evaluado por un corrector que usa una escala.

En consecuencia, en las pruebas objetivas el procedimiento de corrección es más "seguro" en términos de ausencia de variabilidad entre un corrector y otro, es decir, su confiabilidad es alta pero no así su validez, ya que nos dan escasa información sobre la habilidad del candidato para comunicarse en L2. Las pruebas de desempeño, en cambio, se caracterizan por la similitud de las tareas solicitadas con las que el candidato tendrá que llevar a cabo en la vida real, es decir, se intenta solicitar tareas "auténticas". "La evaluación de desempeño presupone que la mejor manera de evaluar si alguien es proficiente o no es colocarlo en una situación en la que pueda demostrar esa proficiencia directamente" (Schlatter, Scaramucci y otros, 2004).

La corrección de un examen escrito de desempeño

Según McNamara (1996), tradicionalmente se han utilizado diferentes procedimientos para conseguir una medición justa: elaboración cuidadosa de escalas de corrección (escalas métricas que en adelante llamaremos "grillas"), entrenamiento de los correctores, realización de más de una corrección de cada examen, con procedimientos preestablecidos para manejar el desacuerdo. Algunas posibilidades que suelen emplearse para solucionar los desacuerdos son: promediar el puntaje obtenido, sumar otra corrección, tratar de que los correctores discutan hasta llegar a un acuerdo.

Los autores coinciden en la importancia de confeccionar una escala de medición adecuada. De acuerdo con Scaramucci (2005) en las pruebas de desempeño, la "grada" (que en el examen CELU denominamos "grilla") "operacionaliza y preserva la concepción que fundamenta la prueba y también sus objetivos, de forma de mantener su validez de constructo y de contenido; también, entre otros aspectos de la prueba, procura garantizar su confiabilidad, de forma de minimizar la influencia de las creencias y concepciones de los correctores, que no necesariamente están en

sintonía con aquellas que fundamentan la prueba, llevando una mayor estabilidad o confiabilidad a los resultados."

Por otro lado, Alderson (1995) sostiene que la formación de los examinadores es un componente crucial en cualquier programa de evaluación, puesto que si la puntuación de una prueba no es válida y confiable, todo el trabajo llevado a cabo para obtener un instrumento de "calidad" habrá sido una pérdida de tiempo. El reto de los examinadores es el de comprender los principios que subyacen en las escalas de puntuación con las que deben trabajar y el de interpretar los descriptores de forma coherente.

Sin embargo, McNamara señala que el hecho de que haya divergencias entre los juicios de diferentes evaluadores no es algo negativo en sí mismo; es, en verdad, algo inevitable. Además, puede resultar provechoso, en cuanto puede aportar diferentes miradas del mismo hecho, que contribuyen a darnos un panorama más amplio de la habilidad manifestada por el candidato en ese texto. Pero, si el desacuerdo supera ciertos límites, genera problemas de confiabilidad para el examen. Por esa razón, en la evaluación de desempeño siempre se pone énfasis en la necesidad de mejorar la capacidad de la prueba para medir de manera justa la habilidad de los candidatos.

El uso de escalas métricas en la evaluación de exámenes de desempeño

Como plantea Bachman, las escalas métricas se basan en la idea de que la adquisición de proficiencia lingüística se desenvuelve en un continuum que varía de la no proficiencia al desempeño ideal (Bachman, 1996). En la misma línea, también Scaramucci (2000) afirma que la proficiencia no es un concepto "absoluto", de todo-o-nada, sino un concepto "relativo", que trata de tener en cuenta la especificidad de la situación de uso futuro de la lengua. Ese continuum del cual hablan estos autores es expresado en forma de una línea de puntos que representan estadios de ese desenvolvimiento, los cuales corresponden al grado de habilidades demostrado en la resolución de la tarea del test en cuestión. Esos puntos son acompañados por una descripción del comportamiento lingüístico de cada nivel y forman los descriptores de la escala.

Existen distintos tipos de escalas métricas para medir la proficiencia en L2.

Los autores consultados en la bibliografía especializada distinguen entre **escala holística** y **escala analítica**, según la forma en la que se atribuyen valores o se asigna puntaje a la tarea.

La **escala holística** consiste en la asignación de un solo puntaje a la producción escrita del candidato, sobre la base de una impresión del conjunto. Todos los investigadores señalan como ventaja el hecho de que este tipo de escala permite una realización rápida de la corrección. Esto hace posible que cada escrito pueda ser corregido más de una vez.

Bachman y Palmer (1996) señalan como una importante ventaja de este tipo de escala la concentración en los logros del candidato y no en sus deficiencias. Agregan que refleja auténticamente la reacción del lector, aunque no provee información específica acerca de las diferentes habilidades. Sin embargo, puede presentar dificultad de interpretación por parte del

evaluador. Al respecto, Weigle (2001) señala que los correctores no necesariamente usan los mismos criterios para llegar al mismo puntaje. Otro riesgo que destaca Weir es que la calidad entera del trabajo puede verse afectada por sólo uno o dos aspectos de los que deben tenerse en cuenta. La impresión global no toma en cuenta a los estudiantes cuyo nivel de performance varía en términos de los distintos criterios.

Ahora bien, la **escala analítica** requiere un puntaje separado para cada uno de los aspectos de la tarea. La mayoría de los autores consultados mencionan como ventaja que este tipo de medición permite hacer un diagnóstico preciso de las diferentes habilidades necesarias para la producción escrita, identificando virtudes y debilidades de los candidatos; otorga mayor objetividad y más seguridad a los correctores. Así, Bachman y Palmer sostienen que el solo hecho de que los correctores tengan que dar un puntaje tenderá a hacer la puntuación más confiable. Los correctores se ven obligados a considerar aspectos de performance que, en otras circunstancias, hubieran ignorado, es decir, no pasan por alto ninguna de las características ni privilegian una frente a otra. Siguiendo a Weir, el corrector debe "atender a la multidimensionalidad de la producción escrita".

Entre las desventajas de la medición analítica, la principal es el tiempo que toma aplicar esta escala. Incluso teniendo práctica, la puntuación tomará más tiempo que la holística. La segunda desventaja es que la concentración en los distintos aspectos puede distraer la atención del efecto total del escrito. Al respecto, Walkiria Sidi (2002), en un trabajo de investigación sobre el CELPE-BRAS, retoma la idea de Perkins (1983) de que el discurso escrito u hablado es un sistema formado por componentes en interacción y constituye un todo integrado, no es sólo el resultado de la suma de sus partes. Debe ser percibido y evaluado globalmente. El resultado es que un puntaje compuesto puede ser muy confiable pero no válido.

A pesar de las desventajas mencionadas, este método es visto por Weir como una herramienta útil para el entrenamiento y estandarización de examinadores relativamente inexpertos en la corrección de este tipo de exámenes.

Por otro lado, Hughes destaca que no todos los sistemas de puntuación darán resultados igualmente válidos y confiables en todas las situaciones. El sistema tiene que ser adecuado al nivel de los candidatos y el propósito del test. Si la información diagnóstica es requerida directamente de los puntajes dados, entonces la puntuación analítica es esencial. Hughes recomienda, en cualquier caso, usar más de una corrección.

El proceso de corrección en el examen CELU

El examen CELU consiste en una prueba única para todos los niveles. Consta de una instancia escrita y otra oral. En la parte escrita el candidato tiene que producir cuatro textos que responden a tareas planteadas a partir de un input que puede ser oral o escrito. Para la corrección se elabora una grilla única para cada tarea. La función de esta es proveer de una descripción

“objetiva” donde el evaluador pueda ubicar de manera confiable las muestras de examen que evalúa. En dicha grilla, los descriptores se definen según la respuesta esperada en los distintos niveles (avanzado, intermedio, básico, no alcanza). Para ello, se tienen en cuenta cuatro criterios: adecuación contextual, adecuación discursiva, morfosintaxis y léxico (ver material anexo). Según Alderson (1995), es aconsejable usar escalas de no más de siete puntos, puesto que es difícil hacer distinciones más sutiles en cuanto al logro alcanzado por los candidatos en su desempeño.

Los diferentes investigadores coinciden en afirmar que la grilla de corrección tiene que ser lo más clara y explícita posible, ya que representa la definición de la(s) habilidad(es) que la prueba intenta medir. Alderson recomienda que la grilla sea confeccionada por las mismas personas que elaboraron el examen. Además, sugiere el uso de escalas diferentes para distintas tareas: “una escala de nivelación es raramente apropiada para la evaluación de todas las actuaciones lingüísticas” (Alderson, 1995). En este sentido, hay que tener en cuenta qué aspectos de la escritura son más importantes, de acuerdo con la finalidad que tiene el examen. Por ejemplo, cuando el examen va a ser utilizado para evaluar capacidad de uso de la lengua en contextos académicos o laborales, puede darse más importancia a la efectividad comunicativa de un texto que al conocimiento de características específicas de la lengua, y esto debe estar reflejado en la grilla.

Antes de comenzar la corrección, se seleccionan, entre los textos producidos por los candidatos, aquellos que ejemplifiquen las características sobresalientes de cada nivel que mida el examen, tal como están descriptas en la grilla. También se puede tener en cuenta algunos casos problemáticos, que no pueden ser ubicados en un punto concreto de la escala. (Alderson 1995; Weigle, 2001). Normalmente esta tarea está a cargo de los correctores senior, quienes conducen la corrección.

Una vez creados los grupos de correctores entre los que habrá un especialista en la actividad que se debe corregir, se procede a leer la grilla de corrección. Luego de haber tomado conocimiento de sus contenidos, los correctores leen las muestras seleccionadas por la comisión técnica. Una vez que le adjudican un nivel a cada una, se discuten las decisiones. En caso de no haber acuerdo, se efectúan las modificaciones necesarias en la grilla. Alderson (1995) insiste en la importancia de que los examinadores hagan su propio juicio de valor antes de conocer los motivos de la comisión, para que estos no influyan en su decisión. También considera aconsejable que se deje constancia escrita de los motivos de cada una de las decisiones tomadas en este consenso y que el jefe de examinadores proceda a incorporar las modificaciones en la grilla. Los autores insisten en la necesidad de que todo el equipo llegue a un acuerdo sobre esa primera corrección antes de continuar la tarea. En esta etapa de unificación de criterios suelen surgir elementos inesperados no descriptos en la grilla que crean dificultades al momento de asignar nivel, ya que lo descripto en ella no siempre se corresponde con lo que se observa en la producción textual de los candidatos. A partir de ese momento, Alderson sostiene que no es conveniente introducir nuevas modificaciones, ya que podría tornar inaceptable cualquier variación en la puntuación.

Comienza, entonces, la corrección propiamente dicha. Cada actividad es corregida una vez en forma analítica y otra de manera holística, tareas que están a cargo de correctores diferentes. Cada corrector debe realizar su tarea por separado, volcando los resultados en una planilla especialmente elaborada, sin que el otro corrector tenga acceso a esa información. En esa planilla consignan el puntaje y nivel asignados y hacen las observaciones que consideren necesarias también por escrito. Estas observaciones resultan sumamente útiles en caso de que las puntuaciones de uno y otro corrector no coincidan. Cuando hay divergencia en el nivel asignado, se procede a una revisión de la corrección, que está a cargo de miembros del mismo equipo que está corrigiendo la actividad en cuestión. Si se hace difícil llegar a un consenso, es un miembro de la comisión quien dictamina el puntaje final.

Al finalizar el proceso de corrección, los autores recomiendan realizar un análisis del trabajo de cada corrector. A veces este análisis muestra que un corrector tiene tendencia a asignar puntajes más altos o más bajos que el resto. Este es un punto que debe ser trabajado con ese corrector. En la medida en que se pueda descubrir una tendencia, es decir que el comportamiento del corrector es consistente en su asignación de puntaje, se puede intervenir para tratar de compensarla (McNamara, 1996; Weigle, 2001).

Los correctores pueden variar entre sí, por ejemplo, en el grado de indulgencia o severidad para evaluar. Esta actitud puede ser generalizada o puede estar relacionada con un grupo especial de candidatos o con una tarea en particular. McNamara informa, por ejemplo, que se ha hallado que los correctores evalúan en forma más severa los aspectos del desempeño de un candidato vinculados con los recursos formales del lenguaje, particularmente la estructura gramatical.

También se puede presentar divergencia en la forma en que los correctores interpretan la escala de corrección que están utilizando. Normalmente en estas escalas se le asigna un puntaje discreto a cada candidato. Lo que ocurre cuando un candidato, por su habilidad, está ubicado en la intersección entre dos niveles, es que un corrector puede ubicarlo en el superior y el otro en el inferior.

Otro aspecto para tener en cuenta mencionado por Cumming (1997) es el de las diferentes expectativas que pueden tener los correctores de textos en segunda lengua, de acuerdo con su experiencia en la enseñanza de la lengua como materna o segunda.

Entrenamiento de los correctores

El entrenamiento de los correctores es uno de los puntos centrales que los diferentes autores mencionan entre los procedimientos para lograr un nivel aceptable de confiabilidad inter-corrector y, de esa manera, conseguir que los textos escritos sean corregidos de una manera justa y consistente.

Alderson recomienda que este proceso de formación se realice a intervalos regulares, no sólo cuando se administren pruebas por primera vez. Esto vale tanto para los examinadores nuevos

como para los experimentados, ya que a veces estos últimos desarrollan formas propias e individuales de examinar y son los responsables de la ausencia de confiabilidad en la corrección.

Debemos destacar, sin embargo, que los autores coinciden en afirmar que, si bien el entrenamiento de correctores es muy importante, nunca podrá reducir completamente la divergencia entre los puntajes o niveles asignados. Por otra parte, Weigle (1994) menciona que estos intentos para aumentar la confiabilidad de un proceso de corrección a través del entrenamiento y el consenso entre los correctores pueden llevar a juicios no válidos sobre la calidad de un texto porque, de alguna manera, distorsionan la interacción normal entre lector/texto que es una de las partes del proceso de lectura. Cita a Huot (1990) quien señala que "los correctores tienen inevitablemente reacciones individuales, personales frente a un texto que son intrínsecas a la lectura fluente; a través del entrenamiento esta contribución personal del corrector puede quedar reducida a una serie de principios negociados. Esto significaría sacrificar una corrección verdadera por una corrección confiable". Otros autores, en esta misma línea, indican que de esta manera estaríamos sosteniendo que hay una única lectura "correcta" de un texto.

Por otra parte, Lumley (2002) sostiene que las reglas y la escala no cubren todas las eventualidades, forzando a los correctores a desarrollar varias estrategias para ayudarse con los aspectos problemáticos del proceso de corrección. Una de ellas es mantenerse cerca de la escala, aunque esto no puede eliminar la fuerte influencia producida por la impresión intuitiva obtenida en la primera lectura del texto, los correctores parecen sentirse obligados a tomar una decisión en cuanto al puntaje en términos de lo que la escala dice, incluso cuando ellos sienten que no están conformes con ello.

Lumley agrega que el corrector y no la escala es quien descansa en el centro del proceso. Es el corrector quien decide: qué rasgos de la escala deben ser tenidos en cuenta, cómo arbitrar entre los inevitables conflictos que surgen a partir del contenido de los descriptores y cómo justificar su impresión del texto en términos de los requerimientos institucionales representados por la escala y el entrenamiento de correctores.

Los efectos del entrenamiento en los correctores

Éste es un campo en el que no hay muchas investigaciones realizadas. Este tipo de investigaciones normalmente utiliza protocolos verbales ("think-aloud protocols"); se les pide a los correctores que verbalicen lo que piensan mientras corrigen; estas verbalizaciones son grabadas para su posterior análisis y comparación.

Una investigación realizada por Huot (1988) halló que no había diferencias en los criterios de corrección que utilizaban dos grupos de correctores: uno con experiencia previa y otro sin experiencia; sin embargo, encontró que los correctores más experimentados tenían estrategias más eficientes y una gama de respuestas más amplia para analizar los ensayos que los otros correctores.

Otra investigación llevada a cabo por Cumming (1990) para comparar el comportamiento de correctores experimentados/inexpertos demostró que los correctores sin experiencia eran más indulgentes cuando juzgaban el contenido y la organización de los textos y que los experimentados hacían una mayor distinción entre las categorías para evaluar, mientras que los inexpertos se centraban más en la detección de los errores.

En general estos estudios sugieren que hay algunas diferencias generales entre correctores experimentados e inexpertos con respecto a las estrategias que usan para evaluar la escritura, pero también hay variabilidad entre los correctores individuales.

Weigle (1994) realizó un estudio para explorar los efectos del entrenamiento tanto sobre correctores experimentados como inexpertos. A partir del análisis de los resultados de las divergencias entre el puntaje asignado en la corrección previa y la corrección posterior al entrenamiento, llegó a la conclusión de que el proceso de entrenamiento que tuvo lugar entre la primera y la segunda corrección fue exitoso, al permitir a los correctores corregir de manera más consistente en la segunda ocasión. Llegó a las siguientes conclusiones:

1. En la primera corrección, los correctores basaron algunas de sus decisiones en una interpretación idiosincrásica de ciertos aspectos de la grilla. Luego, en el entrenamiento se clarificaron ciertas expresiones utilizadas en las descripciones y las estrategias para otorgar peso relativo a los diferentes descriptores de cada nivel. También surgieron durante el proceso de entrenamiento algunos criterios de corrección que no estaban explicitados en la grilla y que posteriormente fueron utilizados por ellos para la segunda corrección. Por ejemplo: qué hacer con ensayos muy cortos, con textos que no se adecuan a ninguna banda en particular dentro de la escala, etc.

2. Los correctores ajustaron sus expectativas al tipo de escritura aceptable para la tarea propuesta para la población a la que estaba dirigido el examen; en este sentido, se encontró que muchas veces los correctores eran más exigentes y estrictos en la corrección previa al entrenamiento.

3. El entrenamiento les dio a los correctores la noción de cómo cada uno evaluaba los textos en comparación con los otros miembros del equipo, aun cuando esto no les haya hecho cambiar en algunos casos el puntaje asignado a los textos en la corrección previa.

Por otro lado, este estudio no encontró lo que algunos autores han señalado como un efecto negativo del entrenamiento: una preocupación dominante por lograr un acuerdo con los otros correctores. Los comentarios de los correctores en este sentido no fueron tan abundantes; por ejemplo la comparación entre la tendencia a ser más severo o indulgente en la corrección nunca fue utilizada como una razón para compensar esa tendencia. Además, comparativamente, hay mayor cantidad de comentarios sustantivos referidos a la calidad de los textos corregidos.

La investigadora llega, entonces, a la conclusión de que el entrenamiento fue efectivo para que los correctores que habían tenido grandes divergencias entre la corrección previa y la posterior pudieran aplicar la grilla en la forma propuesta.

El caso del CELU 206

El Celu 206 recibió la inscripción de 151 candidatos, de los cuales solo 139 se presentaron a la prueba. Como ya se ha mencionado, en este examen los candidatos deben realizar 4 actividades que integran las distintas destrezas. Se analizaron, entonces, los resultados de los 139 candidatos correspondientes a cada una de las 4 actividades. Para ello, se registraron en una planilla los puntajes finales correspondientes a cada una de las 4 actividades, tanto en la evaluación holística como en la analítica. Luego se identificaron los exámenes que habían sido revisados por la comisión técnica en una o varias de sus actividades. En estos casos, se registró el nivel asignado por el corrector holístico, el analítico y el revisor y se seleccionaron, para su estudio, sólo los exámenes que habían sido revisados en 3 ó 4 actividades.

Las hipótesis de trabajo a partir de las que se analizaron los resultados fueron:

- Cuando una tarea presenta un mayor grado de complejidad ya sea en cuanto al tipo textual solicitado o a la comprensión que exige el input (oral o escrito), es posible encontrar mayor diversidad en la producción de los candidatos no ajustada a los descriptores de la grilla. Esto podría producir un mayor grado de divergencia entre los correctores.

- En los casos en que es necesaria la revisión del resultado, se podría encontrar un mayor porcentaje de coincidencia entre esta revisión y uno de los dos tipos de corrección realizados (holístico o analítico).

El siguiente cuadro muestra los resultados obtenidos a partir de nuestro análisis sobre la divergencia entre correctores:

Actividad	Exámenes revisados	Divergencia sobre el total de exámenes corregidos	Coincidencia entre revisor y corrector analítico	Coincidencia entre revisor y corrector holístico
1	49	35%	28	20
2	55	40%	31	23
3	46	33%	23	23
4	57	41%	28	29

Actividad 1: se revisaron 49 exámenes de un total de 139. En 28, la puntuación del revisor coincidió con la del corrector analítico y, en 20, con la del holístico. La divergencia entre correctores fue de un 35,25%.

Actividad 2: de los 55 exámenes revisados, en la corrección de 31 de ellos el revisor coincidió con el corrector analítico y en 23 con el holístico. El porcentaje de divergencia fue de un 39,56%.

Actividad 3: de 46 revisados, en la corrección de 23 de ellos el revisor coincidió con el corrector analítico y en 23 con el holístico. El porcentaje de divergencia fue de 33,09%.

Actividad 4: de 57 revisados, en la corrección de 28 exámenes el revisor coincidió con el corrector analítico y en 29 con el holístico. La divergencia fue del 41%.

A partir de estos datos, se puede concluir que no se encontraron diferencias significativas en cuanto a la evaluación de tareas con mayor o menor complejidad. Si bien en las actividades 1 y 2 se observa mayor coincidencia del revisor con el corrector analítico, no podemos asegurar que se trate de una tendencia, ya que no ocurre lo mismo con el resto de las tareas. Para poder confirmar nuestra segunda hipótesis sería necesario, entonces, contar con una muestra mayor.

Por otro lado, no podemos afirmar que la revisión de los resultados donde hubo divergencia muestre una mayor correspondencia con la corrección analítica que con la holística.

De los exámenes que presentan discrepancias en las 4 actividades (21), se observó que el mayor porcentaje (66% - 15 candidatos) corresponde a candidatos brasileños, es decir, hablantes de portugués. En las planillas de corrección de estos exámenes, las observaciones consignadas por los correctores y revisores en las planillas de corrección coinciden en cada caso en cuanto a los descriptores en que se focalizan (por ej.; morfosintaxis y léxico). De esta observación surge la necesidad de estudiar más a fondo el desempeño de los candidatos brasileños en este examen, tema para un futuro trabajo.

En cuanto al análisis de divergencias en 3 actividades, se pudo registrar que si bien muchas veces las observaciones de los correctores analítico y holístico coinciden, no ocurre lo mismo con el nivel adjudicado por cada uno de ellos. Esto significa que lo que para uno implica un punto de corte, para otro no lo es. Esta observación nos lleva a un tema que preocupa a los correctores en general y es la necesidad de unificar criterios.

Encuesta realizada a los correctores

Como parte del trabajo se realizó una encuesta entre los correctores que participaron de la corrección del CELU 206 (ver material anexo). Para la confección de la misma se tuvieron en cuenta los factores de divergencia señalados por la bibliografía consultada y las inquietudes que fueron surgiendo por parte de los correctores a lo largo de las distintas experiencias de corrección (mayor o menor grado de indulgencia del corrector, definición del tipo textual, determinación de

los puntos de corte entre los niveles descriptos en la grilla de corrección, entrenamiento de correctores, entre otros). Entre los comentarios recogidos, podemos destacar los siguientes:

Con respecto a los factores que influyen en la divergencia entre correctores, la mayoría coincidió en que uno de los principales está relacionado con la determinación de puntos de corte en la grilla de evaluación. También apareció como inquietud la necesidad de acordar el peso relativo que se le otorga a cada uno de los criterios.

Otras necesidades manifestadas por algunos correctores fueron: contar con una definición clara de los tipos textuales, poder realizar una puesta en común de dudas, disponer de una descripción ajustada de la grilla de corrección.

Contrariamente a lo esperado, los correctores consultados no atribuyeron una importancia significativa a la mayor o menor indulgencia del corrector como factor de influencia en la divergencia.

En cuanto al uso de los diferentes tipos de escalas para la corrección (analítica y holística), los correctores no encuentran que una otorgue más puntaje que otra. Sin embargo, manifestaron tener más dificultad para evaluar con la escala holística, dado que cuesta mucho asignar el nivel mediante esta modalidad, sin reconstruir mentalmente la escala analítica.

Por último, de los cuatro aspectos descriptos en la escala de evaluación de escritos CELU, las adecuaciones contextual y discursiva son las que resultan de mayor relevancia para el corrector. Esto demuestra una buena comprensión del enfoque CELU por parte de los correctores. Como este examen mide la habilidad del candidato para usar la lengua en situaciones similares a las de la vida real, será fundamental la adecuación del texto al contexto planteado en la consigna, sobre todo en lo que se refiere a la caracterización de los interlocutores y el propósito funcional de la tarea. Este es un principio que debe ser tenido en cuenta por los correctores en el momento de evaluar la producción. La adecuación contextual se presenta como el criterio a partir del cual se evaluarán todos los demás. Es decir que un candidato que obtenga intermedio en adecuación contextual no podrá obtener avanzado en los otros criterios.

Conclusiones

Luego de haber expuesto la diferencia de enfoque que tienen las pruebas de desempeño con respecto a otros tipos de evaluación de proficiencia en L2, resulta evidente que es fundamental que esta visión quede claramente registrada en los descriptores de la grilla que se va a utilizar en la corrección de este tipo de exámenes. Sin embargo, coincidimos con la duda que plantea Sidi (2002) sobre si todos los correctores toman en consideración de manera apropiada y consistente los criterios de las escalas. Hemos visto cómo algunas investigaciones realizadas sobre el comportamiento evaluativo de los correctores han encontrado, por ejemplo, que la falta de

experiencia y entrenamiento hace que algunos de ellos tiendan a concentrarse en la adecuación gramatical y la cuenta de errores como factores decisivos para la asignación de niveles, cuestión que resultaría contraria al enfoque del examen CELU, tal como ha sido descripto en este trabajo. Si tomamos en cuenta las respuestas a la encuesta realizada, esta dificultad parece no afectar a los correctores del CELU. Sin embargo, a partir de las observaciones consignadas en las planillas de evaluación, hemos notado algunas dificultades en la interpretación de la grilla, ya que, por ejemplo, a veces se justifica la asignación de niveles diferentes utilizando los mismos argumentos, tomados de los descriptores. Esto demuestra, en forma coincidente con las inquietudes planteadas por los correctores, que es necesario continuar con el trabajo de entrenamiento y discusión de los diferentes criterios con los que se evalúa la producción escrita y los lugares donde situar los puntos de corte entre los niveles.

Con respecto al entrenamiento, hay que concluir en que, si bien es necesario, no logran reducir por completo las divergencias. Hay que continuar con las investigaciones referidas al impacto que tienen en los resultados obtenidos en las correcciones posteriores. Para esto son muy útiles los protocolos verbales. Sin embargo, hay que considerar que las muestras sobre la que se investiga a menudo no son lo suficientemente amplias y que no es posible tomar como prueba contundente lo mencionado por los correctores en los protocolos verbales, ya que es imposible realizar un registro completo de la actividad cognitiva. No se pueden tomar en cuenta todos los factores que afectan el proceso de la toma de decisiones en una corrección; por este motivo es importante contar con más de una corrección. La confrontación de los diferentes puntos de vista permite que salgan a la luz las ideas que están sosteniendo para cada corrector este proceso, que pueden contribuir a lograr la construcción de un perfil más completo del desempeño de los candidatos. En este sentido, tomando como ejemplo la película “Rashomon”, del director japonés Kurosawa, McNamara (1996) se pregunta dónde reside la verdad cuando hay diferentes visiones de un mismo hecho. Llega a la conclusión de que, a pesar de que no coincidan en muchos aspectos, paradójicamente, todas contribuyen a echar luz sobre el suceso.

MATERIAL ANEXO

	AVANZADO (4)	INTERMEDIO (3)	BÁSICO (2)	NO ALCANZA (1)
Adecuación contextual				
Adecuación discursiva				
Morfosintaxis				
Léxico				

Encuesta a correctores de escritos Celu

Tema: divergencias en asignación de nivel

- **Marque con una cruz en qué correcciones de exámenes Celu intervino**

204	
105	
205	
UAP	
106	
206	

- **Responda las siguientes preguntas**

1) ¿Qué aspectos del texto que debe evaluar (morfosintáctico, discursivo, contextual, léxico) tienen más relevancia, según su criterio, en la formación de una primera impresión?

2) En una corrección holística, ¿a qué estrategias recurre? (impresión global, reconstrucción mental de la grilla analítica, etc)

3) ¿Cuál de las dos correcciones (holística y analítica) le presenta mayor dificultad y por qué?

4) ¿Piensa usted que una de las dos modalidades (analítica u holística) otorga un nivel más alto que la otra?

5) En los casos de divergencia que usted pudo observar, ¿qué factores cree usted que influyeron?

.

- La formulación de la consigna de la tarea que realizó el candidato.

- La definición del tipo textual solicitado.

- La descripción de los criterios en la grilla de corrección (adecuación contextual, adecuación discursiva, morfosintaxis y léxico).

- La determinación de los puntos de corte entre los diferentes niveles.

- El peso relativo atribuido a los criterios de corrección para llegar al resultado final.

- El entrenamiento de los correctores.

- La personalidad severa/indulgente del corrector.

- Otros:

- **Escriba a continuación sus observaciones y/o sugerencias con respecto al tema de la encuesta**

REFERENCIAS BIBLIOGRÁFICAS

ALDERSON, J.Charles; CLAPHAM, Caroline y WALL, Dianne (1995) *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

BACHMAN, Lyle; PALMER, Adrian (1996) *Language testing in practice*. Oxford: Oxford University Press.

CERTIFICADO DE ESPAÑOL – LENGUA Y USO. Universidad de Buenos Aires, Universidad del Litoral, Universidad de Córdoba, 2005/2006.

CUMMING, Alister (1997) The Testing of L2 Writing. En *Enciclopedia of Language and Education. Volume 7. Language testing and assessment*. Kluwer Academia Publishers.
.
HAMP-LYONS, Liz (1995) *Rating nonnative writings, the trouble with holistic scoring*. TESOL Quarterly, n.29.

HUGHES, Arthur (1989) *Testing for language teachers*. Cambridge: Cambridge University Press.

HUOT, B.A. (1998) The validity of holistic scoring: a comparison of the talk-aloud protocols of expert and novice holistic raters. Indiana University of Pennsylvania.

LUMLEY, Tom (2002) Assessment criteria in a large-scale writing test: what do they really mean to the raters? En *Language Testing*, vol.19

MCNAMARA, Tim (1996) *Measuring second language performance*. London: Longman.

MCNAMARA, Tim (2001) Performance testing. En *Enciclopedia of Language and Education. Volume 7. Language testing and assessment*. Kluwer Academia Publishers.

MCNAMARA, Tim (2000) *Language Testing*. Oxford: Oxford University Press.

MORROW, Keith (1979) Communicative language testing: revolution or evolution? In: BRUMFIT, C.J.

JOHNSON, K. (Org.). *The communicative approach to language teaching*. Oxford: Oxford University Press.

PERKINS, Kyle.(1983) *On the use of composition scoring techniques, objetive measures, and objective tests to evaluate ESL writing ability*. TESOL. Quarterly.

SCARAMUCCI, Matilde (2000) Proficiencia em LE: consideraçoes terminologicas e conceituais. Trabalhos. En *Lingüística Aplicada*, n.36, lul/dez.

SCARAMUCCI, Matilde (2005) Prova de redaçao nos vestibulares: Educacionalmente benéfica para o ensino/aprendizagem da escrita?

SCHLATTER, Margarete (1998) CELPE-BRAS: Certificado de lengua portuguesa para estangeiros – breve histórico. In CUNHA, Maria, Santos, Percilia. Ensino e pesquisa em portugues para estrangeiros. Brasilia: UNB.

SCHLATTER, M.; GARCEZ, P.M. y SCARAMUCI, M. (2004) *O papel da interaçao na pesquisa sobre aquisiçao e uso da lengua estrangeira: implicaçoes para o ensino e para a avaliacçao*. Letra de Hoje, 39 (3).

SIDI, Walkiria (2002) Niveis de proficiencia em lectura e escrita de falantes de español no exame Celpe-Bras. Dissertacao de Mestrado, Universidad Federal do Rio Grande do Sul.

WEIGLE, Sara Cushing (1994). Effects of training on raters of ESL compositions. Cp. En *Language Testing*.

WEIGLE, Sara Cushing (2001) Scoring procedures for writing assessment. En *Assessing writing*. Cambridge: Cambridge University Press. Cp.6

WEIGLE, Sara Cushing (2001) Designing writing assessment tasks. En *Assessing writing*. Cambridge: Cambridge University Press.Cp.5

WEIR, C. J. (1995) *Understanding and Developing Language Tests.* Hemel Hempstead: Prentice Hall International (UK) Ltd.