PSYKHE
2003, Vol.12, N° 2, 3-13

ISSN 0717-0297

Estudio Acerca del Uso de Preguntas de Ensayo en Pruebas de Selección Universitaria

Study on the Use of Essay Questions in College Admission Tests

Jorge Manzi, María Paulina Flotts, Erika Himmel y Ricardo Rosas
Pontificia Universidad Católica de Chile

David Bravo
Universidad de Chile

Este estudio presenta evidencia empírica para respaldar el eventual uso de preguntas de ensayo en las pruebas de selección universitaria en Chile. En el marco de un proyecto de investigación que ha estado revisando tales pruebas (Proyecto SIES), se desarrollaron, aplicaron y corrigieron cinco versiones alternativas de preguntas de ensayo en una muestra de 458 estudiantes de último año de enseñanza secundaria. Cada pregunta fue corregida con dos tipos de rúbrica (analítica y holista) por dos jueces independientes en cada caso. Los resultados muestran un alto grado de consistencia y acuerdo entre los jueces con ambos tipos de rúbrica. Por otra parte, las correlaciones de estas preguntas con otras variables cognitivas y sociodemográficas mostraron que ellas aportan información adicional a la que proveen las pruebas tradicionales de admisión y evidenciaron resultados positivos en términos de equidad, considerando tanto el género como el capital sociocultural de los examinados.

In this article we present evidence supporting the use of open-ended essay questions in college admission tests. Data was obtained from a larger study whose main focus was to provide empirical evidence for the validity of the newly developed college admission tests in Chile (SIES, Sistema de Ingreso a la Educación Superior). The present study examined the answers given by 458 high school seniors to five different essay questions. Two independent judges graded each question by using two different types of rubrics (holistic and analytic). Results showed that consistency and agreement among judges were high irrespective of the type of rubric used to grade the questions. In addition, correlations of the scores in these questions to cognitive and demographic factors revealed that they provide information above and beyond the information provided by the current college admission tests. Finally, the results are also positive in view of the sociocultural and gender differences associated with educational tests.

Las pruebas para el ingreso a la educación superior en Chile han estado bajo análisis en los últimos tres años. En este período se ha producido un importante proceso de cambio, que ha llevado a la modificación de los instrumentos vigentes desde hace más de 30 años en el país.

Este cambio comenzó formalmente con el trabajo de una comisión ad hoc que durante el año 2000 hizo una revisión y análisis del sistema PAA-PCE (Prueba de Aptitud Académica y Pruebas de Conocimientos Específicos), trabajo que resultó en un informe que proponía: (a) desarrollar nuevos instrumentos de selección para la educación superior que estuvieran alineados con el currículum de la enseñanza media, y (b) actualizar la base cognitiva y metodológica en que se basan las pruebas o instrumentos (Berríos et al., 2000).

El proyecto SIES (Sistema de Ingreso a la Educación Superior) recogió las principales recomendaciones hechas por la comisión y desarrolló bancos de preguntas correspondientes a estas nuevas orientaciones. En concreto, este proyecto produjo ítemes alineados con el currículum de la enseñanza media, correspondientes a cuatro sectores de aprendizaje: Lenguaje, Matemática, Ciencias (que incluye Biología, Física y Química) e Historia y Ciencias Sociales.

El desarrollo de estas preguntas se basó en un modelo de representación del conocimiento que explicita

Jorge Manzi, María Paulina Flotts y Ricardo Rosas, Escuela de Psicología. Erika Himmel, Facultad de Educación. David Bravo, Facultad de Ciencias Económicas y Administrativas. La correspondencia relativa a este artículo deberá ser dirigida a Jorge Manzi, Escuela de Psicología, Facultad de Ciencias Sociales, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Santiago, Chile. E-mail: jmanzi@puc.cl
Este estudio contó con el financiamiento del Fondo Nacional para el Desarrollo Científico y Tecnológico, FONDEF (Proyecto D00I1080).

la relación entre habilidades y contenidos, organizando ambas dimensiones en continuos que admiten diversos grados de complejidad (Rosas, Flotts & Saragoni, 2002). A partir de dicho modelo es posible distinguir, en un extremo, el conocimiento básico (combinación de contenidos y habilidades elementales) y en el otro el conocimiento avanzado (combinación de contenidos y habilidades superiores). Las dos dimensiones mencionadas definen el espacio que sirvió de referencia para la producción de las preguntas del proyecto SIES.

Aunque la gran mayoría de las preguntas producidas en dicho proyecto emplean el formato de preguntas de selección múltiple, se decidió también construir y evaluar preguntas de respuesta abierta, puesto que se entendía que el formato de selección múltiple no permite evaluar habilidades también fundamentales para el éxito en la enseñanza superior, como es el caso de la capacidad para organizar y exponer ideas en forma escrita. La capacidad argumentativa representa una habilidad recurrente en todos los campos del saber, por lo que su incorporación a la batería de selección universitaria representaría una adición deseable.

La dificultad técnica con este tipo de preguntas, como se expone más adelante, está referida a la posibilidad de estimar en forma confiable y objetiva los puntajes, en el entendido que ellos deben ser determinados con procedimientos que involucran un riesgo de subjetividad. Adicionalmente, existe la potencial complejidad práctica asociada al tiempo que demanda la corrección de este tipo de preguntas en un contexto masivo como el que supone su uso en pruebas de selección universitaria. Estos aspectos se abordan y analizan en este trabajo, junto con los fundamentos que hacen recomendable la incorporación de preguntas de este tipo a la selección de estudiantes para la enseñanza superior.

El Uso de Preguntas de Ensayo en Pruebas de Selección Universitaria

A nivel empírico, y comparando la experiencia internacional sobre el tema, es posible constatar que en diversas partes del mundo se emplean las preguntas de ensayo en la selección para la educación superior, ya sea en forma complementaria a la aplicación de "pruebas objetivas" (basadas en preguntas de selección múltiple), como es el caso del SAT norteamericano, o en forma principal, es decir, como el único tipo de evaluación empleada, como es el caso del bachillerato francés.

La batería de selección universitaria más investigada a nivel internacional, el SAT de EE.UU.[1], incorpora actualmente un Writing Test como parte de las pruebas complementarias (SAT-Subject Tests), y a partir del año 2005 será incorporada la escritura de un ensayo como parte de las pruebas obligatorias. Este cambio en el SAT, que incorporará también algunas modificaciones en la prueba de matemáticas, busca mejorar y ampliar la batería de selección, tanto en la perspectiva de su capacidad predictiva, como en términos de las señales que las pruebas envían a la enseñanza secundaria. Según lo expresa el organismo responsable de dicha prueba, "las investigaciones han demostrado que el añadir una prueba de redacción incrementa la validez del pronóstico de éxito universitario, pero lo más importante es que también lanza un claro y contundente mensaje, que la buena redacción es imprescindible para el éxito en la universidad y en el futuro" (College Board, 2002, p. 1).

Producción de Ensayos, Razonamiento y Argumentación

La expresión escrita a través de la producción de un ensayo constituye una herramienta que permite conocer los procesos de razonamiento de un individuo. En un sentido general, el razonamiento "puede ser definido como procesos de construcción y evaluación de argumentos" (Shaw, 1996, p. 52). En términos de la lógica clásica, se entiende que los argumentos están formados por conclusiones y premisas, y la calidad de una inferencia depende de la relación que exista entre las premisas y las conclusiones (Shaw, 1996). Pero esta no es la única forma de analizar la logicidad de una argumentación: "los teóricos de la argumentación estudian la forma en la que las personas aceptan y defienden puntos de vista, mientras que los estudiosos de la lógica tienden a concentrarse en la manera en que las conclusiones se han derivado de las premisas" (Van Eemeren, Grootendorst & Snoeck, 1996, p. 6). Así la teoría de la argumentación (o el estudio del razonamiento informal) constituye un marco de referencia fundamental para comprender los procesos de razonamiento.

La argumentación se entiende como "una actividad verbal y social de razonamiento que apunta a aumentar (o disminuir) la aceptabilidad de un punto

[1] Modelo en el que, dicho sea de paso, se basa el antiguo sistema chileno de PAA-PCE.

de vista controversial para el oyente o lector, a través de proponer una constelación de proposiciones que buscan justificar (o refutar) el punto de vista antes de un juicio racional" (Van Eemeren et al., 1996, p. 5). En este mismo sentido, el razonamiento informal se entiende como "el proceso intencional[2] que involucra generar o evaluar (o ambos) evidencia relacionada con una declaración o conclusión" (Means & Voss, 1996, p. 139), y es un tipo de razonamiento que "adquiere importancia cuando la información es poco accesible, o cuando los problemas son muy abiertos, debatibles, complejos o mal estructurados, y especialmente, cuando el asunto requiere que los individuos construyan un argumento para sustentar una declaración" (Means & Voss, 1996, p. 140).

Mientras que para la lógica clásica proposicional el interés está puesto en distinguir argumentos válidos e inválidos en función de ciertas premisas abstractas que se estructuran formalmente en constantes lógicas, la argumentación o razonamiento informal tiene como objetivo "desarrollar criterios para determinar la validez de la argumentación considerando sus puntos de partida y la disposición de su presentación, e implementar la aplicación de estos criterios a la producción, análisis y evaluación del discurso argumentativo" (Van Eemeren et al., 1996, p. 22).

Investigación en psicología cognitiva y educacional ha puesto de manifiesto la relevancia de las habilidades de razonamiento informal en el aprendizaje, especialmente en dominios como las ciencias sociales y humanidades, en las que se deben desarrollar, defender y evaluar ideas y perspectivas (Glaser, 1984; Perkins, 1985; Resnick, 1987).

La importancia de expresar ideas y argumentar a favor (o en contra) de ellas ha sido explícitamente considerada en el currículum de Comunicación y Lenguaje chileno en tercer año medio, donde se propone el aprendizaje de discursos de tipo argumentativo. La inclusión y explicitación de este tipo de discurso es justificada aludiendo a su funcionalidad en distintos contextos comunicativos (Ministerio de Educación [MINEDUC], 1988).

Dificultades Técnicas en el Uso de Preguntas de Ensayo

Asignar puntajes o calificaciones a ensayos de estudiantes no es una tarea exenta de dificultades, desde las más obvias como la "subjetividad" de las puntuaciones o el tiempo requerido para hacerlo, hasta algunas más complejas, como la adecuación de los puntajes a ciertos estándares de desempeño.

Hacer evaluaciones del desempeño de un estudiante implica contestar preguntas del tipo ¿cuáles son los aspectos o dimensiones que serán puntuados?, ¿cuáles son los criterios que deben aplicarse para evaluar esos aspectos del desempeño?, ¿cómo se deben desarrollar los criterios para la puntuación? y ¿cómo deben aplicarse estos criterios? (Clauser, 2000). Todas estas preguntas impactan directamente en la confiabilidad de las puntuaciones, en tanto la claridad respecto de qué evaluar en un producto y cómo hacerlo contribuye a la estabilidad de dichas puntuaciones.

El estudio de la estabilidad de las puntuaciones ha estado orientado a delimitar cuáles son los posibles factores que interferirían en ella, tanto a nivel intraindividual (es decir, estabilidad de las puntuaciones de un mismo juez en el tiempo) como interindividual (consistencia entre las puntuaciones asignadas por distintos jueces). A nivel intraindividual se ha estudiado, por ejemplo, el efecto de la longitud del período de evaluación en el nivel de exigencia con que los jueces califican (Congdon & McQueen, 2000), encontrando que hay un efecto del tiempo en que los ensayos son corregidos, lo que hace recomendable la práctica de recalibrar a los jueces durante el proceso de corrección de preguntas cuando este proceso se extiende en el tiempo.

A nivel de factores interindividules, se han estudiado las diferencias entre jueces expertos y novicios, encontrándose que ambos tipos de correctores difieren en la comprensión que tienen de la rúbrica, y los procedimientos de toma de decisiones que emplean (Wolfe, Kao & Ranney, 1998). En términos generales, los jueces expertos tienden a leer el ensayo completo y después puntuar según la rúbrica, mientras que los jueces novicios suelen interrumpir el proceso de lectura para monitorear cuán bien satisface el ensayo los criterios de puntuación (Wolfe et al., 1998).

Consistencia y Acuerdo entre Jueces

Los estudios acerca de la corrección de preguntas de respuesta construida o abierta frecuentemente se han centrado en establecer el grado en que diversos evaluadores concuerdan en sus apreciaciones con respecto a los objetos evaluados. En este tipo de estudios, dos o más jueces evalúan un conjunto de respuestas y se registra el grado en que las calificaciones concuerdan. Tinsley y Weiss (2000) plantean que

[2] "Goal-dependent" en la versión original.

dos aspectos deben ser diferenciados y evaluados cuando se estudia la concordancia interjueces: uno es el grado de consistencia o confiabilidad entre ellos y el otro es el grado de acuerdo o concordancia. Ambos aspectos son parcialmente una función de los objetos evaluados, la escala de evaluación y los jueces que efectúan las evaluaciones.

La confiabilidad o consistencia apunta a establecer el grado en que las calificaciones otorgadas son sensibles a las diferencias entre los objetos evaluados. En este caso, lo que se busca establecer es si diversos jueces concuerdan en el ordenamiento de los objetos evaluados, aún si las puntuaciones absolutas que empleen sean discrepantes. El indicador más apropiado para este tipo de análisis cuando las escalas de evaluación son ordinales o intercalares es el índice de correlación intraclase.

Por otra parte, el acuerdo interjueces refleja el grado en que dos o más jueces coinciden en asignar la misma categoría de puntaje a un mismo objeto. Cuando las categorías de puntaje empleadas son cualitativamente diferentes, este aspecto se puede determinar con el índice kappa de Cohen. Cuando se emplean escalas ordinales o intervalares, lo recomendable es recurrir al índice T de Lawlis y Lu (1972), que posee una prueba de significación estadística basado en la distribución Chi-cuadrado.

Dado que en este estudio se emplean escalas de puntuación ordinales, la correlación intraclase así como el índice T de Lawlis y Lu serán empleados para estimar la consistencia y acuerdo de los jueces respectivamente.

Objetivos del Estudio

En este trabajo se analiza el valor de preguntas de ensayo como parte de una batería de selección de postulantes a la educación superior. En este contexto, se evalúa el grado en que este tipo de preguntas producen resultados equivalentes al ser juzgadas por distintos jueces. Complementariamente, se aporta evidencia acerca del grado en que este tipo de preguntas aporta información adicional a la que proveen los instrumentos tradicionales de selección universitaria, basados en preguntas de selección múltiple. Finalmente, este trabajo también apunta a estimar la relación entre los puntajes producidos con este tipo de preguntas y variables sociodemográficas, como el sexo de los evaluados, o la escolaridad de sus padres. Estos últimos resultados son analizados en la perspectiva de la validez y equidad de mediciones basadas en preguntas de ensayo.

Método

Este estudio se llevó a cabo en el contexto del estudio de campo principal del proyecto SIES que involucró una muestra de casi 12.000 estudiantes de cuarto año medio. En este estudio, además de evaluar el funcionamiento de las preguntas de selección múltiple construidas para los cuatro sectores de aprendizaje previamente mencionados, se evaluó en una submuestra el funcionamiento de preguntas de ensayo. En concreto, cada estudiante participante en esta parte del estudio respondió inicialmente una de las formas de la prueba de lenguaje y a continuación una de las preguntas de ensayo. A continuación se presentan las características metodológicas de este estudio.

Participantes

Los participantes fueron 458 alumnos de cuarto año de enseñanza media de nueve establecimientos educacionales del país, de los tres tipos de dependencia administrativa: municipales, particulares subvencionados y particulares pagados. La composición de la muestra, incluyendo su distribución según el género de los estudiantes, se presenta en la Tabla 1.

Tabla 1
Composición de la muestra de estudiantes que respondieron la pregunta de ensayo

	Particular pagado	Particular subvencionado	Municipalizado	Total
Hombres	28	116	103	247
Mujeres	59	83	69	211
Total	87	199	172	458

Instrumentos

1. *Pregunta de ensayo*: se prepararon cinco versiones diferentes de estas preguntas. Su estructura era siempre la misma, variando sólo el tema. Cada versión incluía los siguientes elementos: (a) presentación de una afirmación acerca de algún tema de interés general, cercano a la realidad juvenil y que admitiera diversos puntos de vista (por ejemplo, medios de comunicación y violencia, desarrollo económico y preservación de la naturaleza, diferenciación de roles masculinos y femeninos)[3], (b) instrucciones para el ensayo, donde se enfatizaba que se esperaba que el examinado entregara una opinión fundamentada con respecto a la afirmación que se le había presentado; se le solicitaba que en su argumentación considerara al menos un punto de vista alternativo al suyo, y (c) explicitación de aspectos que serían tomados en cuenta al momento de evaluar el ensayo. Se mencionaban aspectos de contenido y estructura (tales como la organización de las ideas) así como aspectos formales (ortografía, vocabulario y redacción).
2. *Prueba de Lenguaje SIES*: los estudiantes respondieron también una prueba con 84 preguntas de selección múltiple

[3] Dado que el foco del estudio se centró en la evaluación de habilidades argumentativas, se optó por temas relativamente familiares, de manera que el contenido no fuera un obstáculo para argumentar. Esta es una decisión que puede ser modificada en aplicaciones reales.

correspondiente al sector de Lenguaje. Las preguntas fueron presentadas en dos bloques, el primero de 52 y el segundo de 32 preguntas. El puntaje en esta prueba se calculó empleando una escala con promedio y desviación estándar equivalentes al que ha sido empleado tradicionalmente en las pruebas de selección universitarias chilenas (promedio 500, desviación estándar 100). Estos puntajes se basaron en la distribución obtenida con el total de estudiantes que respondió la prueba de Lenguaje en el estudio de campo del proyecto SIES.

Procedimiento

Todos los sujetos participantes respondieron primero la prueba del SIES de Lenguaje (para la que se asignaron 215 minutos), y luego contestaron una de las cinco versiones de la pregunta de ensayo, para la que asignó un tiempo máximo de 30 minutos. La versión de la pregunta fue asignada al azar a los estudiantes de cada curso evaluado.

1. *Corrección de los ensayos*: cada uno de los ensayos fue corregido en dos instancias sucesivas. En un primer momento, se ocupó una rúbrica analítica para la corrección y luego una rúbrica holista. En ambos casos cada ensayo fue corregido por dos evaluadores independientemente.
2. *Elaboración de rúbricas para la evaluación de los ensayos*: una rúbrica es un conjunto de pautas que permiten evaluar el trabajo de un alumno y que guardan relación con ciertos criterios y estándares de rendimiento (Goodrich, 1997; Hart, 1994; Hermann, Aschbacher & Winters, 1992; Tombari & Borich, 1999). La rúbrica comprende una escala de medición fija y una lista de criterios que describen distintos niveles de desempeño.

La clasificación más común de las rúbricas es aquélla que distingue rúbricas analíticas y holistas. Las rúbricas holistas son aquéllas que abordan el desempeño como un todo, basando la puntuación en una apreciación global, mientras que las rúbricas analíticas son aquéllas que evalúan el desempeño en un conjunto de dimensiones previamente especificadas (Tombari & Borich, 1999).

Para los efectos de la corrección de las preguntas de ensayo probadas en la aplicación experimental del SIES, se diseñó una rúbrica analítica y una rúbrica holista que tomaron en consideración aquellos aspectos del desempeño que son relevantes desde una perspectiva curricular y a partir del conocimiento e investigación en torno a la teoría de la argumentación. Para ambas rúbricas se decidió trabajar con cuatro niveles de desempeño.

La rúbrica analítica contempló las siguientes diez dimensiones para la evaluación del ensayo:

1. Vocabulario: grado de variedad y precisión del vocabulario empleado y adecuación de éste al registro de habla esperado.
2. Ortografía: grado en que se emplean correctamente las reglas ortográficas del español (ortografía literal, puntual y acentual).
3. Redacción 1: grado de cohesión textual, es decir, uso adecuado de la gramática oracional y de los conectores que vinculan las oraciones.
4. Redacción 2: coherencia global del texto, incluyendo la estructuración del texto en párrafos.
5. Pertinencia: grado en que el contenido de la argumentación se adecúa al enunciado.
6. Tesis o idea central: grado en que el texto identifica un tema o idea central que refleja un punto de vista o posición acerca del tema enunciado.
7. Argumentos que apoyan la idea central: grado en que el texto identifica argumentos para apoyar la idea central. Los argumentos pueden provenir de experiencias personales, conocimientos, lecturas, estudio, etc.
8. Consideración de puntos de vista alternativos: grado en que el texto incluye la consideración de contrargumentos.
9. Organización global: grado de coherencia del texto, en que las ideas siguen una secuencia lógica. En esta dimensión también se considera que el texto incluya un inicio y un cierre (aunque no necesariamente una conclusión).
10. Evaluación global: apreciación general acerca del contenido y calidad del texto.

Por su parte, la rúbrica holista se construyó de tal manera que los evaluadores pudieran hacer un juicio global acerca de la calidad del texto atendiendo a cuán claro, coherente y convincente fuera; si exponía o no un argumento central; si era capaz de relacionarlo adecuadamente con contrargumentos; si el texto estaba bien organizado; y si cumplía con los requisitos formales de vocabulario, ortografía y redacción. En esta rúbrica se utilizó también una escala de 4 puntos, pero se permitió que los evaluadores cualificaran positiva o negativamente el puntaje asignado (agregando el calificador "+" o "-") atendiendo a aspectos del desempeño que fueron explicitados en la rúbrica.

3. *Selección de benchmarks o ejemplares*: para ejemplificar los distintos niveles de desempeño que distinguen ambas rúbricas, se escogieron ensayos prototípicos de cada uno de estos niveles, de manera que los evaluadores pudieran contar con ejemplares frente a los cuales contrastar sus juicios. El empleo de benchmarks es un recurso habitual en el entrenamiento y calibración de correctores de preguntas abiertas, puesto que especifican la forma en que puede aplicarse la pauta de corrección. En este caso, los benchmarks fueron previamente seleccionados por el equipo técnico del proyecto.
4. *Selección y capacitación de los evaluadores*: se seleccionaron ocho evaluadores, todos ellos profesores de enseñanza media en Lenguaje y Comunicación, con experiencia en sala de clases. Cuatro de ellos trabajaron en la rúbrica analítica y cuatro en la holista.

Con cada uno de los equipos de evaluadores se realizó una capacitación para aprender el manejo de la rúbrica y acordar criterios para la corrección. En esta capacitación se llevó a cabo un proceso de calibración, donde se realizó un trabajo conjunto de evaluación de algunos ensayos ocupando para ello la rúbrica y los benchmarks. Una vez concluido el entrenamiento, los evaluadores fueron organizados para corregir las cinco preguntas de ensayo según el plan que especifica la Tabla 2. De acuerdo a éste, cada corrector corrigió completamente dos preguntas, y la mitad de la última. Al mismo tiempo, cada pregunta fue corregida por dos correctores, alternándose las parejas en las cuatro primeras preguntas.

La aplicación de cada una de las rúbricas se hizo en forma independiente y sucesiva.

Tabla 2
Plan de corrección de los ensayos

	Evaluador 1	Evaluador 2	Evaluador 3	Evaluador 4
Pregunta 1	X	X		
Pregunta 2			X	X
Pregunta 3	X		X	
Pregunta 4		X		X
Pregunta 5	X	X	X	X

Resultados

Consistencia y Acuerdo entre Jueces

Tal como se indicó más arriba, una condición fundamental para el empleo de preguntas de ensayo, es asegurar que diferentes evaluadores entrenados puedan otorgar evaluaciones equivalentes a una misma respuesta. Aunque en condiciones reales de aplicación se emplean mecanismos de seguimiento de las evaluaciones de cada evaluador, estableciéndose instancias de reentrenamiento (recalibración), cada vez que se detectan evaluaciones atípicas, en este caso se procedió a asignar puntaje a todas las preguntas, sin reentrenar a los evaluadores. Esto se hizo así, pues interesaba conocer el nivel básico de consistencia y acuerdo interjueces que era posible obtener a partir de las rúbricas y entrenamiento inicial.

Al comparar las evaluaciones hechas por dos o más jueces es necesario distinguir el grado de consistencia y el grado de acuerdo de las puntuaciones asignadas por ellos. La consistencia interjueces indica el grado en que la varianza en las puntuaciones es atribuíble a diferencias entre los objetos puntuados, mientras que el acuerdo interjueces representa el grado en que los diferentes jueces tienden a asignar la misma puntuación a cada elemento (Tinsley & Weiss, 2000). Por ello, y considerando que la escala de respuesta empleada en este caso es de tipo ordinal, se procedió a calcular la correlación intraclase y el índice de concordancia T de Lawlis y Lu (1972) para estimar ambos aspectos de la concordancia entre los jueces.

Los resultados de las correlaciones intraclase, que aparecen en la Tabla 3, muestran valores uniformemente altos para la rúbrica holista en todas las preguntas[4]. En el caso de la rúbrica analítica, los resultados son también positivos aunque algo menos favorables. Las dimensiones que se asocian a menores niveles de consistencia en este caso son: vocabulario, pertinencia, tesis y punto de vista alternativo. La columna de la Tabla 3 que especifica los resultados para dos jueces corresponde a la proyección de la consistencia, basada en la fórmula de Spearman-Brown, asumiendo que la puntuación es producida por dos jueces. En su conjunto, los resultados obtenidos revelan que los evaluadores ordenaron en forma confiable los ensayos.

Para evaluar el grado de acuerdo entre jueces se empleó el índice T de Lawlis y Lu (1972). Este permite estimar el grado de acuerdo entre jueces variando el nivel de concordancia admisible. En este caso se evaluó para dos condiciones: cuando la discrepancia entre los jueces es de 0 puntos (es decir, cuando hay pleno acuerdo), y cuando la discrepan-

Tabla 3
Correlaciones intraclase para cada pregunta según rúbrica analítica y holista

	Pregunta 1		Pregunta 2		Pregunta 3		Pregunta 4	
	1 juez	2 jueces	1 juez	2 jueces	1 juez	2 jueces	1 juez	2 jueces
Vocabulario	.54	.70	.54	.70	.74	.85	.31	.47
Ortografía	.86	.92	.69	.82	.90	.95	.74	.85
Redacción 1	.62	.77	.59	.74	.80	.89	.54	.70
Redacción 2	.63	.78	.74	.85	.57	.73	.73	.84
Pertinencia	.23	.38	.42	.59	.52	.68	.55	.71
Tesis	.33	.49	.53	.69	.61	.76	.58	.74
Argumentación	.54	.70	.40	.57	.48	.65	.50	.66
Punto de vista alternativo	.49	.66	.41	.58	.41	.58	.40	.57
Organización	.74	.85	.59	.74	.51	.67	.42	.59
Evaluación global	.66	.79	.69	.81	.83	.91	.63	.77
Rúbrica analítica	.73	.84	.75	.86	.82	.90	.75	.86
Rúbrica holista	.91	.95	.89	.94	.89	.94	.90	.95

[4] Para simplificar la presentación de resultados, se incluyen en las tablas las cuatro primeras preguntas, puesto que en la quinta se mezclan los cuatro correctores.

cia es de 1 punto. La Tabla 4 muestra que en términos generales se constataron altos grados de acuerdo en todos los indicadores de la rúbrica analítica, así como en la rúbrica holista. De hecho, el porcentaje de preguntas que resultan clasificadas en la misma categoría de puntaje es superior al 70% en ambos casos. Cuando se suman los casos donde la discrepancia llega a 1 punto, el porcentaje de acuerdo se acerca al 100%. Esto revela que en las condiciones evaluadas en este estudio se obtienen niveles adecuados de acuerdo entre jueces.

Cabe mencionar que la pauta holista produce, en términos globales, niveles algo superiores de consistencia y acuerdo, lo que sumado a su mayor rapidez de aplicación, sugiere que esta pauta es preferible para mediciones masivas que requieran obtener puntuaciones en poco tiempo, como es el caso del uso de preguntas de ensayo en pruebas de selección universitaria.

Relación entre las Preguntas de Ensayo y otros Factores de Selección

Dado que la aplicación de estas preguntas de ensayo se dio en el contexto de un estudio de campo referido a pruebas de selección universitaria, se obtuvo información de los examinados que permite establecer la relación que se produce entre las preguntas de ensayo y otros factores de selección. Se dispone de información de dos pruebas de lenguaje: la que se estaba evaluando en el estudio de campo del proyecto SIES[5] y la PAA Verbal. Adicionalmente, se cuenta con la información de las otras pruebas de admisión regulares[6], así como el promedio de notas de la enseñanza media de los estudiantes.

Los resultados que aparecen en la Tabla 5 muestran que las preguntas de ensayo se correlacionan positivamente con todos los factores de selección universitarios. Las correlaciones son moderadas, lo que es consistente con la suposición que las preguntas de ensayo aportan información con respecto a habilidades o capacidades que no miden las pruebas regulares, las que sólo emplean preguntas de selección múltiple. Adicionalmente, se observa que las correlaciones son levemente superiores con las pruebas de lenguaje, lo que reafirma que las habilidades involucradas en la pregunta de ensayo dicen relación con el área de lenguaje. Sin embargo, las correlaciones con las otras pruebas son cercanas, lo que indica que las preguntas de ensayo comparten con todos los factores de selección varianza común.

Tabla 4
Resumen de índices de acuerdo interjueces (Coeficiente T de Lawlis & Lu)

	Pregunta 1		Pregunta 2		Pregunta 3		Pregunta 4	
	0 discr	1 discr	0 discr	1 discr	0 discr	1 discr	0 discr	1 discr
Vocabulario	.76	.97	.75	1	.93	1	.81	1
Ortografía	.72	1	.72	.91	.78	1	.51	.88
Redacción 1	.58	.91	.52	.97	.77	1	.41	.94
Redacción 2	.60	.94	.54	.97	.50	.97	.46	.86
Pertinencia	.37	1	.39	.94	.73	1	.36	.86
Tesis	.55	1	.54	.91	.73	1	.52	.94
Argumentación	.61	1	.58	.94	.70	.94	.65	.97
Punto de vista alternativo	.58	.79	.30	.80	.24	.94	.33	.71
Organización	.75	1	.55	.97	.63	1	.68	.94
Evaluación global	.72	1	.72	.97	.95	1	.85	1
Rúbrica holista	.79	.97	.65	1	.73	.97	.65	1

*Todos los valores de T son estadísticamente significativos ($p < .01$).

[5] El diseño del estudio de campo del proyecto SIES implicó que quienes respondieron las preguntas de ensayo respondieron adicionalmente sólo la prueba de lenguaje de ese proyecto. Por ello no es posible estimar la correlación de las preguntas de ensayo con otras pruebas del proyecto SIES.

[6] Dado que casi todos los participantes en el estudio de campo rindieron posteriormente las pruebas regulares de admisión, fue posible contar con los resultados de ellas. Sin embargo, no se incluyeron en los análisis las pruebas de conocimientos específicos, dado que el número de casos disponibles para las correlaciones, por el carácter optativo de las pruebas, es sustancialmente menor.

Con el propósito de profundizar la relación de las preguntas de ensayo con los otros factores de selección, se llevó a cabo un análisis de componentes principales con los factores de selección ya mencionados. Este análisis reveló, tal como cabía esperar, que existía importante evidencia a favor de un factor común, el que explica un 58.2% de la varianza total. Sin embargo, el segundo factor también explicaba un porcentaje importante de dicha varianza (14.6%), por lo que se decidió retenerlo. Luego de rotar en forma ortogonal los dos factores seleccionados (con rotación Varimax), se obtuvieron los resultados que se presentan en la Tabla 6. Allí se aprecia que, antes de la rotación, todos los factores tienen una carga positiva en el factor 1, aunque las dos rúbricas de la pregunta de ensayo lo hacen en una magnitud algo menor. Al rotar los factores, la pregunta de ensayo se distingue nítidamente en un segundo factor, donde las dos rúbricas tienen una muy alta carga factorial. Las notas de enseñanza media, y en menor medida las pruebas de lenguaje, aparecen con cargas cercanas o superiores a 0.3 en este factor. El primer factor rotado, en cambio, se asocia con claridad a las pruebas de selección basadas en preguntas de selección múltiple, especialmente las tres pruebas obligatorias del sistema tradicional. Las notas de enseñanza media poseen en este factor un peso algo menor que las pruebas.

En suma, los análisis efectuados revelan que las preguntas de ensayo se asocian a los factores convencionales de selección, pero aportan información adicional. La naturaleza del aporte adicional probablemente es una combinación de las habilidades cognitivas involucradas, así como del método implicado en su evaluación (preguntas de respuesta cerrada versus preguntas de respuesta construida).

Diferencias entre Hombres y Mujeres

Estudios efectuados en EE.UU. han revelado que las mujeres obtienen rendimientos superiores a los hombres en preguntas de ensayo que forman parte de pruebas de selección universitaria. Utilizando los resultados del ACT Assessment Program, Doolittle y Welch (1989) encontraron que los hombres tienen mejores rendimientos en matemáticas, y las mujeres en las preguntas de selección múltiple que miden habilidades de escritura y también en el ensayo

Tabla 5
Correlaciones entre pregunta de ensayo y factores de selección universitaria

	1	2	3	4	5	6	7
1. Rúbrica Holista	1.00	0.61	0.41	0.41	0.37	0.34	0.38
2. Rúbrica Analítica	0.61	1.00	0.45	0.46	0.35	0.36	0.41
3. Prueba SIES Lenguaje	0.41	0.45	1.00	0.76	0.60	0.63	0.44
4. PAA Verbal	0.41	0.46	0.76	1.00	0.75	0.71	0.51
5. PAA Matemática	0.37	0.35	0.60	0.75	1.00	0.65	0.52
6. Prueba de Historia de Chile	0.34	0.36	0.63	0.71	0.65	1.00	0.52
7. Notas E. Media	0.38	0.41	0.44	0.51	0.52	0.52	1.00

Tabla 6
Resultados del análisis factorial

	Antes de Rotación		Luego de Rotación Varimax	
	Factor 1	Factor 2	Factor 1	Factor 2
Rúbrica Holista	.63	.63	.21	.87
Rúbrica Analítica	.65	.61	.24	.86
Prueba SIES Lenguaje	.82	-.14	.77	.30
PAA Verbal	.88	-.22	.87	.26
PAA Matemática	.81	-.30	.85	.17
Prueba de Historia de Chile	.81	-.31	.85	.16
Notas E. Media	.70	.02	.59	.38
% de Varianza Explicada	58.2%	14.6%		

o *writing test*. Asimismo, el mejor rendimiento de las mujeres en preguntas de ensayo ha sido constatado para pruebas como el GMAT[7] (Zwick, 2002) y el SAT II (Stumpft, 1998).

El hecho que las mujeres obtengan mejores puntuaciones en pruebas de ensayo, ha llevado a algunos a considerar que el uso de este tipo de mediciones permite compensar el menor rendimiento de las mujeres en pruebas estandarizadas de admisión universitaria. Por ejemplo, Zwick (2002) plantea: "una prueba de escritura tiene otra ventaja [además de tomar en consideración la importancia de la escritura en el trabajo escolar]: podría contribuir a la equidad en los procesos de admisión para las mujeres" (p. 183). De acuerdo a numerosas investigaciones en el área, la incorporación de los puntajes del SAT II Writing Test en los estudios de validez predictiva, aumenta la predicción en el caso de las mujeres (Zwick, 2002). De hecho, los cambios en el SAT I que consideran la obligatoriedad de las preguntas de ensayo, se fundamentan, entre otros, en los argumentos aquí expuestos (College Board, 2002).

Por esta razón, y considerando además que tradicionalmente las mujeres han obtenido resultados inferiores a los hombres en todas las pruebas de admisión universitarias en Chile (Bravo & Manzi, 2002), resultaba importante analizar los resultados con la pregunta de ensayo desde esta perspectiva. Para llevar a cabo el análisis se comparó en forma estandarizada el promedio de hombres y mujeres, dividiendo la diferencia de los promedios por la desviación estándar común. De esta manera es posible comparar la magnitud de las diferencias con independencia de las escalas empleadas en cada caso. Según se puede constatar en la Tabla 7, las preguntas de ensayo resultan claramente favorables a las mujeres, en magnitudes que exceden lo observado con las otras pruebas de lenguaje. Estos resultados son coherentes con los resultados de los estudios norteamericanos antes mencionados.

Tabla 7
Comparación del promedio de hombres y mujeres en pregunta de ensayo y pruebas de habilidad verbal

	Holista	Analítica	SIES-Len	PAAV
Hombres	2.095	2.725	511.5	545.8
Mujeres	2.588	2.928	526.3	558.2
Desviación Estándar	0.766	0.38	92	114.4
Diferencia Estandarizada	-0.64	-0.53	-0.16	-0.11

[7] Una prueba de selección para estudios de postgrado en EE.UU.

Cabe precisar que en la muestra empleada para este análisis las diferencias entre hombres y mujeres en las pruebas de lenguaje son algo mayores que las observadas en el conjunto de los estudiantes examinados en el estudio de campo. Por ello, cabe esperar que en una muestra más representativa la diferencia aquí observada se atenúe algo. Sin embargo, el patrón debiera ser el mismo: mayores diferencias favorables a las mujeres en este tipo de preguntas que en las de otras pruebas, donde lo convencional es constatar mejores rendimientos de los hombres.

Relación con Escolaridad Paterna

Dado que este estudio se llevó a cabo con una muestra relativamente pequeña de establecimientos educacionales, no es razonable llevar a cabo un análisis que permita establecer diferencias en los resultados según el tipo de dependencia. Sin embargo, sí es pertinente analizar la relación de los puntajes en la pregunta de ensayo cuando se consideran indicadores socioeconómicos individuales. El que resulta más relevante en este caso es la escolaridad de los padres, que como es bien sabido, presenta una importante correlación con los resultados de mediciones educacionales (Mizala & Romaguera, 2000; MINEDUC, 2002). De hecho, para el conjunto de estudiantes evaluados en el estudio de campo del proyecto SIES, se observó que en el caso de las pruebas de lenguaje, la correlación entre la escolaridad de la madre y los puntajes de la PAAV era de .37 y .26 con la prueba SIES de Lenguaje. Las correlaciones respectivas con la escolaridad del padre son similares y levemente superiores. La Tabla 8 muestra los resultados referidos a la muestra de estudiantes que respondió la pregunta de ensayo. Se aprecia en ella que las correlaciones con las pruebas de selección múltiple de lenguaje (PAAV y SIES Lenguaje) son equivalentes a las obtenidas en el conjunto de la muestra del estudio de campo. Lo interesante es que la escolaridad del padre o de la madre se asocia en menor grado con las puntuaciones de las preguntas de ensayo, lo que sugiere que este tipo de mediciones depende en menor medida del capital cultural de la familia.

Tabla 8
Correlaciones con la escolaridad de los padres

	Holista	Analítica	SIES-Len	PAAV
Escolaridad de la madre	.05	.15	.23	.34
Escolaridad del padre	.11	.16	.26	.35

Discusión

Este estudio presenta evidencia acerca del empleo de preguntas de ensayo en la batería de mediciones de selección universitaria. Desde el punto de vista conceptual, no hay duda que este tipo de preguntas involucran habilidades y capacidades relevantes en el contexto académico, por lo que su eventual incorporación debiera enriquecer la información que se obtiene a través de las pruebas de admisión tradicionalmente empleadas en Chile.

Si hay ventajas conceptuales tan evidentes, cabe preguntarse por qué no se están usando en Chile este tipo de preguntas. La respuesta que probablemente se daría es que su uso conlleva problemas prácticos mayores (tiempo de corrección) y riesgos técnicos importantes (subjetividad de los puntajes). El estudio reportado muestra que estas potenciales dificultades pueden ser superadas. Bajo condiciones controladas, y empleando procedimientos ya establecidos a nivel internacional para la corrección de este tipo de preguntas, ha quedado demostrado que tanto con rúbricas analíticas como con rúbricas holistas, es posible lograr altos niveles de acuerdo y consistencia entre distintos jueces, por lo que el riesgo de subjetividad puede ser controlado. Por otra parte, la dificultad práctica asociada al tiempo que demanda la corrección de este tipo de preguntas no es un impedimento relevante, puesto que el ejercicio práctico realizado muestra que la corrección de una pregunta de este tipo en la escala que se requeriría para usarla en pruebas nacionales de admisión, puede ser llevada a cabo en un plazo inferior a las 6 semanas. Naturalmente, la solución de los dos problemas recién expuestos supone necesariamente al menos dos condiciones: el desarrollo de buenas rúbricas de corrección y el entrenamiento riguroso de los correctores.

Aunque las dos rúbricas mostraron buenos índices de consistencia y acuerdo, la holista parece ser preferible para su uso en pruebas de selección, puesto que sus índices técnicos son al menos tan altos como los de la rúbrica analítica y los tiempos que demanda su entrenamiento y aplicación son considerablemente menores.

Adicionalmente, cabe mencionar que cuando se trabaja con jueces en aplicaciones reales, se incorporan procedimientos regulares para monitorear y optimizar el acuerdo y consistencia entre jueces, que incluyen su reentrenamiento y recalibración cada vez que se detectan discrepancias. Por ello, cabe esperar que en tales condiciones los niveles de consistencia y acuerdo sean aún mayores que los aquí reportados.

El estudio muestra que las preguntas de ensayo poseen correlaciones sustantivas con los factores de admisión tradicionales, pero con clara evidencia que aportan información diferente de la que hoy se obtiene con pruebas basadas en preguntas de selección múltiple. Por lo tanto, hay evidencia inicial acerca del aporte incremental de este tipo de preguntas.

Finalmente, los resultados obtenidos en este estudio son positivos en la perspectiva de los problemas de equidad asociados a las pruebas de admisión universitaria. Por una parte, en la pregunta de ensayo las mujeres logran rendimientos sistemáticamente superiores a los de los hombres, lo que al menos parcialmente puede compensar la ventaja que tradicionalmente estos últimos han tenido en todas las pruebas de admisión. Por otra parte, el rendimiento en este tipo de preguntas parece estar menos influido por el capital sociocultural familiar, en comparación con lo que es usual observar con las pruebas tradicionales.

En suma, este estudio entrega evidencia sustantiva que respalda la factibilidad de incorporar preguntas de ensayo a la batería de selección universitaria en Chile. Dado que su empleo es costoso e involucra tiempos de corrección mayores que los que demandan las preguntas de respuesta cerrada, se sugiere que al menos en una etapa inicial se contemple la inclusión de una pregunta de este tipo en las pruebas de selección a la enseñanza superior.

Referencias

Berríos, R., Claro, F., Cox, C., Donoso, G., Flores, R., Himmel, E., Lorca, C., Manzi, J., Passalacqua, A., Quadri, S., Riquelme, S., Rodríguez, C. & Vaisman, L. (2000). *Comisión nuevo currículum de la EM y pruebas del sistema de admisión a la educación superior*. Informe sometido en consulta previa a la Ministra de Educación. Aprobado por el Consejo de Rectores de las Universidades Chilenas. Santiago.

Bravo, D. & Manzi, J. (2002). *El SIES, la equidad y la elevación de los aprendizajes* [En red]. Disponible en: http://www.sies.cl

Clauser, B. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24*(4), 310-324.

College Board. (2002). *El College Board presenta un nuevo SAT* [En red]. Disponible en: http://www.collegeboard.com/about/newsat/press-espanol.html

Congdon, P. & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

Doolittle, A. & Welch, K. (1989). *Gender differences in performance on a college-level achievement test.* Iowa, IA: American College Testing Program.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 49,* 294-303.

Goodrich, H. (1997). Understanding rubrics. *Educational Leadership, 54*(4), 14-18.

Hart, D. (1994). *Authentic assessment: A handbook for educators*. New York, NY: Addison-Wesley Publishing Company.

Herman, J. L., Aschbacher, P. R. & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Lawlis, G. & Lu, E. (1972). Judgment of counseling process: Reliability, agreement and error. *Psychological Bulletin, 78,* 17-20.

Means, M. & Voss, J. (1996). *Who reasons well? Two studies of informal reasoning among children of different grade, ability and knowledge levels*. Pittsburgh, PA: Lawrence Erlbaum.

Ministerio de Educación. (1988). *Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media.* Santiago: Autor.

Ministerio de Educación. (2002). *Prueba SIMCE 2° medio 2001: Factores que inciden en el rendimiento de los alumnos*. Nota Técnica, Departamento de Estudios y Estadística. Santiago: Autor.

Mizala, A. & Romaguera P. (2000). *Determinación de factores explicativos de los resultados escolares en educación media en Chile.* Documento de Trabajo 85, Centro de Economía Aplicada, DII, Universidad de Chile.

Perkins, D. (1985). Post-primary education has little impact upon informal reasoning. *Journal of Educational Psychology, 77,* 562-571.

Resnick, L. (1987). *Education and learning to think.* Washington, DC: National Academy Press.

Rosas, R., Flotts, M. & Saragoni, C. (2002). Modelo de representación del conocimiento para las nuevas pruebas de selección para el ingreso a las universidades chilenas. *Psykhe, 11*(1), 3-14.

Shaw, V. (1996). The cognitive process in informal reasoning. *Thinking and Reasoning, 2,* 51-80.

Stumpft, H. (1998). Stability and change in gender-related differences on the College Board placement and achievement tests. *Current Directions in Psychological Science, 7,* 192-196.

Tinsley, H. & Weiss, D. (2000). Interrater reliability and agreement. En H. Tinsley & D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 96-124). San Diego, CA: Academic Press.

Tombari, M. & Borich, G. (1999). *Authentic assessment in the classroom.* New Jersey, NJ: Prentice-Hall.

Van Eemeren, F., Grootendorst, R. & Snoeck, F. (1996). *Fundamentals of argumentation theory.* Mahwah, NJ: Lawrence Erlbaum.

Wolfe, E., Kao, C. & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scores. *Written Communication, 15*(4), 465-492.

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. New York, NY: Routledge Falmer.