

## **La confiabilidad de las evaluaciones del aprendizaje conceptual: Índice Spearman-Brown del método split-halves (Reliability of the evaluation of conceptual learning: index of Spearman-Brown and the split-halves method)**

**Rositas, J., M. H. Badii & J. Castillo**

UANL, San Nicolás de los Garza, N.L., México, [irositasm@yahoo.com.mx](mailto:irositasm@yahoo.com.mx)

**Key words:** Conceptual learning, index of Spearman-Brown, reliability index, Split-halves

**Abstract.** Based upon the curiosity to evaluate in a reliable and consistent fashion, the acquisition of knowledge, several methods of evaluating conceptual learning are presented, and among these, the method of Index of Spearman-Brown within the split-halves method is used for the estimation of the reliability of the measurement instrument. The related problems in this respect are laid out, taking into account the ease of determination of this index in a practical manner with the aid of the concrete applied cases. Final considerations are raised with regards to application of this methodology in the USA and the New Zealand. Finally, the utilization of these methods are encouraged.

**Palabras claves:** Aprendizaje conceptual, Índice de confiabilidad, Índice de Spearman-Brown, Split-halves

**Resumen.** En concordancia con la inquietud de evaluar siempre en la forma más completa, confiable y consistente posible la adquisición de conocimientos, se presentan varios métodos de evaluar el aprendizaje conceptual y explican y justifican la selección del Índice Spearman-Brown en el método split-halves que nos permite medir la confiabilidad de los instrumentos de medición (exámenes) en este tipo de aprendizaje. Se expone la problemática y la conveniencia de determinar este índice y se indica la forma de poner en práctica el método de evaluación seleccionada, mostrando casos concretos de aplicación. Se presentan resultados obtenidos en varias materias impartidas. Se incluyen consideraciones finales sobre la aplicación de estas metodologías en EUA y Nueva Zelanda y se invita a la utilización de estos enfoques.

### **Introducción**

La mayoría de los cursos impartidos son bajo el ue constructivista del aprendizaje significativo. Lo ideal en cualquiera de cualquier curso es que los estudiantes presenten por escrito un proyecto terminado en el que demuestren

que han puesto en práctica exitosamente los conocimientos adquiridos. La calificación obtenida en este proyecto generalmente tiene una ponderación importante.

El éxito del estudiante en el transcurso del curso es ir desarrollando y aplicando habilidades del pensamiento o competencias que bajo la guía del maestro va obteniendo durante el semestre. Estas habilidades de pensar consisten en estructurar y desarrollar en forma congruente y clara los distintos apartados de un protocolo de investigación o de solución a un determinado problema. Pero como reza acertadamente un aforismo del Dr. Edward N. Hay: “*se piensa con lo que se sabe*” (Bellak, 1989), la calidad del protocolo desarrollado estará condicionada por los conocimientos adquiridos.

En su exitoso método para evaluar empleados y ejecutivos Hay declara que prácticamente cualquier pensamiento para la solución de problemas se basa en el conocimiento de principios, hechos y medios. Por lo tanto, es necesario que el estudiante sepa y distinga claramente y con anterioridad los principios, teorías y procedimientos (primeramente en una forma conceptual) en que se basa una determinada materia. Este argumento se identifica con el aprendizaje conceptual que algunos autores lo han definido como la “incorporación de datos, conceptos y principios a la estructura mental de comprensión” que “permite describir, entender, explicar, fundamentar y proyectar la acción”.

El estudiante aprueba o *acredita* la materia si el maestro *crea* razonablemente que el estudiante conoce o domina al menos el 70% u 80% de la materia cursada según sea de licenciatura o de postgrado. Las evaluaciones que hemos aplicado y que son la base para acreditar o no el curso, tienen dos partes, una parte conceptual y una parte aplicada. Nos enfocaremos en el presente artículo a la parte conceptual.

## **Problemática**

Un concepto es una idea compartida. Sobre esta idea, según cada campo de estudio, es posible e incluso necesario llegar a tener una definición común, aunque de hecho tenga distintos matices de significado según los antecedentes y experiencias previas del estudiante. Para la adquisición y distinción clara de estos conceptos, cada grupo e incluso cada persona recurre a diversas estrategias, pero para ir verificando que se han ido aprendiendo en forma compartida estos conceptos, es necesario ir acordando y demostrando qué es, y qué no es, a lo que se refiere cada uno de ellos.

La importancia de tener claros los conceptos predomina no solamente en el campo educativo, si no en las organizaciones que tienen enfoque de calidad y

mejora continua, en las que se llega a afirmar que no se puede mejorar lo que no se puede medir y que no se puede medir lo que no se ha definido claramente.

## Objetivos

Por los antecedentes y problemática arriba expuesta, los objetivos principales del presente artículo son:

- a. Presentar varios métodos de evaluar el aprendizaje conceptual y seleccionar un método que nos permite medir la confiabilidad de los instrumentos de medición (exámenes) del aprendizaje de conceptos.
- b. Indicar la manera de poner en práctica el método evaluación seleccionada.
- c. Mostrar casos concretos en los que se han aplicado estas metodologías.
- d. Un propósito adicional, es promover e invitar a la aplicación de estos enfoques y metodologías de evaluación.

## Justificación del instrumento de evaluación

Una manera de verificar esto, aunque no muy práctica, sería que el estudiante mediante exámenes de respuesta abierta expresara con sus *propias palabras* cada concepto importante de la materia estudiada y que mostrara mediante mapas conceptuales cómo se interrelacionan estos conceptos.

Así por ejemplo, para poder saber si un estudiante ha llegado a comprender lo que es *confiabilidad* y *validez de contenido* de un instrumento de medición, esperaríamos que su definición de confiabilidad se acercara lo más posible, si se ha estado llevando el texto de Hernández-Sampieri et al. (2003), a la definición de estos autores que afirman que es “*el grado en que la aplicación repetida del instrumento de medición al mismo sujeto u objeto produce iguales resultados;*” mientras por definición de *validez de contenido*, el estudiante hiciera referencia al “*grado en que el instrumento representa el concepto medido*” (Bohrstedt, 1976), o el grado en que un instrumento “*refleja un dominio específico de lo que se mide.*” No se descarta, e incluso se invita al estudiante a que aporte durante las sesiones definiciones o puntos de vista adicionales que complementen, enriquezcan o critiquen las definiciones del texto básico.

Si el estudiante escribe que confiabilidad es *el grado de objetividad que tiene el instrumento* mientras que *validez* significa que *lo que se ha medido es relevante*, sus definiciones tienen algo de cierto, pero ¿son completas y correctas? ¿Qué calificación le pondría el maestro en base 100 a cada pregunta? ¿50, 70, 85? Realmente, no habría un criterio consistente para saber cuánto. El

problema de acreditación del aprendizaje de conceptos, se complicaría tremendamente, si quisiéramos evaluar decenas de conceptos, en decenas de estudiantes, para que el examen tuviera una buena validez. Imaginemos simplemente que tenemos 30 estudiantes y se requieren evaluar en un examen parcial 50 conceptos. Habría que revisar 1,500 “definiciones personales”. Bajo este enfoque, el maestro termina conformándose con aplicar un examen de 10 ó 15 preguntas, perdiendo de esta forma validez de contenido y dificultándose el dar puntuaciones consistentes a cada pregunta y no siendo posible estadísticamente estimarle la confiabilidad a este tipo de exámenes.

La evaluación de la estructuración de ideas y la expresión escrita del estudiante es muy importante, pero debemos dejar en una segunda parte del examen o en la revisión de los avances del proyecto de investigación y en el reporte final.

Una manera de registrar la comprensión de varias docenas de conceptos considerados importantes en la primera parte de un examen parcial, digamos 50 conceptos en máximo 2 horas, es aplicar exámenes estandarizados basados en definiciones ostensibles o mostrables, o simple identificación del concepto con que se conoce tal definición. En contraste con los exámenes de respuesta abierta, en los exámenes conceptuales estandarizados con este enfoque, lo que se le proporciona al estudiante es una declaración en la que se sitúa el contexto de lo que se le preguntará y luego la definición de uno o varios conceptos por párrafo, pidiéndosele que identifique el nombre o la palabra correcta con que se identifica cada concepto de una lista de palabras ordenadas alfabéticamente. Si se trata de un ejemplo o frase en la que se aplica el concepto se le dan varias opciones entre las que seleccionará la respuesta correcta. Una forma de evaluar la comprensión, con el enfoque de definición ostensible, es ir presentando en el examen párrafos del siguiente tipo:

En cualquier investigación es importante que todo instrumento cumpla con ciertas características. Una de estas características es el grado en que un instrumento representa el concepto medido y es la característica conocida ¿con qué nombre? → [1: \_\_\_\_], mientras que el grado en que la aplicación repetida del instrumento de medición al mismo sujeto u objeto produce iguales resultados, es la característica conocida ¿con qué nombre? → [2: \_\_\_\_].

Con este enfoque, el autor ha llegado a integrar en dos o tres páginas 40 ó 60 ítems para evaluar la comprensión de los principales conceptos estudiados en la mitad de un curso de metodología en el que el texto básico es el libro citado de Hernández-Sampieri et al. (Cap. 1 al 8 en 5 sesiones de 3 horas efectivas). Un ítem es cada elemento mínimo o reactivo de una medición que estimula una respuesta del sujeto evaluado. Un ítem puede ser una pregunta, una frase que hay que completar, una figura o cifra a interpretar, etc.

Bajo este enfoque ha sido posible obtener altos coeficientes de confiabilidad (entre 0.91 y 0.95).

### **Aspectos contextuales y metodológicos**

Al evaluar los exámenes desarrollados y aplicados a los alumnos con el enfoque ostensible, se ha encontrado que cumplen con las características de validez y confiabilidad.

Hernández-Sampieri apoyándose en Wiersma (1999) y Gronlund (1990) hace referencia a tres tipos de evidencia de la validez: de contenido, de criterio, y de constructo, subdividiendo a su vez la validez de criterio en validez concurrente y predictiva. Aunque todos estos tipos de validez son importantes, la básica es la validez de contenido que se cumple al incluir en el examen preguntas en cantidades proporcionales a los contenidos conceptuales de cada capítulo.

En los cursos de metodología de investigación, como se mencionó, se estudian los primeros ocho capítulos de Hernández-Sampieri et al. (2003) en cinco sesiones de tres horas efectivas (la primera es introductoria y la séptima para repaso y examen parcial), dejando la segunda parte del curso (6 sesiones) para los últimos tres capítulos y para ejercicios de análisis de resultados y revisión final del proyecto de investigación. En cuanto a contenido del primer examen parcial se busca que contenga alrededor de 7–8 ítems por capítulo (variando este promedio según la extensión del capítulo) o sea alrededor de 60 ítems en total. Al tomar este tipo de exámenes, el estudiante aprende por vivencia personal lo que significa la validez de contenido.

En cuanto a los métodos de confiabilidad o grado en que la aplicación repetida del instrumento de medición al mismo sujeto u objeto produce iguales resultados, cualquier procedimiento debe arrojar en el cálculo de tal grado valores dentro del rango de 0 a 1.

Hernández-Sampieri et al. (2003) citan cinco procedimientos posibles para el cálculo de la confiabilidad. Los primeros dos: *test-retest* y *formas alternativas* requieren cuando menos dos administraciones del instrumento (examen o prueba) al mismo grupo de personas e imponen una serie de requisitos difíciles de cumplir. Los otros tres procedimientos: *split-halves* (mitades partidas o partición binaria), *coeficiente alfa de Cronbach* y *coeficiente de Kuder-Richardson*, requieren una sola aplicación. De estos tres procedimientos el que seleccionamos por ser el más amigable en su elaboración y por tener una interpretación más comprensible es el primero o sea el *split-halves* o partición binaria. Una limitación adicional del método de Cronbach-alfa es que los reactivos o elementos de las escala deben ser de tipo intervalo (Davis, 2000).

## Evaluación práctica de la confiabilidad

En el presente apartado se expone el procedimiento de cálculo del índice de confiabilidad seleccionado (Spearman-Brown) con un ejemplo concreto, se hace una interpretación de resultados y se presenta un par de toma de decisiones a partir de los resultados del índice de confiabilidad, error estándar de medición y pruebas de hipótesis.

### Procedimiento de cálculo

El método de split-halves o partición binaria es un método muy práctico que se lleva a cabo desarrollando los siguientes pasos.

Crear una tabla en Excel (Tabla A del Apéndice) en la que cada línea representa una pregunta del examen y cada columna representa los resultados de un estudiante. En el cuerpo de esta tabla de datos-base se codifican con "0" (ceros) y "1" (unos), las fallas y aciertos obtenidos respectivamente al responder los estudiantes cada ítem.

Agrupar los resultados de los exámenes (que han sido aplicados a cada estudiante una sola vez) en dos grupos. Un grupo con los resultados de las preguntas o ítems impares o nones y otro con los resultados a las preguntas pares. Esto se logra fácilmente anteponiendo en la tabla de resultados (Tabla B del Apéndice) una "N" a las preguntas nones o impares, o una "P" a las preguntas pares y usando el Menú de Excel → Datos, → Sub-Menú = → Ordenar. Los dos grupos de preguntas equivalen a la aplicación subsecuente de dos exámenes paralelos o sobre temas similares (mismos capítulos). De acuerdo a Anastasi y Urbina (1998), la separación en grupos de ítems nones y pares producen calificaciones equivalentes para ambas mitades al estar dispuestos en el mismo orden de dificultad.

En la última columna de la tabla puede revisarse el grado de dificultad de cada pregunta y tomar la decisión apropiada (eliminarla, explicarla mejor en el curso, o mejorar redacción).

Sumar en la misma Tabla B del Apéndice para cada estudiante el número de aciertos en las preguntas nones y el número de aciertos en las preguntas pares. Copiar estas líneas de sumas como columnas, recurriendo a Menú → Edición → Copiar → Pegado especial → Transponer, obteniendo así la Tabla 1.

Correlacionar las puntuaciones que los estudiantes obtuvieron en la preguntas nones versus las pares (Tabla 1 y Figura 1).

El instrumento es más confiable en la medida que haya mayor correlación entre las puntuaciones de ambas mitades. Esto es, cuando existe

una considerable correlación en el diagrama de dispersión (Figura 1) se observa que los estudiantes con bajas puntuaciones en las preguntas nones tienden a mostrar también puntuaciones bajas en las preguntas pares, y lo mismo para los estudiantes con altas puntuaciones.

Tabla 1. Resultados de los exámenes.				
Alum	Nones	Pares	Totales	Calif.
1	23	25	48	96
2	22	23	45	90
3	24	24	48	96
4	21	22	43	86
5	25	23	48	96
6	25	24	49	98
7	21	18	39	78
8	24	25	49	98
9	25	24	49	98
10	21	20	41	82
11	24	22	46	92
12	25	23	48	96
13	20	21	41	82
14	22	23	45	90
15	17	18	35	70
16	22	21	43	86
17	17	13	30	60
18	22	20	42	84
19	24	23	47	94
20	25	23	48	96
21	25	23	48	96
22	24	24	48	96
23	25	25	50	100
	Mín.		60	
	Máx.		60	
	Prom.		89.6	
	S=Desv		10	

Fuente: Tabla 1 de Datos Básicos del Apéndice

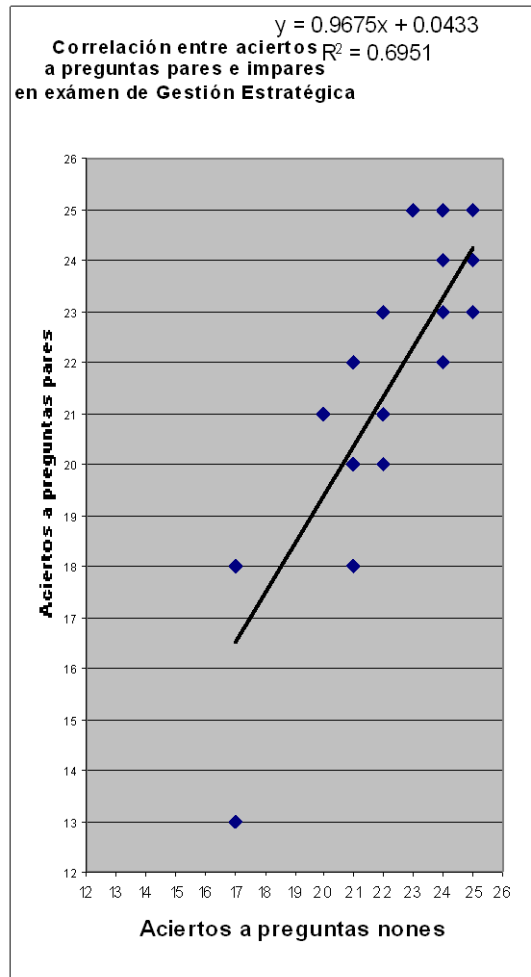


Figura 1. Correlación entre las variables.

A partir del valor del coeficiente de correlación “ $r$ ” (o raíz del valor de  $R^2$  que aparece en Figura 1) Thomas & Young (1995) en su obra sobre estadística educacional sugieren que la estimación numérica del grado de confiabilidad esperado ( $r_e$ ) se haga mediante la fórmula de Spearman-Brown:  $r_e = 2*r/(1 + r)$ .

Dado que  $R^2 = 0.6951$ , el valor correspondiente de  $r = 0.8337$  y por lo tanto  $r_e$  esperado resulta:  $r_e = 2 * 0.8337 / (1 + 0.8337) = 0.91$ .

### Interpretación de resultados

De acuerdo a Nunnaly y Berstein (1994), los valores mínimos recomendados de los coeficientes de confiabilidad son de 0.7 para investigaciones exploratorias, 0.80 para investigación básica y 0.90 para toma de decisiones, por lo que el nivel obtenido de 0.91 en el presente examen se considera adecuado para el uso que se le dará más adelante.

¿Cómo pudiéramos interpretar el valor  $r_e = 0.91$ ?

Una manera de interpretarlo es la siguiente. Si fuera posible aplicar a este grupo de estudiantes, cien veces el mismo examen, la tabla de posiciones (tabla jerarquizada por calificaciones de mayor a menor) saldría en el mismo orden en 91 ocasiones y solo en las nueve ocasiones restantes el orden diferiría. Una de las razones por las que el índice de confiabilidad pueda resultar bajo (inferior a .90) puede ser debida a que en la redacción de algunas preguntas haya ambigüedad, por lo que la confusión al contestar equivaldría, en el caso de que se aplicara dos veces consecutivas el examen, a que en la primera ocasión pondrían unas respuestas y en la segunda ocasión pondrían respuestas diferentes. Otra razón pudiera ser debida a un número bajo de preguntas o ítems. (Anastasi & Urbina, 1994).

Por lo tanto, un alto índice de confiabilidad nos puede dar la tranquilidad de que la redacción del examen es clara, prácticamente, sin ambigüedades, por lo que si un estudiante obtiene baja calificación lo más probable es que su conocimiento de la materia sea bajo y no debido a que el examen “*venía muy revuelto*.” El índice de confiabilidad Spearman-Brown además de reportarnos la calidad del examen como instrumento de evaluación, también nos sirve para tomar decisiones.

### Toma de decisiones basadas en el índice

Si el examen fuera perfectamente confiable, esto es si la correlación  $r$  fuera igual a 1 y consecuentemente, el coeficiente de confiabilidad  $r_e$  sería igual también 1; esto es, la calificación del examen en cualquiera ocasión que se aplicara sería la calificación verdadera sin presentarse ningún error de medición, lo cual es prácticamente imposible.

Cuando el índice de confiabilidad es menor a 0.90, una decisión práctica es que hay que mejorar la redacción del examen y/o aumentar el número de ítems.



Debido a que la confiabilidad casi siempre resulta menor a 1, pudiera tener sentido el argumento de un estudiante que afirma: “*aunque saqué 70, y el pase es 80, si presento de nuevo un examen equivalente seguramente que pasaré*”. Para tomar una decisión relacionado con esta situación, como por ejemplo, la del estudiante con número de lista 15 que obtuvo 70, es necesario diseñar y aplicar una regla práctica que tome en cuenta el error de medición.

Ya que no es posible que un examen sea perfectamente confiable, los especialistas en psicometría y estadística educacional (Thomas & Young, 1995; Anastasi & Urbina, 1998) han concluido que el error estándar de medición ( $S_m$ ) de una distribución teórica de calificaciones (incluyen errores de medición) es igual a:  $S_m = S \cdot \sqrt{1 - r_e}$ . En esta fórmula  $S$  es la desviación estándar de las calificaciones, y como se observa otro uso de  $r_e$  es en la estimación de error estándar de medición. Tomando  $S$  de tabla 1, el valor  $S_m$  resulta de ser;  $S_m = 10 \cdot \sqrt{1 - 0.91}$ , y por tanto,  $S_m = 3$ .

La estimación de este error es importante porque en ocasiones el resultado de un examen estandarizado es utilizado para tomar una decisión respecto a si un solicitante a un empleo será seleccionado como candidato a un puesto o rechazado o si un estudiante en un curso de opción a título egresará o no de una carrera en ese semestre. En muchas decisiones el resultado del examen estandarizado es uno de varios elementos para tomar la decisión.

Aquí podríamos aplicar y decidir en base a una prueba de hipótesis.

### Prueba de hipótesis

Siguiendo el razonamiento de Thomas & Young (1995), para situar la base de la teoría estadística de esta prueba, ellos afirman que si un estudiante pudiera presentar el examen un número ilimitado de veces sin que se viera afectado el individuo por ello, el promedio de todas sus calificaciones en este examen sería la verdadera calificación, representándose tal valor mediante  $\mu$ . Por lo tanto, en el caso del estudiante que nos ocupa:

$$H_0: \text{Calificación} = \mu = 80$$

$$H_1: \text{Calificación} = \mu < 80$$

$$\alpha = 0.05 \rightarrow, \text{consecuentemente, } Z_{\text{tablas}} = -1.64$$

$$Z_{\text{calculada}} = (X - \mu) / S_m = (70 - 80) / 3 = -3.3,$$

Por lo que se rechaza la  $H_0$ , y lo más probable es que la calificación verdadera sea menor a 80.

Otra forma de ver esto mismo, es calculando el error de estimación que, con un 95% de confianza es igual a  $1.96 \cdot S_m = 5.88$ , por lo que la calificación verdadera bien pudiera estar en la franja de  $70 \pm 5.88$ , que no incluye el 80.

Por lo tanto, aun tomando en cuenta un probable error de medición el estudiante que obtiene 70 no debe acreditar.

Claro que no hay que darles tantas explicaciones a los estudiantes, pero como regla práctica se les puede prometer, *por aquello de que alguien haya tenido alguna dificultad especial con algunas preguntas*, hasta 6 ó 5 puntos extra según qué tanto participen en el curso. Recuérdese que la verdadera calificación como puede estar 5 ó 6 puntos hacia arriba, también pudiera estar 5 ó 6 puntos hacia abajo.

### **Consideraciones finales**

La justificación, de acuerdo a experiencias de tomar en cuenta el error de medición, aún con índices de confiabilidad buenos, es cuestionar el criterio de comparación o calificación mínima 80 para acreditar.

En términos prácticos, el atacar el estándar del 80, equivaldría al argumento del estudiante que afirma a que si en las escuelas primarias dejara de ser un requisito la calificación de 60 para “aprobar,” en las secundarias y niveles pre-profesionales debe dejar de ser el 70 el criterio para aprobar, y por lo tanto, en los postgrados debe dejar de ser el 80 la calificación mínima.

La exposición de motivos de los diseñadores de este sistema de estándares mínimos queda fuera del alcance de este trabajo y pudiera ser motivo de otro artículo comparando el enfoque utilizado en el sistema de evaluación actual de “Criterion Referenced Approach” con el enfoque de “Norm Referenced Testing”.

La cuestión es que si a medida que se avanza en los niveles educativos (primaria, secundaria, profesional, etc.) también se avanza en los niveles de la taxonomía de Bloom, suena razonable que los niveles de exigencia y de cumplimiento también se avance.

Un ejemplo del interés por evaluar la confiabilidad y la validez del aprendizaje a nivel de un país utilizando el Índice de Spearman-Brown, lo encontramos en un estudio de certificación de logros educacionales del Ministerio de Educación de Nueva Zelanda realizado por la Universidad de Waitako (2006) sobre que pueden revisar las personas interesadas en profundizar en estas aplicaciones.

### **Conclusiones**

En concordancia con los objetivos del presente artículo, se ha expuesto la conveniencia y la problemática de determinar índices de la confiabilidad de los instrumentos de medición (exámenes) en la parte relativa del aprendizaje

conceptual, explicando y compartiendo las experiencias en la obtención y en la utilización práctica de tales índices.

El ejemplo siguiente demuestra el caso (Tabla 2).

Tabla 2. Índices de Confiabilidad y Errores estándar de medición.

<b>Materia</b>	<b>No. reactivos</b>	<b>Nn. estudiantes</b>	<b>Índice</b>	<b>Error estándar</b>
<b>A</b>	84	29	0.95	3.5
<b>B</b>	62	30	0.95	3.9
<b>C</b>	50	23	0.91	3.0
<b>D</b>	50	13	0.94	4.6

La aplicación de estos enfoques (instrumentos e índices), incluso dosificados durante los cursos mediante exámenes breves con enfoque similar a los exámenes parciales, dan tranquilidad al maestro en cuanto a la dirección del curso y en cuanto a consistencia y confiabilidad en la equidad y al alumno le imponen retos que lo estimulan a aplicarse más de lleno en sus responsabilidades de aprender por lo que se promueve a aprovechar y utilizar esta metodología.

### Referencias

- Anastasi, A. & S. Urbina. 1998. Test psicológicos. 7ª edición. Prentice-Hall. México.
- Bellak, A., O. 1989. Sistemas específicos de evaluación de puestos: el método Hay de perfiles y escalas guías. En Rock, Milton L. Manual de administración de sueldos y salarios. McGraw-Hill, Mexico.
- Gronlund, N. E. 1990. Measurement an evaluation in teaching (6th ed.). Macmillan. N. Y.
- Hernández-Sampieri, R. H., C. Fernández-Collado & P. Baptista-Lucio. 2003. Metodología de la Investigación. McGraw-Hill. México.
- Thomas, F. L. & J. I. Young. 1995. An Introduction to Educational Statistics: The Essential elements, 3rd ed. Simon and Schuster Custom. Needham heights, Massachusetts.
- Wiersma, W. 1999. Research Methods in Education: An Introduction. 7th ed. Allyn & Bacon, Boston.
- Nunnaly, J. C. & I. H. Berstein. 1994. Psychometric Theory. McGraw-Hill. N. Y.
- Waikato University. 2006. National Certificate of Educational Achievement: Issues related to reliability, validity and manageability. Disponible en Internet en el Enero de 2006 en <http://www.so.e.waikato.ac.nz/certstudies/English/CSEngandNCEA/NZARE2000.html>.

Apéndice: Tabla A. Datos-Base: Resultados de un curso de alumnos a nivel de postgrado.

**Tabla Básica: Calificaciones de la parte conceptual de un curso de Gestión Estratégica (Revisión de 8 capítulos de Bateman (2004).**

Núm. preg.	Núm. de lista del alumno ( 0=incorrecto, 1=correcto)																							Acie rto s	Porc.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23		
1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	21	0.91
2	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	21	0.91
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	22	0.96
4	1	1	1	1	1	1	0	1	1	0	1	1	1	1	0	1	1	1	1	1	1	0	1	19	0.83
5	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	19	0.83
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	22	0.96
7	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0	1	1	0	1	1	1	1	1	18	0.78
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
9	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	21	0.91
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
11	0	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	0	1	1	1	1	1	1	18	0.78
12	1	0	0	0	1	1	1	1	1	0	1	1	0	1	1	0	0	0	1	1	0	1	1	14	0.61
13	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	18	0.78
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	22	0.96
15	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	20	0.87
16	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	8	0.35
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	22	0.96
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	22	0.96
21	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	20	0.87
22	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	20	0.87
23	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	21	0.91
24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
25	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	19	0.83
26	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	21	0.91
27	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	21	0.91
28	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	20	0.87
29	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	22	0.96
32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	21	0.91
33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
34	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	22	0.96
35	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	22	0.96
36	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
37	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	22	0.96
38	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	22	0.96
39	1	1	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	20	0.87
40	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	0	1	1	1	1	1	19	0.83
41	1	1	1	1	1	1	0	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	19	0.83
42	1	1	1	1	0	1	0	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	18	0.78
43	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	22	0.96
44	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	22	0.96
45	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	21	0.91
46	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	21	0.91
47	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	22	0.96
48	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	1	1	1	1	19	0.83
49	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
50	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	0	1	1	0	1	1	1	1	17	0.74
<b>Total</b>	<b>48</b>	<b>45</b>	<b>48</b>	<b>43</b>	<b>48</b>	<b>49</b>	<b>39</b>	<b>49</b>	<b>49</b>	<b>41</b>	<b>46</b>	<b>48</b>	<b>41</b>	<b>45</b>	<b>35</b>	<b>43</b>	<b>30</b>	<b>42</b>	<b>47</b>	<b>48</b>	<b>48</b>	<b>48</b>	<b>50</b>	Calif. Prom .	
<b>Calif</b>	<b>96</b>	<b>90</b>	<b>96</b>	<b>86</b>	<b>96</b>	<b>98</b>	<b>78</b>	<b>98</b>	<b>98</b>	<b>82</b>	<b>92</b>	<b>96</b>	<b>82</b>	<b>90</b>	<b>70</b>	<b>86</b>	<b>60</b>	<b>84</b>	<b>94</b>	<b>96</b>	<b>96</b>	<b>96</b>	<b>100</b>	<b>89.6</b>	

Tabla B : Datos de la Tabla "A" clasificados en dos grupos de preguntas nones y pares.

		CALIFICACIONES DE UN CURSO DE GESTION ESTRATEGICA																								
		Núm. de lista del alumno ( 0=incorrecto, 1=correcto)																								
Non ó Impar	Núm. preg.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Aciertos	Porc.
N	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	21	0.91
N	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	22	0.96
N	5	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	19	0.83
N	7	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0	1	1	0	1	1	1	1	1	18	0.78
N	9	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	21	0.91
N	11	0	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	0	1	1	1	1	1	1	18	0.78
N	13	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	18	0.78
N	15	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	20	0.87
N	17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
N	19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
N	21	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	20	0.87
N	23	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	21	0.91
N	25	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	19	0.83
N	27	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	21	0.91
N	29	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
N	31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	22	0.96
N	33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
N	35	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	22	0.96
N	37	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	22	0.96
N	39	1	1	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	20	0.87
N	41	1	1	1	1	1	1	0	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	1	19	0.83
N	43	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	22	0.96
N	45	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	21	0.91
N	47	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	22	0.96
N	49	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
Suma de nones		23	22	24	21	25	25	21	24	25	21	24	25	20	22	17	22	17	22	24	25	25	24	25		0.91
P	2	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	21	0.91
P	4	1	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	19	0.83
P	6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	22	0.96
P	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
P	10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
P	12	1	0	0	0	1	1	1	1	0	1	1	0	1	1	0	0	0	1	1	0	1	1	1	14	0.61
P	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	22	0.96
P	16	1	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	8	0.35
P	18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	22	0.96
P	20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	22	0.96
P	22	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	20	0.87
P	24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
P	26	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	21	0.91
P	28	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	20	0.87
P	30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
P	32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	21	0.91
P	34	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	22	0.96
P	36	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	1.00
P	38	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	22	0.96
P	40	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	0	1	1	1	1	1	1	19	0.83
P	42	1	1	1	1	0	1	0	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	18	0.78
P	44	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	22	0.96
P	46	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	21	0.91
P	48	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	19	0.83
P	50	1	1	1	0	1	1	1	1	1	1	1	0	1	1	0	0	0	1	1	0	1	1	1	17	0.74
Suma pares		25	23	24	22	23	24	18	25	24	20	22	23	21	23	18	21	13	20	23	23	23	24	25		0.88
<b>SUMA TOTAL</b>		48	45	48	43	48	49	39	49	49	41	46	48	41	45	35	43	30	42	47	48	48	48	50		
<b>Calificación</b>		96	90	96	86	96	98	78	98	98	82	92	96	82	90	70	86	60	84	94	96	96	96	100	Med=	89.6