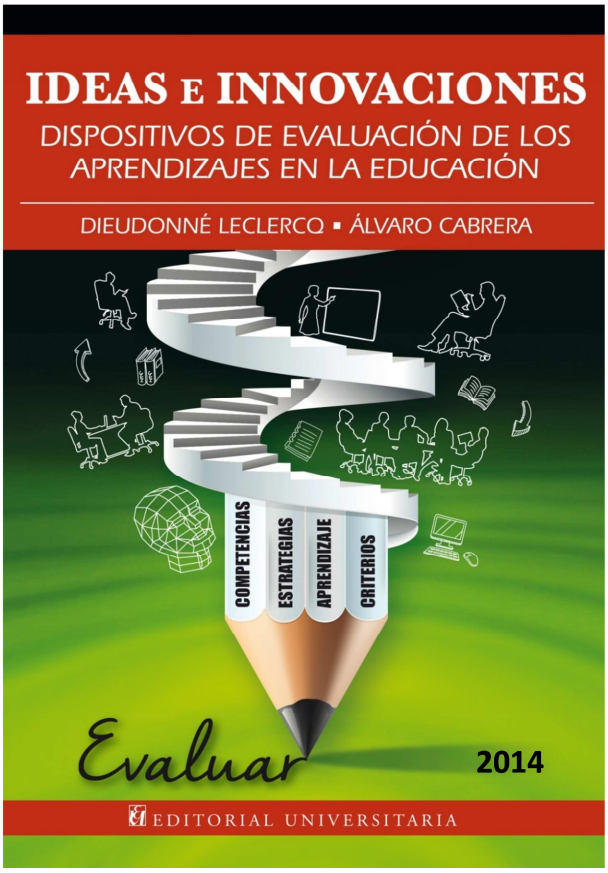


Dieudo LECLERCQ

Álvaro CABRERA MARAY

Universidad de Chile

www.universitaria.cl

Directores de la publicación:

Dieudonné Leclercq
Universidad de Liège (ULg)

Álvaro Cabrera Maray
Universidad de Chile (UCH)

IDEAS e INNOVACIONES
Innovaciones en Dispositivos de Evaluación de los Aprendizajes en la enseñanza Superior
2014

Se pueden bajar gratuitamente
desde http://orbi.uliege.be, después Leclercq D., o
desde www.evaluaraprendizajes.cl
-Los **resúmenes** de los 23 capítulos
del libro IDEAS http://hdl.handle.net/2268/173543
-El **índex** de este libro para buscar entre
entradas de 1500 conceptos y
400 de autores http://hdl.handle.net/2268/180060

Dieudonné Leclercq

Dr. en Educación (1975) en « La Metacognición vía la autoevaluación con grados de certeza » y con postdoctorales en las universidades de Pittsburgh y UCLA. Fue profesor en las Universidades de Namur (1975-1980) y de Liège (1980-2010). Es emérito desde 2010. Ensena como invitado en las Ues. de Liège y Paris 13. Recibió el título de *Honorary Member of the World Cultural Council* (México). Ha colaborado, en Chile, con la U de Chile (UCH -Santiago), la UMCE, la UCT (Temuco), la UC del Maule, la UNAB y la UCSC (Concepción). En Perú con la PUCP y el SINEACE (Lima), la UNSAAC (Cusco) y la UNTRM (Chachapoyas). En México con la U A Chapingo. En España con la U de Sevilla y la U de Deusto (Bilbao). d.leclercq@uliege.be

Álvaro Cabrera Maray

Licenciado en Artes mención Teoría de la Música, y Master en Pedagogía en Educación Superior de la U. de Liège (Bélgica). Ha sido profesor en la Facultad de Artes y en Cursos de formación General, trabajando en el Depto. Estudios de Pregrado de la U. de Chile a cargo del Área de Formación. Integró la Red nacional de Centros de Enseñanza-Aprendizaje y la de expertos SCT-Chile sobre sistema de créditos transferibles. Trabajaba en el Ministerio de Educación de Chile, coordinando los programas de la reforma educacional en Educación Superior. alvarocabreramaray@gmail.com

Contenidos del libro IDEAS:

ES: Calificación ; Evaluación ; Productos ; Meta-cognición ; Resolución de problemas ; Proyectos ; Trabajo de grupo ; Portafolio ; Vigilancia cognitiva ; Pruebas de Progreso ; Taxonomía de Bloom ; Auto-evaluación ; Grados de certeza ; Test de Concordancia de Script ; Retroinformación ; calidades ; validez

EN : Assessment ; Evaluation ; Outcomes ; OSCE ; MCQ ; PARMs ; Metacognition ; Problem solving ; Projects ; Group produced work ; Portfolio ; Cognitive vigilance ; Progress Tests ; Bloom's Taxonomy ; Self-assessment ; Confidence Degrees ; Concordance Script Test ; Feedbacks ; Edumetrics ; Metacognitive Spectral Test ; ETIC PRAD ; quality ; validity

FR : Notation ; Evaluation ; Résultats ; ECOS ; QCM ; PARMs ; Métacognition ; Résolution de problèmes ; Projets ; Travail de groupe ; Portfolio ; Vigilance cognitive ; Tests de progression ; Taxonomie de Bloom ; Auto-évaluation ; Degrés de certitude ; Test de Concordance de Script ; Rétro-information ; Edumétrie ; Test Spectral Métacognitif ; qualités d'une évaluation ; validité d'une mesure

IDEAS = **I**nnovaciones en **D**ispositivos de **E**valuación de los **A**prendizajes en la educación **S**uperior

La lista de los capítulos y el resumen de caca uno

aparece a continuación después de este capítulo.

G.4. La Desviación Estándar (DE): el principal índice de dispersión

En el ejemplo dado, la distribución (ordenada) de las notas es 2, 3, 4, 5, 5, 5, **6**, 6, 7, 7, 8, 9 y 10.

Tabla 18: Definición de medidas o índices de dispersión y sus valores en el ejemplo

ÍNDICES DE DISPERSIÓN	EN EL EJEMPLO (FIGURAS 10 Y 11), CORRESPONDEN A...
El rango (range) o amplitud de la distribución de notas corresponde a la diferencia entre la nota máxima y la nota mínima.	8 = (10–2)
Los cuartiles (Quartiles) o Q1, Q2 y Q3, corresponden a las notas obtenidas por los estudiantes que ocupan la 25ª posición sobre 100 (cuartil 1 o Q1), la 50ª posición (cuartil 2 o Mediana o Q2) y la 75ª posición (cuartil 3 o Q3).	Q1 = entre 4 y 5 (= 4,5) Q2 = 6 Q3 = entre 7 y 8 (=7,5)
La Distancia Inter–cuartiles = Q3–Q1	=(7,5–4,5) = 3,0
La Distancia Semi Inter–cuartiles (DSI) = (Q3–Q1)/2	=(7,5–4,5)/2 = 1,5
La Desviación Estándar (DE), también anotada como "σ" (sigma) es calculada por la fórmula $\sqrt{\Sigma d^2/NS}$, NS siendo el número de sujetos o estudiantes en este caso[218]; Σd^2 siendo la sumatoria de las distancias (elevadas al cuadrado) entre los valores observados y la media de la distribución[219]. Por su parte: $\Sigma d^2/NS$ corresponde a la Varianza. De esa manera, podemos notar que la DE es la raíz cuadrada de la Varianza.	DE = 2,20 Varianza = 4,84

Téngase en cuenta que la *Media (o promedio)* es el valor que minimiza la suma de los cuadrados (d^2) de las distancias a ella misma, y, por consecuencia, minimiza la Varianza y la Desviación Estándar. La DE permite expresar, de manera sintética, la dispersión que existe en la distribución alrededor de la Media. Al igual que las otras medidas de dispersión, la DE nos informa sobre la magnitud de las diferencias que podemos encontrar en una misma distribución.

G.5. La asimetría (Skewness) de la distribución (en J ó en i)

En el Capítulo 6, sección A2, hemos denunciado el "peligroso mito de la curva de Gauss" en educación. Hemos insistido en el principio de que la "curva ideal" en educación no es la "campana" (o campana de Gauss), sino que una curva donde la gran mayoría de los estudiantes se aproxima a la perfección, es decir, a la nota máxima. De ese modo, aparece una distribución en J (Jota) que es "asimétrica". La fórmula que sigue permite calcular el grado de asimetría o *Skewness*:

[218] Es importante hacer notar que el *"NS"* (número de sujetos, que también puede anotarse solo como *"N"*) corresponde a toda la población aquí analizada, dado que el objetivo no es generalizar estos resultados hacia la población general del país, por ejemplo. Para profundizar en este tema se sugiere remitirse a un libro especializado de estadística.

[219] Es importante notar que la suma de las distancias a la media (d–) de los resultados que están por debajo de aquella es equivalente a la suma de las distancias (d+) de los resultados que están por arriba de la media.

Asimetría = Skew= Sk=

$$\frac{n}{(n-1)(n-2)} \quad \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

donde *n* representa el número de notas (de los estudiantes), *x* las notas y *s* su Desviación Estándar (DE).

La fórmula *Sk2 = 3 (Media – Mediana) / DE* es otro modo de calcular la asimetría (Doane y Seward, 2011, p. 10).

Una asimetría (*skewness*) *positiva* indica una distribución con la "cola" a la derecha. La Moda (la nota más frecuente) está ubicada a la izquierda de la Media. Una asimetría (*skewness*) *negativa* indica una distribución con la "cola" a la izquierda, y la Moda a la derecha de la Media.

Tabla 19: Distribuciones con asimetría perfecta (la 1) y con asimetría negativa cada vez más fuerte (2, 3, 4 y 5)

Distrib.	1	2	3	4	5
0	0	0	0	0	0
10	1	0	1	1	0
20	3	3	1	1	0
30	5	4	2	2	0
40	9	6	3	2	0
50	14	12	4	2	0
60	9	17	5	3	1
70	5	6	9	3	1
80	3	2	17	7	3
90	1	0	6	12	10
100	0	0	2	17	35
Media	50	54,2	68,8	80	95,4
Dev. Est.	15	14,5	20	23	8,5
NS	50	50	50	50	153,9
Asimetría	0	–0,53		–1,4	–2,4

Caso 8: Consideremos 5 distribuciones ficticias en la Figura 12, desde aquella en forma de "campana de Gauss" (a la izquierda y perfectamente simétrica), y 4 distribuciones crecientemente asimétricas.

Los valores de *Skewness* aparecen indicados sobre las curvas (mientras más negativas, más en jota es la curva):

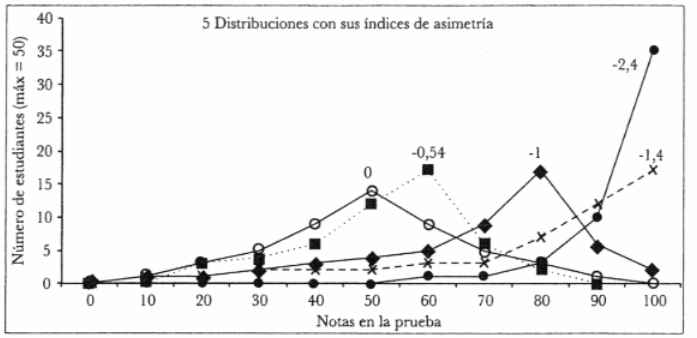


Figura 12: Formas de las 5 distribuciones indicadas en la Tabla 19

G.6. ¿Cuándo se puede rechazar la hipótesis de normalidad (Gauss)?

La respuesta a esta pregunta depende del número de datos que han permitido calcular los valores de Sk o Sk2. La Tabla 20 indica los valores de *Sk2 (Pearson 2 skewness coefficient)* y el número de sujetos que permiten rechazar (a $p<0.10$) la hipótesis de normalidad de la distribución (según Doane y Seward, 2011, p. 11).

Tabla 20: Valores de referencia más allá de los cuales se puede rechazar la hipótesis de normalidad de una curva o distribución, según número de observaciones o sujetos.

Se habla de curva en j a partir de Skew <	Cuando NS =	Se habla de curva en i a partir de Skew >
–0,726	25	0,726
–0,673	30	0,673
–0,594	40	0,594
–0,539	50	0,539
–0,496	60	0,496
–0,462	70	0,462
–0,435	80	0,435
–0,411	90	0,411
–0,391	100	0,391
–0,322	150	0,322
–0,281	200	0,281
–0,23	300	0,23
–0,2	400	–0,2
–0,179	500	0,179

G.7. Ejemplos de valores de Sk en distribuciones espectrales

En la Figura 13 aparecen los valores de Sk de cada uno de los 4 hemispectros obtenidos por Ndabwarukanye (2004, p. 29)[220] en una prueba de 34 preguntas sobre el

[220] Estos resultados ya fueron presentados en el Capítulo 16, sección C.5.

conocimiento de la "cadena de primeros auxilios" (ver dibujo), rendida por 65 profesionales de la salud (nota: no del área de urgencia). Lamentablemente, las distribuciones de respuestas incorrectas (las dos en el hemispectro de izquierda) tienen una Sk positiva (forma de *i*) en vez de negativa (forma de *J*) como las distribuciones de la derecha (respuestas correctas). Afortunadamente, en el post-test la distribución de las respuestas correctas aparece aún más en forma de *J* que en el pre-test. Al mismo tiempo, y de manera consistente con ello, aumenta el número de personas que demuestran haber aprendido. No obstante, en el post-test las respuestas incorrectas aparecen más en *i* (Sk positiva) que en el caso del pre-test, es decir, los errores de algunos de los evaluados parecieron acentuarse durante el proceso de aprendizaje. Evidentemente, toda esta información, dada por el instrumento de evaluación que hemos construido, nos permite reflexionar para mejorar el curso.

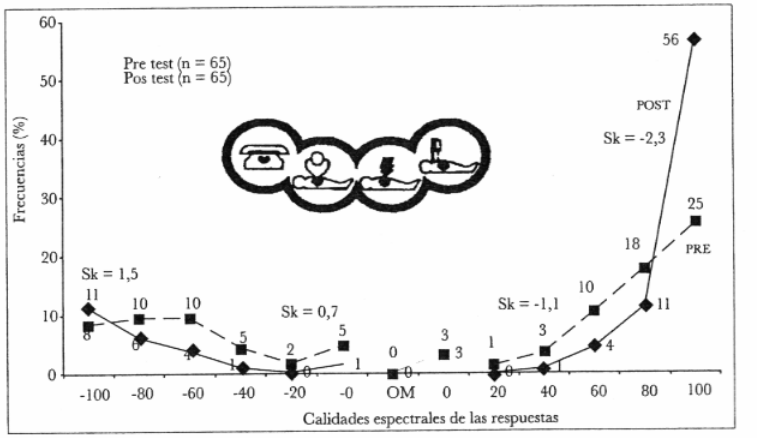


Figura 13: Los dos hemispectros del pre-test y los dos del post-test sobre la cadena de primeros auxilios

G.8. Las formas de las distribuciones

Como ya hemos visto, existen expresiones para describir la forma de una distribución: en "i", en "u", en "j", en "campana". La Figura 14 presenta algunas de estas expresiones, con los valores de la asimetría para cada curva.

Cuando la asimetría es *negativa* (curva en j), la Moda (es decir, la nota más frecuente) está ubicada a la derecha de la Mediana, y la Mediana a la derecha de la Media:

$$Media \leq Mediana < Moda.$$

Cuando la asimetría es *positiva* (curva en i), el orden de los 3 índices es al revés:

Moda < Mediana ≤ Media.

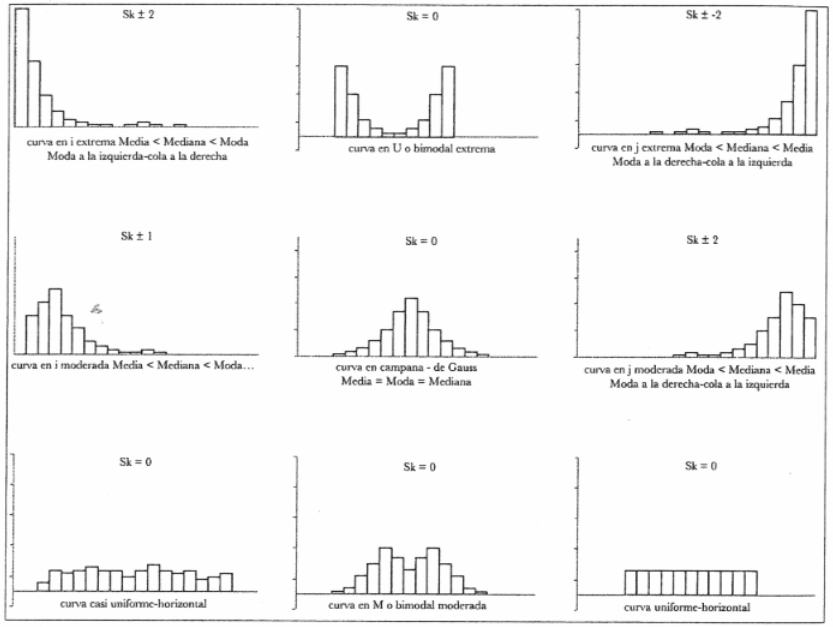


Figura 14: distribuciones típicas y sus grados de asimetría

Parte 3. Índices y decisiones sobre los estudiantes: los dos tipos de referencias para tomar decisiones

Como se mencionó en el Capítulo 2 (sección B), la finalidad de la evaluación puede ser *sancionar* (certificar[221] o seleccionar[222]) o bien *formar* (mejorar los aprendizajes), o una combinación de ambas. Las decisiones pueden basarse sobre dos tipos de *referencias*: normativas (o relativas) y criteriales (o absolutas):

Hablamos de referencias normativas,

- cuando el punto de corte (en un concurso) se hace según el orden en la distribución
- cuando la detección (del nivel de desarrollo) se hace según la distancia a la media

H. Punto de corte según el orden (principio de concurso)

Este caso ocurre cuando el resultado de cada estudiante es atribuido a un determinado nivel en la distribución de las notas de los "n" estudiantes de una cohorte o grupo dado, ordenada de mayor a menor o viceversa. En situación de selección de los "n" estudiantes con las mejores notas, el reto para cada estudiante es lograr posicionarse en un nivel adecuado (arriba del *umbral de posición* o punto de corte definido con anterioridad), comparado con los resultados de los otros estudiantes. En este caso, la *fiabilidad* (ver sección N) *del orden o posición relativa* en la distribución es un asunto crucial.

[221] *Las decisiones* pueden ser referidas a la admisión (la selección) de un estudiante en un año o carrera dado, la atribución a este estudiante de ciertas funciones (como tutor de otros, como moderador de foro, etc.), o bien de posibilidades abiertas (como el derecho a elegir una especialización, etc.).

[222] Por ejemplo, cuando se trata de un concurso donde son admitidos los 50 mejores.

CASO 9: EL EXAMEN DE INGRESO.

En marzo, el profesor P participa junto a otros cuatro colegas de la Facultad de Educación en la selección de 200 candidatos que buscan ser admitidos en esa Facultad. Para ello se construyó una prueba de ingreso con un máximo de 50 puntos y se busca seleccionar las 200 mejores notas. Este año, 327 estudiantes son candidatos. El que tiene el ranking más alto (lugar número uno) obtuvo 47,5 puntos, mientras que los postulantes con peores resultados en esta prueba (entre los 327) obtuvieron 8,3 puntos.

A continuación se presentan los resultados de este año comparados con los del año pasado:

	N° de candidatos	Puntaje (de 0 a 50) del postulante menos bueno	El n° 201 obtuvo:	El n° 200 obtuvo:	El n° 199 obtuvo:	Puntaje (de 0 a 50) del mejor postulante (clasificado en el n°1)
Este año	327	8,3	21,4	21,4	21,4	47,5
El año Pasado	383	6,4	17,7	17,9	18,3	42,2

El año pasado no hubo ningún problema para seleccionar los "top 200": el número 201 tenía una puntuación inferior al del número 200 (2 décimas menos 17,7 v/s 17,9). Sin embargo, este año sí hubo un problema: tres candidatos tienen la misma calificación (21,4). La prueba no era lo suficientemente capaz de discriminar para tomar la decisión. Es decir, no fue capaz de detectar o crear diferencias suficientemente grandes entre los sujetos. En ese caso, se hace necesario que el jurado imagine una solución (por ejemplo, entrevistar a estos tres estudiantes). También podemos ver que un estudiante con una puntuación igual a 20 (con puntaje mayor a 17,9) sí habría sido aceptado el año pasado. Sin embargo, es la referencia al orden de los puntajes del año en curso (la clasificación) lo que sirve de criterio para tomar la decisión (criterio anunciado), puesto que el punto de corte es "el número 200". La referencia de la decisión es llamada entonces "normativa" dado que se refiere a los resultados de un determinado grupo de alumnos, independiente de si el puntaje es mayor o menor a un determinado umbral o varía de año en año.

I. La clasificación según la distancia a la Media de una cohorte o grupo

I.1. Posicionamiento vía Notas Z:

Nota z de un individuo = (X-M) / σ, donde

- *X* es la nota de un sujeto en un test,
- *M* es la media de las notas de dicho test,
- σ es la desviación estándar (sigma o DE) de las notas obtenidas en dicho test.

Generalmente se escribe el signo de la nota *z* (positivo o negativo), ya que eso indica si la distancia a la media (expresada en unidades de desviación estándar) se encuentra por debajo o por encima de la media.

Si en una distribución (por ejemplo, en la prueba G1), la Media (o promedio) es de 5,92 y la desviación estándar (DE) es de 2,20, un estudiante cuya puntuación es 8 tiene una nota Z positiva, puesto que está por encima de la media (8 > 5,92). La nota Z (o z-score) se calcula como: (8 – 5,92) / 2,20) = 2,08 / 2,20 = +0,95. Así, 0,95 viene a ser la nota Z o z–score de ese estudiante en su grupo.

Para el caso de un alumno cuyo puntaje es 5, la nota z es (5 – 5,92) / 2,20) = –0,42. Dado que la nota es inferior a la media su z-score será negativo.

Los valores de la puntuación o notas Z suelen oscilar entre –3 y +3.

La posición (o puntuación Z) de un estudiante en una distribución puede ser utilizada como señal de alarma. Este es el caso de la operación RESSAC (véase el Capítulo 3, sección B4), en donde las calificaciones de los estudiantes se representan en una "radiografía Z" que indica su posición en su grupo-clase. A medida que su puntación Z está más por debajo de la Media, más preocupante es su puntuación, y más urgente se hace indagar en las razones que pueden explicar esa posición.

Sabemos que algunas habilidades se adquieren a través de la vida, muy lentamente y no al mismo tiempo ni desde el mismo punto de partida para todos. Una curva de frecuencias acumuladas muestra el porcentaje real de personas que han logrado una determinada habilidad. El desarrollo motor proporciona ejemplos típicos de curvas (en S mayúscula) tal como se muestra en la Figura 15 (Pikler, 1979) para 600 niños de peso normal al nacer. ¡Claramente, las pendientes de las cuatro curvas no son las mismas!

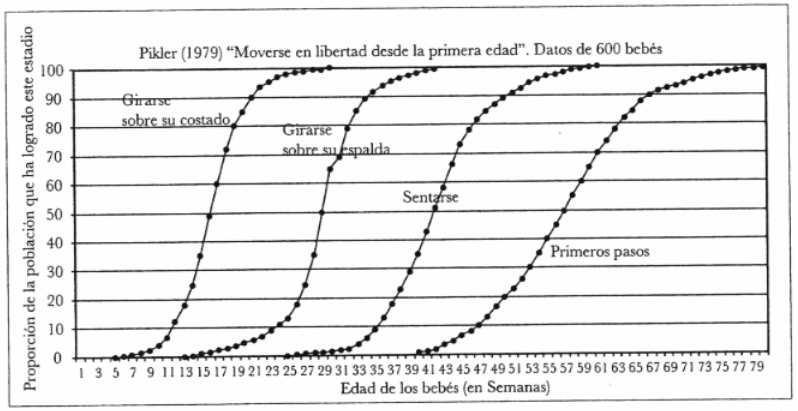


Figura 15: CCI de cuatro capacidades motoras de los bebés (darse vuelta de costado, darse vuelta de espaldas, sentarse, primeros pasos) según Emmi Pikler

I.2. Deciles

Como vimos más arriba, la posición de la nota de una persona en una distribución ordenada puede ser expresada en "notas z". También lo puede ser por deciles o percentiles si tenemos referencias suficientemente finas de los datos. En la Figura 16 la curva de desarrollo de los "primeros pasos" (Fig. 15) se ha rediseñado para permitir la lectura (vertical) en deciles. El decil 1 comprende al primer 10% de los sujetos, en este caso, a los niños más precoces o avanzados. Eso significa que en este ejemplo, solo el 10% de los niños ha dado sus primeros pasos a esa edad. Por su parte, los niños del 10° decil son los más retrasados, puesto que el 90% de su edad ya ha dado sus primeros pasos antes que ellos. Los valores numéricos que se leen en vertical corresponden a frecuencias acumuladas (que van de 0% a 100%) de los niños que dieron sus primeros pasos a una edad determinada, la que se indica (en semanas) en el eje X u horizontal. Los percentiles se asumen bajo la misma lógica pero el conteo es mucho más fino ya que se hace por grupos de 1%, lo que permite tener cifras más detalladas.

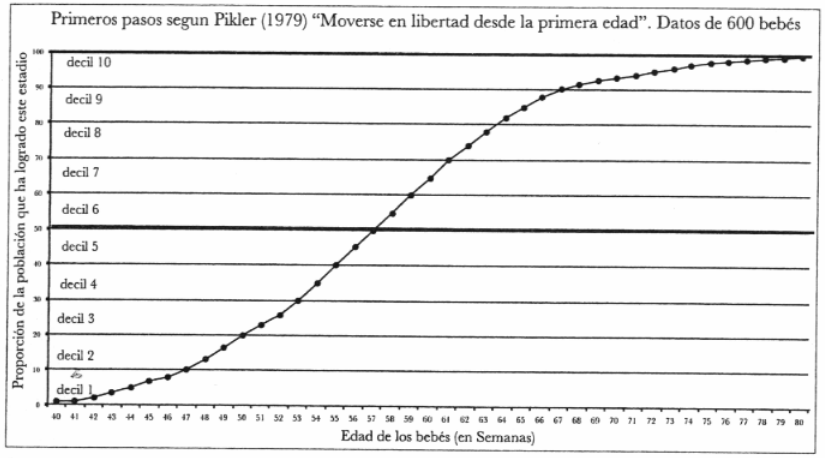


Figura 16: Curva que permite interpretar fácilmente en deciles las frecuencias acumuladas de niños que han efectuado sus primeros pasos a una edad dada (en semanas)

I.3. Aprendizajes escolares

A diferencia de los aprendizajes espontáneos, la adquisición de muchos saberes se acelera vía la escolaridad. En este caso, la curva de aprendizaje se diferencia de la curva natural (u ojiva de Galton).

Caso 10: en 1973 Albert, un niño francés que está finalizando su 6to año escolar, ya sabe cómo se escriben correctamente 1.000 palabras. ¿Está adelantado, atrasado o "a tiempo" en relación con los niños de su edad?

Ters, Mayer y Reichenbach (1973, p. 24) "trazaron" la progresión promedio midiendo el número promedio de palabras[223] que en 1973 un 75% de estudiantes de un nivel escolar determinado (en Francia) han adquirido en ortografía (es decir, que saben escribir sin faltas de ortografía). La unidad de categorización en este caso es el "año escolar" (al final del año). El valor de referencia del 75% fue arbitrariamente elegido, básicamente por tradición, siguiendo la prueba de Binet-Simon de "inteligencia" general, en donde una pregunta se considera típica de una cierta edad cuando el 75% de los niños de esa edad tiene éxito al responderla.

[223] Extraídas de la escala de ortografía Dubois-Buysse, que cuenta con 4.000

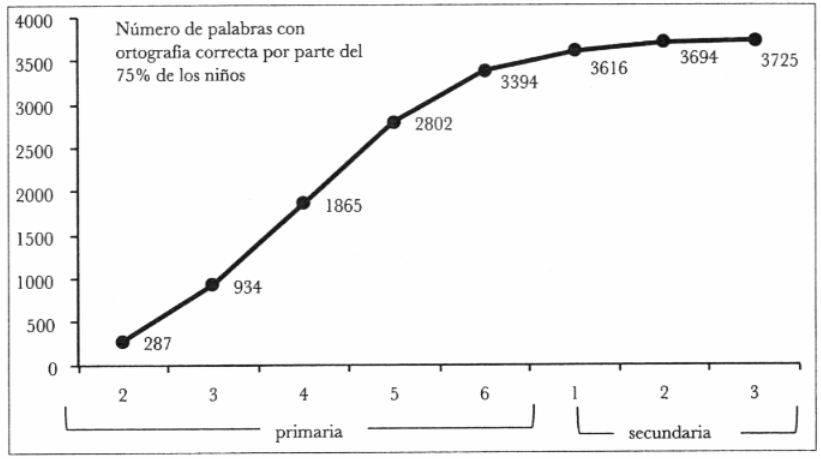


Figura 17:Número de palabras correctamente escritas (ortografía) por el 75% de los niños de un año escolar determinado en Francia (1973)

Nuevamente observamos la forma logística exponencial de la curva (en S mayúscula). Esto significa que hay una zona de inclinación (acantilado), es decir, una zona en donde el progreso es más rápido que antes y en donde la curva se hace más "vertical". En este caso, la aceleración tiene lugar entre el 3° y el 5° grados. Luego los progresos se hacen más lentos. Se trata de un fenómeno bien conocido: muchos adultos, durante años enteros (incluso décadas), no aprenden a escribir correctamente ninguna palabra nueva.

En cada oportunidad hemos recalcado "en el año 1973, en Francia", porque las estadísticas presentadas deben ser siempre actualizadas (al menos cada 10 años) y situadas (en este caso, en Francia). La curva sirve como un ábaco que permite localizar a Albert (el niño de nuestro ejemplo), ubicándolo en 3er año de primaria, es decir, mostrando que tiene 3 años de retraso en relación al 75% de su nivel escolar. Aunque no es objeto de este capítulo, la obtención de informaciones de este tipo nos permite valorar –en su justa medida– la existencia de test estandarizados, los que deben ser usados de manera prudente y pertinente (nunca para "rankear" a Albert y decir a todos que es "un perdedor" o "un ganador", sino más bien para preguntarnos por las razones de su retraso y tratar de actuar sobre ellas).

I.4. Medición del impacto de la enseñanza y la formación

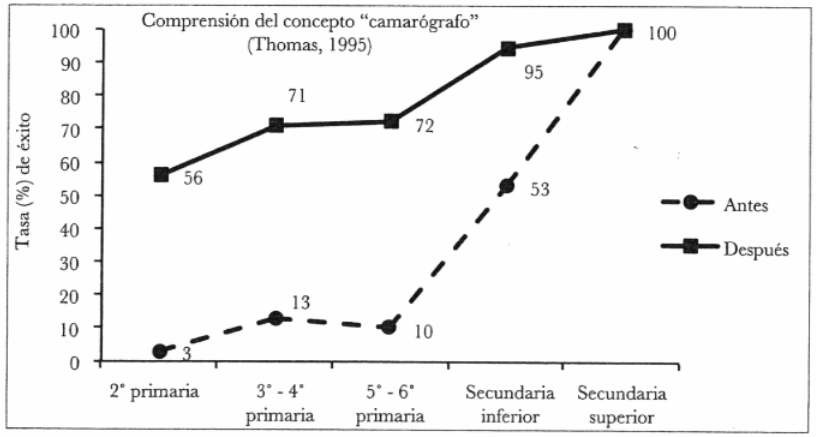
Caso 11: En 1995, como parte de una operación educativa llamada TELECOLE, F. Thomas condujo sesiones de educación desde los medios de comunicación para cursos con niños de cinco grados escolares diferentes (ver Figura 18). Como se trataba de una investigación–acción, se midió el impacto de este programa en la comprensión de variados conceptos como "libreto", "rol", "actor", "director", "productor", "camarógrafo", etc.

Leclercq (asesor de esta investigación-acción) sugirió medir la comprensión de estos conceptos…

- ANTES de las sesiones de educación en los medios de comunicación, para medir la evolución espontánea (sin sesiones de educación) en los últimos años.
- DESPUÉS de estas sesiones, en orden a cuantificar el impacto de la diferencia entre el antes y el después.

Estas sesiones vienen a ser un ejemplo de la aceleración (vía la escuela) del aprendizaje espontáneo.

La Figura 18 muestra los resultados para un concepto (camarógrafo) bajo la forma de dos curvas, ambas en forma de S mayúscula... pero con un punto de inflexión más a la izquierda (por lo tanto, precoz) para el grupo de los estudiantes "después", mientras que la inflexión para el grupo "espontáneo" o "antes (sin intervención)" (medida en los pre-test) se situó recién en el paso de la primaria a la secundaria.



	2° primaria	3°–4° primaria	5°–6° primaria	Secundaria inferior	Secundaria superior
Antes	3	13	10	53	100
Después	56	71	72	95	100

Figura 18: Evolución "espontánea" y "acelerada" (vía sesiones de educación en los medios de comunicación) de la comprensión del concepto "camarógrafo", Bélgica, 1995 (Poumay y Leclercq, 1999)

Una vez más, reiteramos que estos datos deben ser actualizados y localizados.

Estos dos ejemplos (I.3 e I.4) ilustran, en el campo del conocimiento, aquello que debería desarrollarse en el ámbito de las competencias genéricas (como educación para la ciudadanía, educación sexual, en TIC, comunicación, pensamiento crítico, concepciones epistemológicas, etc.), es decir, puntos de referencia del desarrollo en los diversos niveles escolares o en determinadas edades. Todo eso en función de detectar eventuales retrasos en el desarrollo.

J. Referencias criteriales

En este caso, el resultado del estudiante es igual o superior/inferior a un *umbral "de corte"* previamente establecido, con independencia de los resultados de los otros estudiantes. En caso de selección o de certificación, la *fiabilidad* de la nota (es decir, su *error de medición*) es un aspecto crucial.

J.1. Fijar un umbral de éxito *a priori*

A priori en este caso significa "antes de conocer los resultados de los estudiantes" y aun antes de anunciar la prueba. Hemos visto en los capítulos y secciones 16.A.4 y 18.J.1 el método de Angoff para hacer eso. Pero ¡atención! Utilizar grados de certeza aumenta las exigencias porque la nota máxima no la puede lograr un estudiante que ha respondido correctamente todas las preguntas, sino un estudiante que lo ha hecho indicando a la vez un *grado de certeza máxima*. Es la razón por la cual se han definido "techos" o "umbrales de excelencia" con grados de certeza según los ámbitos: escuela, pacientes, pilotos, enfermeras, etc. (ver Capítulo 17).

J.2. Fijar el umbral de éxito *a posteriori*

Una vez conocidos los resultados es posible modificar la severidad. Por ejemplo, si nos parece que el examen ha sido demasiado difícil se puede dar a cada estudiante un nivel adicional de calificación. Aunque teóricamente el problema de la severidad se plantea en la misma forma cuando el examen es más fácil, por razones deontológicas, es casi imposible cambiar el umbral para que sea más alto que lo que fue anunciado. Tampoco parece aceptable disminuir un punto o nivel a la calificación de cada estudiante. Si en un curso la tasa de logro general (de los estudiantes) para la prueba de un curso resulta insatisfactoria, el docente puede pensar en cambiar, en el futuro, el dispositivo de enseñanza, de aprendizaje, de evaluación… incluyendo los objetivos si fuera necesario.

J.3. Criterios múltiples, Amplitud de Déficit y votos

Además de la Media de las notas, se puede considerar otros índices como, por ejemplo, la *Amplitud del Déficit* (*AD*). Aquella se calcula por la suma de las diferencias entre las notas inferiores al umbral y el umbral. Por ejemplo, si el umbral de éxito en un curso es 12/20, un estudiante que ha obtenido un 8 en un curso X tiene, en este curso, una AD de 4 puntos. Si ha obtenido un 10 en otro curso Y, en este curso su AD es 2, y en total su AD = 6.

El caso 12 y la Tabla 21 ilustran un modo de *combinación de criterios múltiples* sugerido por Leclercq a la Facultad de Psicología y Educación de la Universidad de Liège. Facultad que lo ha utilizado desde 1995 con cambios menores de año en año (en las reglas y en los valores de la tabla).

CASO 12: LA DELIBERACIÓN DE FIN DE AÑO

Al final del año escolar (en enero), el profesor P participa en el jurado de los nueve profesores que durante el año impartieron cursos a la cohorte de estudiantes seleccionados al inicio de año. Cada uno de estos 9 profesores otorgó una puntuación entre 0 y 20 a los 187 estudiantes de esta cohorte (13 de los 200 seleccionados inicialmente se retiraron durante el año). El propósito de la reunión de "deliberación" es decidir quiénes serán autorizados para pasar al año siguiente sin necesidad de presentarse ante una junta examinadora. Nadie sabe, antes de la deliberación, cuántos estudiantes van a tener éxito entre los 187 (podrían ser 0 o 187).
En este caso, el resultado de cada uno (pasar de año / no pasar) no depende de los resultados de los demás. Para pasar al próximo año, varios criterios fueron anunciados a los estudiantes al inicio del año:
Regla 1: El umbral promedio o puntaje de corte para la media de puntos es de 12, lo que significa que el promedio de las 9 notas de los estudiantes debe ser igual o mayor a 12.
Regla 2: Ninguna de las 9 calificaciones debe ser inferior a 7 (en una escala de 20) lo que supondría una nota de exclusión.
Regla 3: La amplitud de los déficits (AD) acumulados no puede exceder al valor 6 (valor de corte), siendo 12 la nota de referencia para cada curso. Por ejemplo, un estudiante que tiene un 9 en un curso aumenta en 3 (3 puntos por debajo de 12) la amplitud de su déficit. Si consiguió un 10 en otro curso se le asigna un déficit magnitud 2. Por lo tanto, obtiene un total de cinco para este indicador (la AD acumulada es igual a 5).
Regla 4: Un promedio alto (de las 9 notas) puede "compensar" una amplitud de déficit (alta).
Regla 5: El jurado puede considerar informaciones adicionales a las notas (accidente, drama familiar, etc.).
Regla 6: Para estar seguro de aplicar fácilmente y de manera aceptable y con rigor para todos (equidad) las mismas reglas múltiples, una tabla de doble entrada proporciona zonas de "voto del jurado".

Tabla 21: Reglas y pauta de deliberación del tribunal de la FASPE de la ULg (antes de 2005) donde R = Repitencia, S = Suficiencia, D = Distinción, GD = Gran Distinción, MGD = Distinción Máxima.

MEDIA DE LAS CALIFICACIONES FINALES EN TODOS LOS CURSOS (UMBRAL DE ÉXITO = 12. MÁXIMO = 20)									
AD	11,8–11,9	12–12,4	12,5–12,9	13–13,7	13,8–14,9	15–15,5	15,6–16,4	16,6–17,3	17,4 o +
0-4	S	S	S	S	D	D	GD	GD	MGD
5	Votar S	Votar S	S	S	Votar D	D	Votar GD	GD	Votar MGD
6	R	Votar S	Votar S	S	S	Votar D	D	Votar GD	GD
7	R	R	Votar S	Votar S	S	S	Votar D	D	Votar GD
8	R	R	R	Votar S	S	S	S	Votar D	D
9	R	R	R	R	Votar S	S	S	S	Votar D
10	R	R	R	R	R	Votar S	Votar S	S	S
11	R	R	R	R	R	R	R	Votar S	Votar S

Parte 4. Índices de evolución de progreso

Los índices que se presentan a continuación utilizan conceptos estadísticos que admiten la inferencia. Las diferencias medidas –a partir de los mismos estudiantes– son posibles en el marco de una investigación a nivel de la clase.

K. La Media de las *diferencias* (progresos) individuales (Mdif) en distribuciones MÉTRICAS

Es posible calcular la significación estadística de la diferencia pre-post para un mismo grupo de estudiantes. La Tabla 22 ilustra esta posibilidad a través de la fórmula del indicador *t de Student* para grupos aparejados (es decir, los mismos 9 estudiantes antes y después de una formación). Cada prueba tiene un máximo posible de 100% y un mínimo posible de 0%.

Tabla 22: Cálculo de las medias de dos distribuciones (PRE y POST) de nueve notas

	E1	E2	E3	E4	E5	E6	E7	E8	E9	M	DE	Valor t
PRE	29	32	40	41	43	44	45	47	52	41,4	7,2	
POST	72	81	72	84	78	87	72	83	92	80,1	7,2	
d	43	49	32	43	35	43	27	36	40	38,7	6,76	17,1

Para las muestras aparejadas la fórmula del índice t de Student es: $t = Mdif \;/\; (DEdif \;/\; \sqrt{NS})$

Donde *Mdif* = Media de las diferencias entre las dos notas de una misma persona.
DEdif = La Desviación Estándar de esas diferencias.
NS = Número de sujetos o estudiantes (y, por consecuencia, de diferencias).

Para saber qué nos indica este test debemos consultar la tabla de los valores *t* (Tabla 23) para NS-1 grados de libertad, es decir 8. Para (NS-1)=8 y una probabilidad de error $p<0,05$ (es decir, una probabilidad de 5 sobre 100 de equivocarse), el *t* de referencia es 2,31 (ver Tabla 23 siguiente). Para $p<0,01$ es 3,36 y para $p<0,001$, es 5,04. Todos esos valores son superados por el valor de *t* que obtuvimos en nuestro ejemplo (17,1). Podemos decir entonces que la diferencia es estadísticamente significativa con una probabilidad de error inferior a 1 en 1.000. Así, comparando el PRE-test y el POST-test podemos pensar que el proceso de aprendizaje ha sido fructífero, debido a... (¿las metodologías usadas?, ¿la motivación?, ¿otras razones?).

Tabla 23: Valores referenciales t de Student

Grados de libertad (v)	P = .10 (10%)	P = .05 (5%)	P = .02 (2%)	P = .01 (1%)	P = .001 (0,1%)
1	6,31	12,71	31,82	63,66	636,62
2	2,92	4,30	6,96	9,92	31,60
3	2,35	3,18	4,54	5,84	12,94
4	2,13	2,78	3,75	4,60	8,61
5	2,02	2,57	3,36	4,03	6,86
6	1,94	2,45	3,14	3,71	5,96
7	1,90	2,36	3,00	3,50	5,41
8	1,86	2,31	2,90	3,36	5,04
9	1,83	2,26	2,82	3,25	4,78
10	1,81	2,23	2,76	3,17	4,59
11	1,80	2,20	2,72	3,11	4,44
12	1,78	2,18	2,68	3,06	4,32
13	1,77	2,16	2,65	3,01	4,22
14	1,76	2,14	2,62	2,98	4,14
15	1,75	2,13	2,60	2,95	4,07
16	1,75	2,12	2,58	2,92	4,02
17	1,74	2,11	2,57	2,90	3,97
18	1,73	2,10	2,55	2,88	3,92
19	1,73	2,09	2,54	2,86	3,88
20	1,72	2,09	2,53	2,84	3,85
21	1,72	2,08	2,52	2,83	3,82
22	1,72	2,07	2,51	2,82	3,79
23	1,71	2,07	2,50	2,81	3,77
24	1,71	2,06	2,49	2,80	3,75
25	1,71	2,06	2,48	2,79	3,73
26	1,71	2,06	2,48	2,78	3,71
27	1,70	2,05	2,47	2,77	3,69
28	1,70	2,05	2,47	2,76	3,67
29	1,70	2,04	2,46	2,76	3,66
30	1,70	2,04	2,46	2,75	3,65
35	1,69	2,03	2,44	2,72	3,59
40	1,68	2,02	2,42	2,71	3,55
45	1,68	2,02	2,41	2,69	3,52
50	1,68	2,01	2,40	2,68	3,50
60	1,67	2,00	2,39	2,66	3,46
120	1,66	1,98	2,36	2,62	3,37
∞	1,64	1,98	2,33	2,58	3,29

L. La Ganancia Relativa (GR)

Para calcular este índice (GR), la fórmula que podemos usar es: *GR = (Ganancia / Ganancia Posible) * 100* (Hovland *et al.*, 1949). Esta fórmula exige que exista una máxima conocida. Pero, ¡atención! La Media de las GR de cada estudiante no es la GR calculada sobre las dos Medias. Por su parte, la fórmula de la Pérdida Relativa exige que haya una mínima conocida. *PR = (Pérdida – Pérdida posible) *100.*

Visualización de la AE (Amplitudes de Efecto) y de la GR (Ganancia Relativa) con gráficos de "doble escalera" y polígonos de frecuencia:

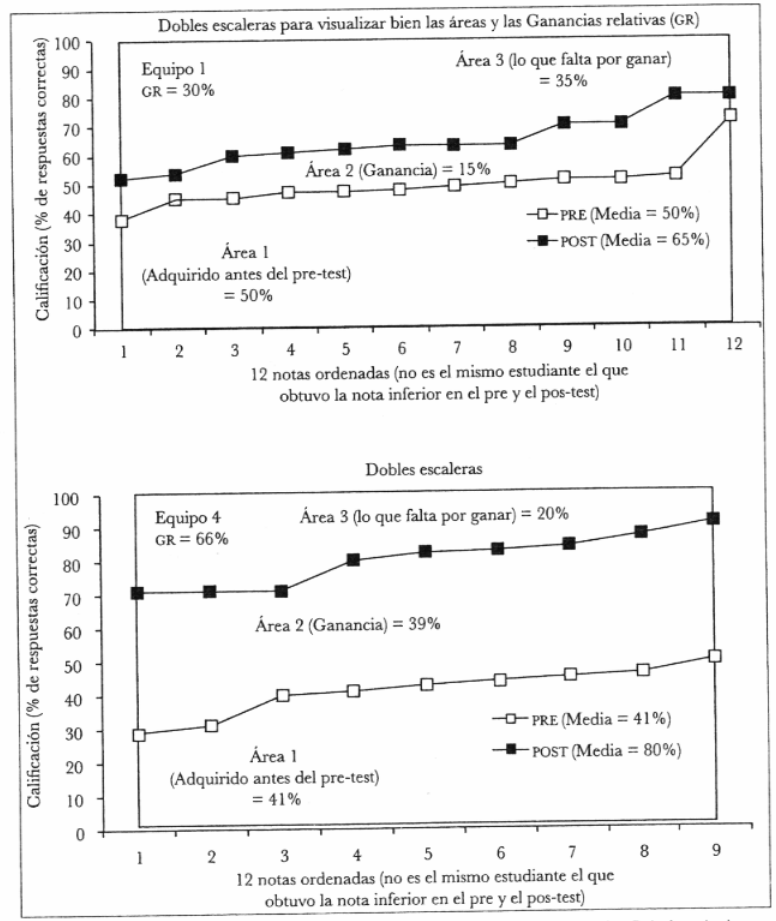


Figura 19: Dobles escaleras para visualizar bien las áreas y las Ganancias Relativas (GR)

Un gráfico de "doble escalera" consiste en ordenar los resultados del PRE-test y del POST-test para dibujar las dos curvas de manera simultánea, tanto para los resultados de un grupo de control (aquí Grupo 1 o G1, con 12 estudiantes), y de un grupo "experimental" (G4, con 9 estudiantes) que han pasado ambos el PRE y el POST.

Las escaleras permiten visualizar, en el total de estudiantes ordenados según sus logros, las tres SUPERFICIES (y su importancia en %): la superficie *(1)* que corresponde al PRETEST y que muestra lo que las personas ya dominan o conocen; la *(2)* que implica la GANANCIA propiamente tal; y la *(3)* que muestra lo que queda por "ganar" o mejorar hasta llegar al techo de un 100% de dominio del tema o área.

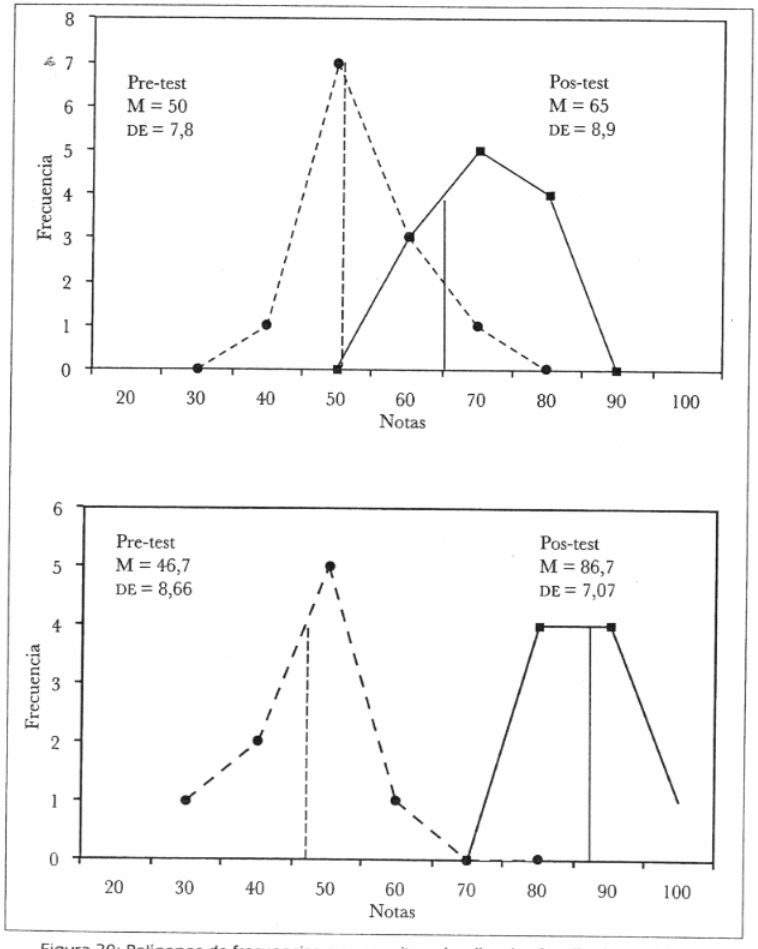


Figura 20: Polígonos de frecuencias que permiten visualizar las Amplitudes de Efecto

Por su parte, las amplitudes de efecto (Figura 20) pueden estimarse así: para el Grupo 1, AE = (65-50)/7,8 = 1,92. Y para el Grupo 4, AE = (80-41)/7,2 = 5,41.

Los polígonos de frecuencia (histogramas) permiten visualizar comparativamente los avances entre PRE y POST para cada grupo, así como las respectivas dispersiones. Estas últimas tienden a ser más fuertes en el G1 que en el G4. Este tipo de gráfico también permite ver a simple vista las posiciones de las Medias y la forma de cada distribución (que aquí tienden a tener forma de campana o curva de Gauss).

M. La Amplitud del Efecto (AE) u Effect Size (ES)

M.1. La significación práctica de las diferencias

En su artículo "La significación práctica, un concepto al que le ha llegado su tiempo", Kirk (1996) recuerda que durante decenios la investigación experimental en psicología ha sido dominada por el concepto de *significación estadística*, es decir, por la estimación de la probabilidad de equivocarse afirmando, por ejemplo, que la existencia de una diferencia observada es causada por el azar. En otras palabras, "aceptar la hipótesis nula", lo que significa decir "no hay diferencias fundamentales aunque, a causa de errores de medición, las diferencias son observadas"–[224]. De manera habitual, en ciencias humanas los investigadores aceptan una probabilidad de error de 5 sobre 100 (lo que se codifica por la expresión $p<0,05$ ó p.05). Sin embargo, lo que Kirk recomienda es preocuparse de la significación práctica de las diferencias, medida a través del Índice de Amplitud del Efecto (*Effect Size*).

Junto a Tukey (1991) y Cohen (1994), Kirk subraya el carácter ritual del enfoque estadístico, citando (p.748) a Rosnow, y Rosenthal (1989), comenta que *A Dios le gusta casi tanto p.06 como p.05*. Insiste además sobre el hecho de que la imposibilidad de rechazar la hipótesis nula no es la confirmación de la ausencia de diferencia. Por ejemplo, en el Capítulo 16 hemos visto muchos casos en los cuales, ante la ausencia de grados de certeza, el observador concluye que no hay diferencias aunque con grados de certeza sí se puedan observar tales diferencias.

M.2. La Amplitud del Efecto (AE) y los metaanálisis

Gene Glass (1976) recomienda calcular la AE porque ella sirve de unidad de medición común entre muchos resultados de investigaciones diferentes (que usan unidades de medición diferentes) pero que tratan un mismo asunto. Establecer una medida de AE facilita el cálculo de valores promedio para los resultados agregados. Eso es lo que se conoce como un *metaanálisis*. Glass argumenta que estos metaanáli-

[224] Los trabajos de Fisher sobre el test de hipótesis datan de 1925 mientras que los de Spearman sobre los *errores de tipo I* (rechazar la hipótesis nula siendo ella verdadera) y los *errores de tipo II* (aceptar la hipótesis nula siendo ella falsa) datan de 1928.

sis permiten a los investigadores asimilar la progresión del conocimiento. De hecho, su frase favorita es:

> Miles de doctorados quedan mudos si nadie los ha leído. Los datos están en los libros, el conocimiento está en los cerebros.

Glass ha propuesto una fórmula simple de cálculo de la *AE: (M2-M1/DE1)*, la que se puede leer como: "La diferencia entre los dos resultados dividida por la Desviación Estándar de los primeros resultados". La AE está por tanto expresada en notas estandarizadas (como la nota Z). Por su parte, Hedges (1981) propone una fórmula que integra las dos Varianzas (VAR1 y VAR2):

$$AE = (M2\text{-}M1) / \left(\frac{\sqrt{(VAR1) + (VAR2)}}{2} \right).$$

M.3. Amplitud del Efecto y año escolar

Glass, McGaw y Smith (1981, p. 103) proponen utilizar "un año escolar" (un grado) como unidad de razonamiento y medida. Consideran que, en la enseñanza primaria, el alumno promedio "gana" en habilidades a lo largo de los 10 meses del año escolar (considerando 2 meses de vacaciones). Ellos muestran que, rindiendo los test de dominio (en Inglés, en Aritmética, en Ciencias) este alumno promedio supera al alumno promedio del año anterior por una Desviación Estándar (DE). En consecuencia, la Amplitud del Efecto de un mes del año escolar es más o menos +0,1 (o +10%). Del mismo modo, una diferencia de +0,3 AE equivale a ganar tres meses de trabajo escolar. Esta manera de razonar –que vale solo para la escuela primaria en el trabajo de estos autores– permite contestar la pregunta: "¿Valen los esfuerzos y recursos invertidos en una reforma que logra, en promedio, una AE de +0,15?".

M.4. Referencias normativas de AE gracias a los metaanálisis

Cuando se trata de reunir datos de muchos grupos en comparación (que es lo que hacen los meta-análisis), los valores promedio de AE son, a menudo, menos altos que en el ejemplo de las figuras 19 y 20, en donde se trata de progresos para un mismo grupo. Por ejemplo, en un meta-análisis de investigaciones que comparan el aprendizaje individualizado con la enseñanza colectiva tradicional, las AE son las siguientes (el número de investigaciones consideradas se indica entre paréntesis): + 0,12 (51) en la primaria (Hartley, 1978), + 0,11 (51) en la secundaria (*ídem*), +0,60 en la educación superior (Kulik, 1979). En un meta-análisis más detallado, Bangert *et al.* (1983) calculan, para el aprendizaje individualizado, las siguientes medias de AE:

+0,10 (49) para el conocimiento
+0,14 (14) para las actitudes relativas al contenido
+0,26 (4) para el pensamiento crítico
+ 0,26 (4) para la autoestima.

En su meta análisis de investigaciones que comparan la enseñanza tradicional y los Usos Pedagógicos del Ordenador (UPO), Bangert *et al.* (1985) calculan, tanto para el nivel de la escuela primaria como de la secundaria:

+0,40 (9) para la Enseñanza Manejada por Computador (prescripciones, pruebas)
+0,36 (17) para la Enseñanza Asistida por Computador (cursos tutoriales)
+0,07 (16) para la Enseñanza Enriquecida por Computador (programación, simulación)

Estos autores también han observado que las AE varían según las competencias de los estudiantes, mostrando las siguientes AE promedio según el tipo de estudiante: con fuertes competencias: +0,24 (4); con competencias de nivel medio: +0,13 (4); con competencias débiles: +0,46 (4).

En relación con la actitud hacia el computador, la AE promedio era de +0,62 (4).

Parte 5. Índices sobre las pruebas (test) y preguntas

N. La fiabilidad (reliability) de la medición que resulta de una prueba

N.1. La importancia de la fiabilidad en educación y en la medición del CI

En psicología diferencial, cuando se trataba de medir el Coeficiente Intelectual (CI) alrededor del año 1900, no había un acuerdo entre científicos sobre lo que era "la inteligencia" (tampoco hay acuerdo hoy en día). En ese marco, la idea que cobró fuerza fue la de medir aptitudes generales que se pueden aplicar a cualquier contenido, tales como la inteligencia numérica, espacial, de razonamiento, etc. Los contenidos no importaban en sí mismos a condición de que el resultado final (el *orden* entre las personas testeadas o sus *posiciones en la distribución* del total de personas) fuera confiable. En ausencia de una definición común de "inteligencia"[225], muchos científicos postularon que no se podía medir una "construcción hipotética" (*construct*). Alfred Binet, el francés que inventó los test de inteligencia en 1904, dijo: *La inteligencia es lo que mide mi test.*

En educación no solo los procesos son importantes, sino que también los contenidos y su relación con los procesos, es decir, el conjunto proceso-contenidos.

La Tabla 24 resume las diferencias entre las dos visiones.

Tabla 24: Resumen de las diferencias entre la evaluación en educación y en medición del CI (Psicología Diferencial)

En una prueba (o test) de N ítems...	... en Educación	... en la medición de un aspecto del CI en psicología diferencial
1 ... cada pregunta (o ítem)...	... puede medir otro proceso mental y/u otro contenido diferente al de las otras preguntas... entregando información sobre este conjunto "proceso-contenido".	... debe medir el mismo proceso mental que los otros ítems, cualquiera sea el contenido; ... solo es un elemento que ayuda a establecer la nota o puntaje total.
2. ...lo que se mide ...	...son varios procesos mentales aplicados a contenidos.	...es una aptitud general (que se puede aplicar a todos los contenidos).
3. ...la validez de cobertura de los contenidos...	...importa si se trata de certificar algo o de seleccionar para un programa definido.	...no importa (los contenidos no son relevantes).
4. ... la consistencia interna de cada pregunta con las otras (con la prueba total)...	...es poco importante porque cada pregunta puede ser un conjunto "proceso-contenido" diferente de las otras preguntas.	...es importante porque todos los ítems deben medir un solo proceso mental, intentando así eliminar la influencia de los hábitos culturales[226].
5. ... el postulado teórico es que...	...las capacidades son resultado del aprendizaje y evolucionan.	...el CI se considera como "fijo", innato, no sujeto a cambios.
6. ... posicionar el resultado de un estudiante al interior de su grupo...	...no importa, a menos que estemos en una situación de concurso o selección.	...es crucial.

[225] Que sea espacial, númerica, verbal, de razonamiento, etc.

[226] Carraher *et al.* (1985) han demostrado que niños que practican la economía informal en las calles pueden lograr buenos resultados en razonamiento matemático si las preguntas que se les hacen tienen la forma de operaciones comerciales auténticas con monedas y pequeños números.

Lo anterior conlleva que el establecer un *índice de replicabilidad* de una prueba (*reliability*) adquiera una importancia diferente, según si usamos la perspectiva de la psicología diferencial o la de la educación.

En educación se quiere obtener *un máximo de informaciones diferentes*, para mejorar, remediar o asegurar la "cobertura" del programa.

En psicología diferencial se quiere medir *un solo proceso mental* (muchas veces), de modo que la media de las numerosas mediciones sea "robusta", es decir, tenga un gran valor de replicabilidad.

Una prueba que tiene un *índice de fiabilidad* alto es una prueba en la cual los ítems miden todos un mismo proceso, permitiendo así un *resultado total* con *un pequeño error de medición*. Si el índice de replicabilidad de una medición o test logra un valor determinado (a menudo 0,8), los que tomarán las decisiones (de selección, etc.) en base a ese test pueden entonces fiarse de sus resultados.

N.2. Cómo reducir el Error de medición

Hay dos maneras de reducir el error de medición de una prueba:

- crear una prueba unidimensional (que mide una sola dimensión). Por ejemplo, si se quiere medir la inteligencia "de los números" se deben eliminar preguntas que tratan sobre conocimientos no relacionados con los números.
- crear una prueba con una cantidad de preguntas suficientemente elevada para obtener un error de medición reducido.

En el Capítulo 13, sección E.3, se ilustra la fórmula (Ebel, 1972) que permite calcular el Número de Preguntas de tipo Selección Múltiple (NPSM) necesarias para alcanzar una fiabilidad dada (*Reliability* ó r_{tt}), teniendo en cuenta el número (k) de soluciones que ofrecen las PSM:

$$N_{PSM} = (9/ (1-r_{tt}) \bullet (k+1) / (k-1)$$

N.3. Medir la fiabilidad con el Alpha de Cronbach

Conceptualmente, la replicabilidad o *reliability* se define como la correlación entre los resultados de dos aplicaciones del mismo test a un mismo grupo de estudiantes. De allí proviene la notación r_{tt}, donde r significa "correlación" y donde t significa "test".

Generalmente, en la práctica es imposible tomar dos veces el mismo test a los mismos estudiantes. Sin embargo hay varias maneras de estimar la fiabilidad. La más utilizada hoy en día es la de calcular el Alpha (α) de Cronbach (1951) según la fórmula:

$$\textit{Alpha de Cronbach} = \alpha = NP /(NP-1) \bullet (1 - (S\ Var\ it\ /\ VarT))$$

Donde:

S Var it es la *Suma* de las varianzas de cada uno de los ítems (NP o número de preguntas)

VarT es la varianza de los resultados en el Total del Test

Para los datos de la Tabla 25, en donde NS e = 13 y NP = 10, el α *de Cronbach* es igual a (10/(10-1)) • (1 – (2,17/4,84)) = *0,61*

Tabla 25: Ejemplo de datos para ilustrar el cálculo del *α de Cronbach*. La última línea contiene las varianzas de cada ítem, que sumadas equivalen a 2,17. La Varianza total de la prueba o test es 4,84

Rref	0,316												
NS=	13		Preguntas										
NP=	10		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Media
	NE	Nep	7	9	5	11	7	10	4	9	7	8	7,7
Estudiantes	NES	TE	0,538	0,692	0,385	0,846	0,538	0,769	0,308	0,692	0,538	0,615	0,59
E1	4	0,4	1	0	1	1	0	1	0	0	0	0	
E2	9	0,9	1	1	1	1	1	1	1	1	0	1	
E3	6	0,6	0	1	0	1	1	1	1	0	0	1	
E4	6	0,6	1	1	1	0	1	1	0	1	0	0	
E5	3	0,3	0	0	0	1	0	0	0	1	0	1	
E6	5	0,5	0	1	1	1	0	0	1	0	1	0	
E7	7	0,7	0	1	0	1	1	1	0	1	1	1	
E8	5	0,5	1	1	0	0	0	1	0	1	0	1	
E9	5	0,5	0	0	0	1	0	1	0	1	1	1	
E10	2	0,2	0	0	0	1	0	0	0	0	1	0	
E11	7	0,7	1	1	0	1	1	1	0	1	1	0	
E12	10	1	1	1	1	1	1	1	1	1	1	1	
E13	8	0,8	1	1	0	1	1	1	0	1	1	1	
VarT=	4,840	TEs	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	S Var it
Alfa =	0,61		0,249	0,213	0,237	0,130	0,249	0,178	0,213	0,213	0,249	0,237	2,17

El Alpha de Cronbach o medida de Consistencia interna *(Internal consistency) se interpreta así:*

$\alpha \geq 0,9$	Excelente
$0,9 > \alpha \geq 0,8$	Buena
$0,8 > \alpha \geq 0,7$	Aceptable
$0,7 > \alpha \geq 0,6$	Cuestionable
$0,6 > \alpha \geq 0,5$	Débil
$0,5 > \alpha$	Inaceptable… (sobre todo para un test psicológico).

O. Determinar límites de confianza[227]

O.1. El Error Estándar de Medición (EEM)

Caso 13

Un estudiante ha rendido dos pruebas cuyo máximo puntaje en cada una es 100. Estas pruebas tienen una fiabilidad (estimada a través del Alpha de Cronbach o α): de
Prueba A, α = 0,8, en la cual el estudiante ha obtenido un puntaje de 55.
Prueba B, α = 0,6, en la cual el estudiante ha obtenido también un puntaje de 55.
Las dos pruebas tienen una Desviación Estándar (DE) = 10.
¿Cuáles son los respectivos errores de medición de los 55 puntos obtenidos en cada prueba?

El Error Estándar de Medición (EEM) de la nota es $EEM = DE * \sqrt{1-r_{tt}}$
R_{tt} siendo la fidelidad o fiabilidad.

Con la prueba B, el EEM = $10 * \sqrt{1-0,6} = 10 * \sqrt{0,4} = 10 * 0,63 = 6,3$ puntos
Con la prueba A, el EEM = 4,4 puntos

O.2. Los Límites de Confianza (LC)

Asumimos que los errores de medición se distribuyen al azar (según una curva de Gauss). De esa manera, el valor más probable de la nota "verdadera" (y que no conocemos) sigue siendo la nota medida u observada. Sin embargo, podría tratarse de otro valor dentro de un cierto rango. La curva de Gauss nos permite saber cuál es la probabilidad de que "la nota verdadera" esté incluida en un intervalo alrededor de la nota observada o medida (Figura 21).

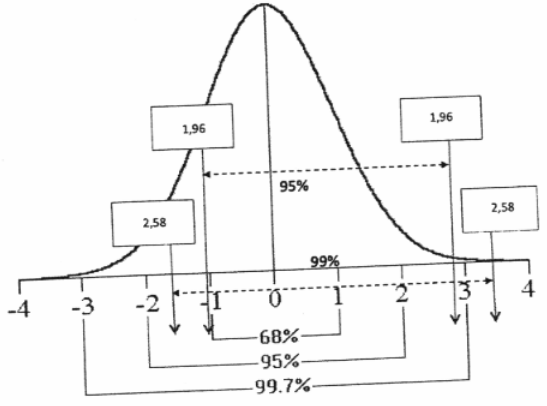


Figura 21: Probabilidades de incluir el valor verdadero en un intervalo de límites de confianza (LC) conociendo el error de medición

[227] Esta sección aborda un tema similar a lo ya visto en C.3.A y C.3.B.

El intervalo entre –1,96 • EEM (bajo la nota observada) y +1,96 • EEM (sobre la nota observada) indica una zona donde existe un 95% de probabilidad de incluir la nota verdadera (o un 5% de probabilidad de equivocarse (p.05)). Si el intervalo va desde –2,58 • EEM hasta +2,58 • EEM tendríamos un 99% de probabilidades de incluir allí la nota verdadera o, lo que es lo mismo, un 1% de probabilidad de equivocarnos (p.01). En las ciencias humanas habitualmente se usa como referencia p.05 para decidir. De esa manera y en nuestro ejemplo:

- Para la *prueba A*, el *LC* inferior es 55 – (4,4 * 1,96) = 55 – 8,6 = *46,4* puntos y el *LC* superior es 55 + 8,6 = *63,6* puntos
- Para la *prueba B*, el *LC* inferior es 55 – (6,3 * 1,96) = 55 – 12,3 = *42,7* puntos y el *LC* superior es 55 + 12,3 = *67,3* puntos

Finalmente, es posible decir que el verdadero puntaje del estudiante en la prueba A se ubica entre los 46,4 y los 63,6 puntos, siendo los 55 puntos la mejor estimación de su puntaje real.

P. Análisis de los ítems

P.1. La facilidad de la pregunta

En la literatura se usa la expresión *dificultad de una pregunta*, lo que constituye un hábito que deberíamos cambiar puesto que la relación entre el porcentaje de éxito y la dificultad es inversa. Mientras más alta es la tasa de éxito, más fácil es la pregunta. Si se usan grados de certeza la expresión *facilidad subjetiva* (certeza media que acompaña las respuestas a esa pregunta) pasa a ser un índice distinto de la *facilidad objetiva* (tasa o porcentaje de éxito), el que puede ser combinado con aquella. En efecto, la multiplicación de los dos términos expresaría la *superficie de éxito de la pregunta* (Leclercq, 2003, p. 47).

P.2. La capacidad de discriminación de un ítem para un grupo de estudiantes

Los índices de discriminación miden la capacidad de una pregunta, o bien de cada solución o alternativa propuesta en una PSM para separar (diferenciar) a los estudiantes de mayor capacidad de aquellos con menos capacidad frente a ese ítem. Hablaremos aquí de "ítem" para referirnos ya sea a la pregunta completa o a cada una de las alternativas que esta considera.

Se habla de *discriminación con criterio externo* si la mayor (o menor) capacidad de un estudiante es identificada a través de una medición en la cual el ítem (de una prueba escolar) no ha participado anteriormente. Por ejemplo, los resultados del año pasado medidos con otros ítems, etc. El índice de discriminación se calcula entonces vía fórmulas de correlación. A menudo, los docentes no conocen estos valores externos.

Se habla de *discriminación con criterio interno* si la mayor (o menor) capacidad de un estudiante es identificada a través del resultado (o nivel de éxito) de cada estudiante en el *total de la prueba*, incluyendo al ítem en cuestión. En este caso se habla de la homogeneidad o consistencia del ítem (*internal consistency*) con respecto a la prueba entera o, dicho de otra forma, se habla de la magnitud de la contribución del ítem para establecer el resultado total.

P.3. Los índices de discriminación: las Correlaciones punto biseriales (r_{pbis})

a) La palabra *biserial* significa que para cada ítem se comparan las notas (más exactamente las Medias de las notas en el total de la prueba) de aquellos estudiantes que han contestado el ítem de manera correcta ("1" o "logrado") con las notas de aquellos que no han contestado correctamente ("0" o "no logrado"). Consideremos en la Tabla 26 las dos opciones o modos de respuesta a la Pregunta 1 en una escala del 1 al 10 y donde cada X es un estudiante. El modo 1 (que aquí significa "éxito") y el modo de respuesta 0 (que aquí significa "fracaso u omisión").

En la Figura 22 se posicionan, sobre dos escalas verticales (equivalentes), los resultados en el total de la prueba de los siete estudiantes que no han logrado la Pregunta *1* (P1) (escala de la izquierda) y se calcula la Media de sus 7 notas totales. En este caso, la Media es de 4,86. A continuación se hace lo mismo en la escala de la derecha con los cinco estudiantes que respondieron correctamente la Pregunta 1. La Media de los resultados totales de este grupo es 6,60. Es posible apreciar que la pendiente que reúne las dos Medias sube, lo que significa que los (5) estudiantes que han respondido correctamente a la Pregunta 1 ("1") tienen mejores resultados en el total del test (es decir, tienen una Media superior en el total de la prueba) comparados con los (7) estudiantes que han respondido de manera errónea o han omitido ("0"). Lo anterior resulta esperable cuando "1" es éxito y "0" fracaso.

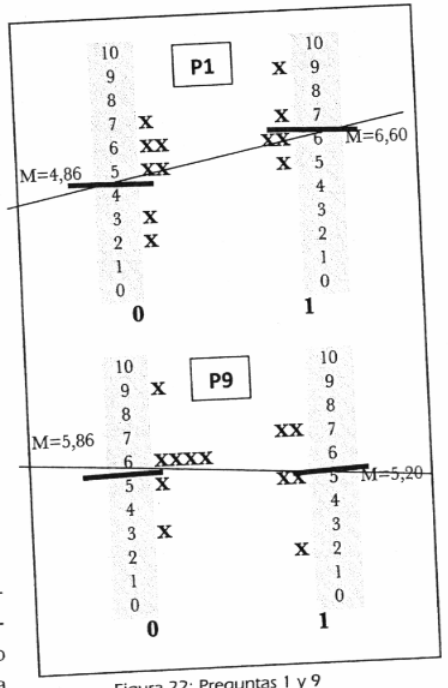


Figura 22: Preguntas 1 y 9

Si hacemos lo mismo con la *Pregunta 9* (P9) vemos que ocurre lo contrario: los cinco que han contestado correctamente la P9 tienen una nota

media en el total (5,20) inferior a la de aquellos que han contestado incorrectamente o no contestado (5,86), lo que parece "un contrasentido".

P.4. La fórmula del r_{pbis} (o correlación punto biserial)

$r_{pbis}\ (1) = (Mx\text{-}Ma) / DE_T . \sqrt{pq}$

En donde (1) señala una opción de respuesta, y donde:

p es la proporción de uso de esa opción de respuesta o modo de respuesta;
q es la proporción de NO uso de esa opción de respuesta *(q=1–p)*;
Mx es la Media de los resultados en el Total del test de los sujetos que responden "1";
Ma es la Media en el total de los que NO responden "1";
DE_T es la desviación estándar de la distribución de las notas en el total.

En la penúltima línea de la Tabla 26 aparecen las proporciones de éxito (1) o índice de facilidad para cada pregunta; y en la última línea aparecen los r_{pbis} (o coeficientes de discriminación) asociados a la opción u opción de respuesta *1*. En la Tabla 26 vemos que todos estos índices son positivos tal como esperado, con la excepción de las preguntas 4 y 9.

Tabla 26: r_{pbis} de las opciones 1 (éxito) para cada una de las diez preguntas de una prueba contestada por doce estudiantes

	RREF	PREGUNTAS									
	0,32	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*	*P8*	*P9*	*P10*
Estudiantes		5	9	4	10	5	9	5	8	5	7
E1	*5*	1	1	1	1	0	1	0	0	0	0
E2	*9*	1	1	1	1	1	1	1	1	0	1
E3	*6*	0	1	0	1	1	1	1	0	0	1
E4	*6*	1	1	1	0	1	1	0	1	0	0
E5	*3*	0	0	0	1	0	0	0	1	0	1
E6	*5*	0	1	1	1	0	0	1	0	1	0
E7	*7*	0	1	0	1	1	1	0	1	1	1
E8	*6*	1	1	0	0	0	1	1	1	0	1
E9	*5*	0	0	0	1	0	1	0	1	1	1
E10	*2*	0	0	0	1	0	0	0	0	1	0
E11	*7*	1	1	0	1	1	1	0	1	1	0
E12	*6*	0	1	0	1	0	1	1	1	0	1
		1	2	3	4	5	6	7	8	9	10
MEDIA	*5,58*	0,42	0,75	0,33	0,83	0,42	0,75	0,42	0,67	0,42	0,58
Dev. Est.	1,83	*0,47*	*0,71*	*0,26*	*–0,10*	*0,65*	*0,71*	*0,38*	*0,42*	*–0,18*	*0,27*

P.5. La correlación "de referencia"

Consideremos que a cada pregunta de una prueba se atribuya una nota 0 o 1. Si son dos preguntas en la prueba, la nota total puede variar de 0 a 2, y la pregunta 1 "pesa" la mitad en el puntaje total (el mismo peso de la pregunta 2). Si en una prueba hay 100 preguntas, cada una con el mismo peso, el puntaje de cada una tendrá un peso de 1/100 en la nota total. Se ve que mientras más grande el número de preguntas en una prueba, menor será el peso de cada pregunta en el puntaje total. Al contrario, cuando el número de preguntas es pequeño, existe una correlación "automática" (positiva en principio) entre las notas de cada pregunta (1 o 0) y la nota total, que puede ser importante. Por esa razón se calcula esta "correlación automática" (la misma para cada pregunta si todas las preguntas tienen el mismo peso). Si el RPBIS de la respuesta correcta en una pregunta es inferior a esta correlación de referencia, hay un problema con esta pregunta. Esta correlación (R) de referencia (ref) se calcula a través de la siguiente fórmula:

$$Rref = 1 / \sqrt{NP}$$

En nuestro ejemplo en cual NP = 10, $Rref = 1/\sqrt{10} = 1 / 3{,}16 = 0{,}31$.

Volviendo a la Tabla 26 vemos que las preguntas P3 y P10 poseen un $r_{pbis}(1)$ positivo pero que se ubica bajo el valor de 0,31. Así, la correlación de referencia establece un umbral para calificar el aporte real de cada ítem al total del test.

P.6. Las interpretaciones de un análisis de ítems

Dejamos al lector el trabajo de interpretar una parte de los datos entregados por el dispositivo SMART de la ULg (los primeros 8 ítems; ver Capítulo 22) y extraídos de una prueba construida con PSM y SGI[228]. Ello sabiendo que:

- en gris se identifican los índices de la solución correcta;
- la primera línea de cada pregunta indica el % de elección de cada ítem, siendo el total 100%;
- la segunda línea indica los r_{pbis} de cada alternativa u opción de respuesta;
- la tercera línea indica la certeza media (en %) de los que han elegido cada solución.

[228] Aquí, las 4 SGI son numeradas como sigue: Solución (6) Ninguna, (7) Todas, (8) Faltan datos y (9) Absurdo (ver Capítulo 13).

Tabla 27: Análisis del SMART–ULg para las respuestas de una cohorte de estudiantes de primer año de universidad en 2008 que responden a las ocho primeras PSM (con cuatro Soluciones Generales Implícitas: 6, 7, 8 y 9) en una prueba de treinta preguntas referidas al curso "Introduction aux Sciences de l'Éducation" (ISE) de D. Leclercq.

Sistema Metodológico de Ayuda para la Realización de Tests SMART Responsable académico: Dieudonné Leclercq Director: Jean-Luc Gilles ESTADÍSTICAS DE PREGUNTAS A0810018 ISE formativo ocubre										
	SOL 0	SOL 1	SOL 2	SOL 3	SOL 4	SOL 5	SOL 6	SOL 7	SOL 8	SOL 9
P1	0,65	19,93	49,35	19,93			4,25	0,33	2,61	2,94
Rbis	–0,12	0,40	–0,07	–0,32			0,01	0,04	0,07	–0,03
Cert. Media	55,47	69,18	57,48	55,74			43,08	60,00	42,50	48,89
P2	1,31	10,46	81,70	2,61			0,98	0,00	0,33	2,61
Rbis	–0,07	–0,27	0,35	–0,13			–0,15	0,00	0,04	–0,06
Cert. Media	55,00	58,75	66,08	70,00			33,33	0,00	40,00	35,00
P3	1,31	7,84	0,98	2,94			7,19	0,00	1,63	77,78
Rbis	–0,10	–0,14	–0,12	–0,22			–0,30	0,00	–0,09	0,46
Cert. Media	50,48	53,33	40,00	51,11			50,00	0,00	44,00	73,36
P4	2,61	2,29	13,40	17,65			53,27	2,94	3,92	3,59
Rbis	–0,23	–0,07	–0,13	–0,10			0,40	–0,13	–0,04	–0,19
Cert. Media	41,64	25,71	43,41	44,44			54,72	57,78	33,33	27,27
P5	1,96	0,00	0,98	20,59			68,30	2,29	1,96	3,27
Rbis	–0,12	0,00	–0,05	–0,27			0,45	–0,15	–0,18	–0,19
Cert. Media	55,28	0,00	46,67	54,92			63,73	68,57	40,00	60,00
P6	0,65	0,65	0,65	5,56	0,65		4,90	81,37	4,90	0,65
Rbis	–0,12	–0,13	–0,14	–0,16	0,03		–0,07	0,31	–0,15	0,10
Cert. Media	55,64	50,00	20,00	56,47	90,00		57,33	74,22	54,67	50,00
P7	0,00	0,00	5,23	0,98			74,84	0,65	3,59	14,71
Rbis	0,00	0,00	–0,14	–0,18			0,33	–0,09	–0,17	–0,16
Cert. Media	55,84	0,00	55,00	53,33			69,43	40,00	56,36	56,89
P8	0,65	1,31	32,03	3,92			3,27	50,00	0,33	8,50
Rbis	–0,12	–0,21	–0,11	–0,04			–0,01	0,27	–0,01	–0,14
Cert. Media	56,16	45,00	57,76	40,00			42,00	59,61	0,00	66,92

Analicemos en conjunto la Tabla 27.

Primero, focalicemos solo sobre las *soluciones correctas* (solo una por pregunta), que están en un cuadro con fondo gris.

a) Consideremos, en las primeras líneas de estos cuadros grises, las tasas de elección de estas soluciones, es decir, las tasas de éxito de las preguntas (siendo el mínimo posible 0%, y el máximo posible 100%). Se ve que la pregunta más fácil es la P2 (81,70% de éxito) y la de mayor dificultad es la P1 (19,93% de éxito).

b) Consideremos ahora, en la segunda línea de cada cuadro en gris, los r_{pbis} (o índices de discriminación) de estas soluciones correctas para cada una de las ocho preguntas. Allí aparece que todos *son positivos*, lo que era esperable.

c) Preguntémonos ahora si estos r_{pbis} son iguales o superiores a la correlación de referencia (aquí 0,31). Podemos ver que para todas las preguntas sí son superiores, excepto en la pregunta 8, que tiene un r_{pbis} de 0,27 (lo que de todas formas está cerca de 0,31). Eso significa que la prueba es homogénea y que cada pregunta contribuye al total más o menos del mismo modo. De esa manera podemos ver que la pregunta 3 (r_{pbis} = 0,46) es la que más contribuye (es decir, la que está más correlacionada con el total) mientras que la pregunta 8 es la que menos lo hace.

Segundo, interesémonos en las *soluciones incorrectas* (es decir, en los cuadros que no son grises).

a) En la primera línea, que presenta las *tasas de elección de cada solución*. Allí se ve que:
   - Hay soluciones que no fueron elegidas por nadie. Es el caso de las soluciones 7 de las preguntas 2 y 3; de la solución 1 de la pregunta 5; etc.
   - Hay soluciones que fueron elegidas por una gran tasa de estudiantes (aunque son escasas). Es el caso de la solución 2 (incorrecta) de la pregunta 8, solución que fue elegida por 32,03% de los estudiantes.

b) Considerando *los r_{pbis} de las soluciones incorrectas*, se ve que:
   - En P3, P4, P5, P6, P7 y P8, todos son negativos, lo que también era esperado.
   - En las Preguntas P1 y P2, cuando las soluciones erróneas o falsas tienen un r_{pbis} positivo (lo que es contrario a lo que se espera) se aprecia que esos r_{pbis} son bajos en valor absoluto (cercanos a 0) lo que es deseable.

Tercero, interesémonos en las *certezas medias*, ya sean aquellas de las soluciones correctas o incorrectas. Así, mirando la tercera línea en cada pregunta, se ve que:
   - Excepto por la P5 y la P8, la Certeza media (Cert. Med.) de la respuesta correcta es más alta que la Certeza media de los distractores (lo que es esperado).
   - NOTA: la certeza media que aparece en la columna de la Omisión (solución "0") es la Media de las certezas de los distractores (no había otro lugar en la página para imprimirla). Por supuesto, aquí no se incluye la certeza de aquellos que omiten, puesto que no han dado certeza ni respuesta alguna.

¿Para qué sirve este tipo de análisis?

1. Ayuda al profesor a detectar preguntas y soluciones que deben ser mejoradas para un uso futuro. NOTA: Cuando una pregunta no está correlacionada con el puntaje total a la prueba (el r_{pbis} de la respuesta correcta es inferior a la correlación de referencia) ello puede significar: o que esta pregunta está mal formulada (los estu-

diantes con mejor rendimiento se han equivocado), o que ella no mide las mismas características que las otras y por lo tanto no contribuye a la homogeneidad de la prueba.

Importante: En las evaluaciones de un curso es normal tener pruebas heterogéneas, con preguntas que buscan medir diferentes aprendizajes esperados en ese curso. Por eso, los r_{pbis} pueden no siempre ser más altos que la correlación de referencia (incluso pueden ser un poco inferiores). Es el profesor, con toda esta información y buscando mejorar su curso y evaluaciones, quien debe juzgar sobre la utilidad de cada pregunta o ítem.

2. Ayuda al profesor a detectar las soluciones erróneas que han atraído a muchos estudiantes, en especial a aquellos que han tenido un puntaje total mejor que el de los demás, para luego intentar comprender el porqué de ese fenómeno.
3. Ayuda al profesor a decidir cuáles preguntas deben ser eliminadas (después de la prueba, pero antes de asignar la calificación) para poder así calcular el puntaje final de cada estudiante.
4. Por último, es un insumo para el momento del debate de retroalimentación después de la prueba (ver Capítulo 9). En este debate los estudiantes que han elegido otras alternativas de respuesta (no consideradas como correctas por el profesor al momento de concebir la prueba) pueden exponer sus argumentos e intentar convencer al docente para que acepte estas alternativas también como correctas. Datos como los de los análisis de la Tabla 27 pueden ayudar al profesor a redefinir las alternativas consideradas correctas, y volver a calcular los puntajes incorporando estos cambios.

Invitamos al lector a *hacer el mismo trabajo de análisis* con otra serie de datos de las ocho primeras preguntas de otra prueba, presentada en la Tabla 28.

Existen dos (ligeras) diferencias de presentación entre el ejemplo de la Tabla 27 y el de la Tabla 28. En la Tabla 28…

1) … cuando las respuestas de los estudiantes han sido aceptadas como correctas, se indica con el código xxx (ejemplo: la alternativa 7 en la pregunta 5);
2) … las respuestas correctas están enmarcadas y las que tienen valores de r_{pbis} problemáticos están sombreadas.

Por ejemplo, la solución 7 de la pregunta 2 tiene un r_{pbis} de 0,18, inferior a la Correlación de Referencia calculada (0,31)[229]. Sin embargo, y a pesar de tener un r_{pbis} bajo ese umbral, debemos considerar que se trata de la solución correcta. El mismo fenómeno ocurre con la solución 2 de la pregunta 8. Es necesario entonces que el profesor reflexione sobre cómo mejorar la pregunta/ítem.

[229] Notemos que si la prueba tuviera 20 ítems la correlación de referencia sería 0,224.

Tabla 28: Análisis del SMART–ULg y las respuestas de una cohorte de estudiantes de primer año de universidad en 2009 a las ocho primeras PSM (con cuatro Soluciones Generales Implícitas: 6, 7, 8 y 9) de una prueba (de veinte preguntas) del curso "Introduction aux Sciences de l'Éducation" (ISE) de D. Leclercq

	SOL 0	SOL 1	SOL 2	SOL 3	SOL 4	SOL 5	SOL 6	SOL 7	SOL 8	SOL 9
P1	1,4	0,7	66,4	1,4			4,5	3,4	3,1	18,8
Rbis	−0,05	−0,02	−0,28	0,03			0,02	0,02	−0,08	0,38
Cert. Media	60,69	50,00	62,99	55,00				62,00	55,56	70,91
P2	1,4	0,3	37,3	7,9			9,9	36,6	2,4	4,1
Rbis	0,06	−0,12	−0,03	−0,21			−0,00	0,18	−0,03	−0,06
Cert. Media	53,81	60,00	57,25	48,70			46,21	62,99	40,00	58,33
P3	1,7	22,6	5,5				1,4	22,3	40,8	5,8
Rbis	−0,02	−0,22	−0,16				−0,02	−0,20	0,47	−0,05
Cert. Media	59,40	56,67	45,00				40,00	69,23	74,12	50,59
P4	3,1	18,8	21,6	4,1	4,5		29,8	10,3	4,5	3,4
Rbis	−0,03	−0,08	−0,06	−0,06	−0,02		0,34	−0,17	−0,13	0,01
Cert. Media	47,65	48,00	50,16	48,33	33,85		56,55	46,00	47,69	52,00
P5	2,1	30,5	7,2	0,7			1,0	0,0	52,1	6,5
Rbis	−0,09	−0,44	−0,14	−0,01			−0,08	XXX	0,53	−0,02
Cert. Media	55,67	58,43	44,76	40,00			46,67	0,00	71,84	57,89
P6	1,0	88,7	5,8	1,7			1,0	0,3	0,0	1,4
Rbis	−0,05	0,22	−0,08	−0,14			−0,15	−0,10	XXX	−0,07
Cert. Media	37,33	75,68	48,24	32,00			20,00	0,00	0,00	20,00
P7	3,8	14,0	39,7	27,1			4,5	1,4	6,8	2,7
Rbis	−0,03	0,34	0,01	−0,25			0,08	−0,11	−0,07	0,05
Cert. Media	48,58	61,95	46,55	54,68			41,54	35,00	48,00	37,50
P8	3,1	9,9	66,8	15,8			3,1	0,0	0,3	1,0
Rbis	−0,05	−0,03	0,14	−0,11			−0,07	XXX	0,04	0,01
Cert. Media	46,14	49,66	63,90	43,91			40,00	0,00	100,00	46,67

P.7. Las causas de las anomalías

Cuando los valores de los índices no corresponden a la teoría estos constituyen una señal pero no logran indicarnos la causa del problema. Es el docente quien siempre debe investigar para saber el origen o la causa de las anomalías que muestran los índices. Para ello puede preguntarse…

- … si la prueba ha sido mal formulada (palabras ambiguas, un distractor inapropiado, etc.)
- … si el concepto está mal explicado en el libro/sitio de referencia
- … si el concepto ha sido mal explicado en el aula, durante el curso presencial.

Parte 6: Correlaciones y dispersión

Q. La interacción entre variables

Elegir un tipo de gráfico puede facilitar la interpretación de los datos y de las relaciones entre datos. Por ejemplo, las figuras 23 (ya presentada en C.5) y 24 muestran los mismos datos revelando la existencia de una interacción entre las variables tipo de examen y tipo de ayuda. En términos generales, una interacción entre dos variables independientes significa que el efecto de una de ellas varía según la categoría o valor que adopta la otra.

En este caso, la Figura 24 muestra claramente una "interacción cruzada" en donde se aprecia que el efecto de la variable "tipo de ayuda" sobre la medida de "experiencia en hipermedia (ámbito D)" varía según la categoría que adopta la variable "tipo de examen (PSM u Oral)". Lo anterior podría leerse como: el efecto de la "ayuda PSM" será tanto más fuerte –aumentando su impacto en el valor de la "experiencia en hipermedia (ámbito D)"– si el examen usado es de tipo PSM (en vez de Oral). Corresponde al investigador decidir qué tipo de gráfico prefiere usar para representar el problema.

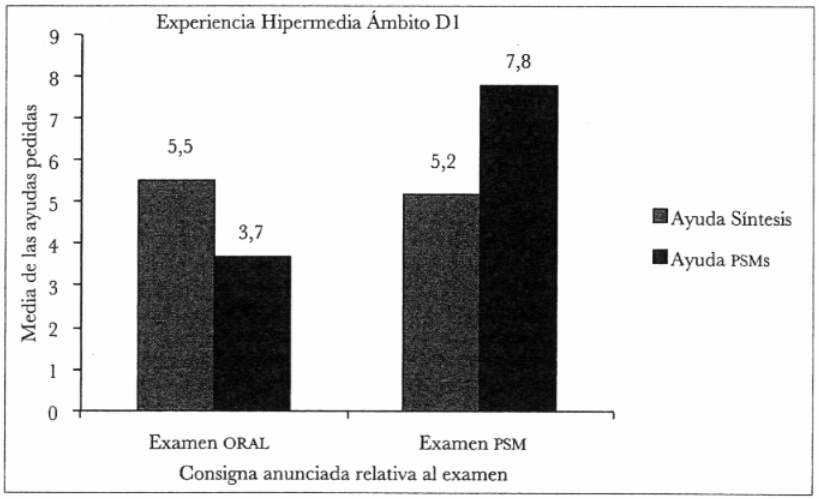


Figura 23: Interacción representada con barras

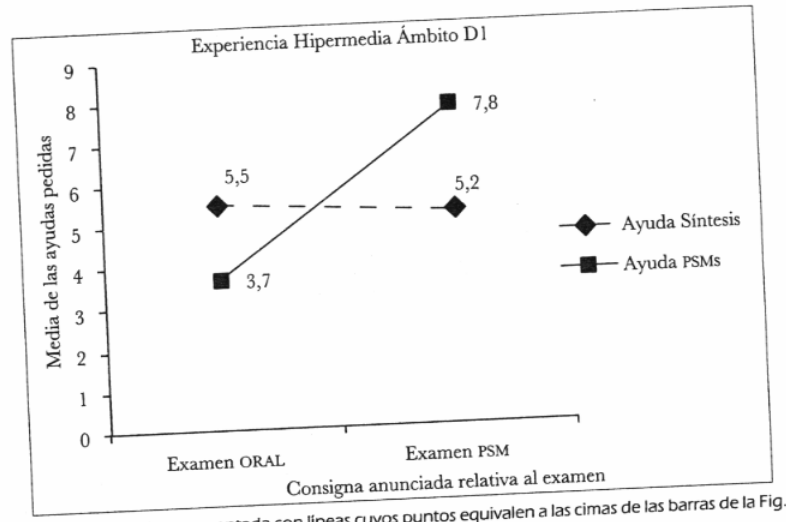


Figura 24: Interacción representada con líneas cuyos puntos equivalen a las cimas de las barras de la Fig. 23

R. El valor indagatorio de las nubes de puntos de las correlaciones

Ya hemos insistido sobre el hecho de que los docentes pueden practicar el principio de la investigación-acción: su primera misión es enseñar. Eso incluye la evaluación, la que, a su vez, les permite tener datos útiles para investigar. Así, por ejemplo, las relaciones entre dos mediciones pueden servir de instrumento para descubrir fenómenos (investigación exploratoria) o para testear hipótesis (investigación confirmatoria).

Habitualmente, calcular el valor de una correlación no es suficiente para estudiar una relación. Visualizar la nube de puntos (lo que Excel puede hacer casi automáticamente) es muy útil, tal como se verá en el siguiente ejemplo. En 1994 Leclercq había sugerido a una de sus estudiantes, F. Lardenoye, visualizar la nube de puntos de la correlación entre el nivel de competencia al inicio del curso de español para nueve estudiantes y el beneficio (expresado en GR) derivado de la utilización de un Hipermedia en el cual estos pudieron explorar lo que necesitaban saber para mejorar su conocimiento. Dos hipótesis estaban en oposición:

1) Mientras más sabe un estudiante al inicio sobre un tema, más aprenderá. Dicho de otra manera, su GR será tanto más elevada en tanto el tema no es totalmente desconocido para ese estudiante, lo que se debería a que posee una cierta "base" o "pre-condiciones" que facilitan la adquisición del nuevo conocimiento.
2) Mientras menos sabe un estudiante sobre un tema, más será su GR al aprenderlo.

Aunque los datos de esta investigación-acción eran las mediciones PRE y POST sobre nueve contenidos para cada estudiante, se calculó la correlación interna por cada uno

y se visualizó su respectiva nube de puntos. La nube que sigue corresponde a un estudiante con solo nueve puntos. En la Figura 25 se ha pedido a Excel calcular la correlación lineal y dibujar la recta que pasa por el centro de la nube. Viendo la forma de la nube, se le pidió a Excel calcular también la *correlación curvilineal*, la cual dibuja la curva de la ecuación (Figura 26).

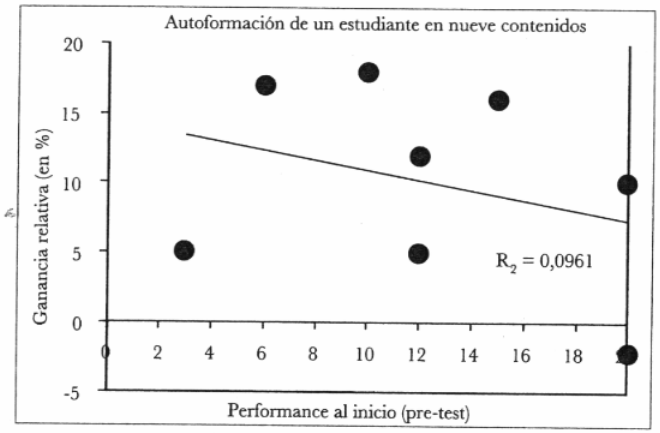


Figura 25: Correlación lineal y recta que la representa

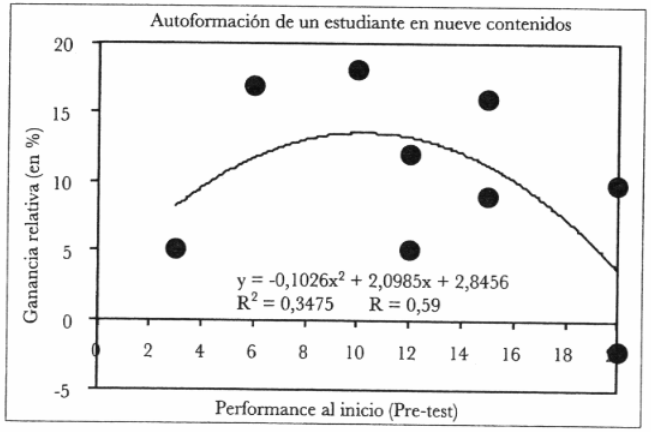


Figura 26: Correlación curvilineal y curva que la representa

Se puede apreciar que la correlación lineal (Figura 25) es tan baja y negativa (–0,1) que es casi inexistente. Al contrario, la correlación curvilineal (Figura 26), para la misma nube, es más alta (0,59). Durante una fase exploratoria esta observación indujo una hipótesis nueva: que la relación no es lineal sino que es curvilineal dado que la relación existente quedaría mejor expresada por esa segunda opción. Esta hipótesis fue testeada por otra estudiante, Jans (1999), esta vez buscando descartarla o confirmarla (investigación confirmatoria) a través de una prueba de 100 preguntas (en inglés), incluyendo en ella grados de certeza.

En las figuras 27 y 28 obtenidas por Jans (1999) se puede apreciar que:

(a) en el eje horizontal se ve la performance al inicio (de Base). Como se utilizaron grados de certeza, dicha performance en cada uno de los 100 ítems puede variar entre –100 (error con certeza 100) y 100 (éxito con certeza 100);
(b) en el eje vertical se aprecia la Ganancia Relativa (GR) que varía de 100% (ganancia máxima) a –100% (pérdida máxima);
(c) los 100 puntos (un punto por ítem) no aparecen en el gráfico porque muchos de ellos se superponen. En este caso, el cálculo de la correlación es más fiable que la inspección visual;
(d) se trata de dos gráficos diferentes para el mismo estudiante: en la Figura 27 las Ganancias Relativas (eje vertical) son medidas con el POST-test inmediato, mientras que en la Figura 28 las GR son medidas con el POST-test diferido (tras unas semanas).

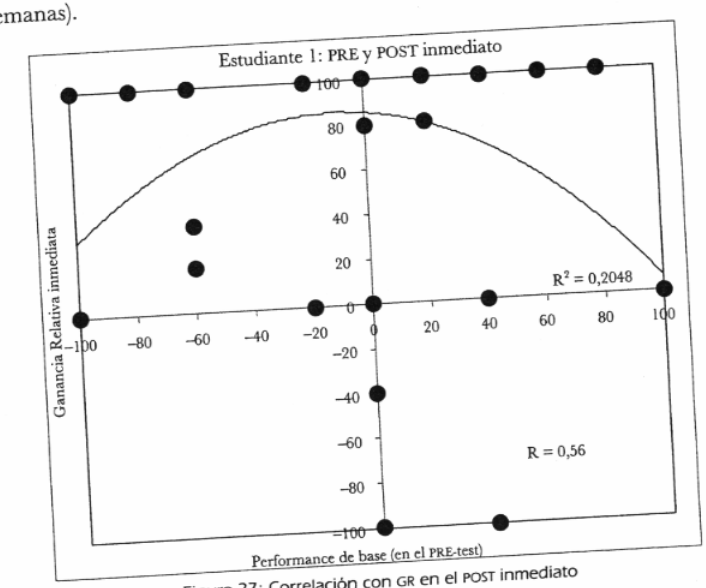


Figura 27: Correlación con GR en el POST inmediato

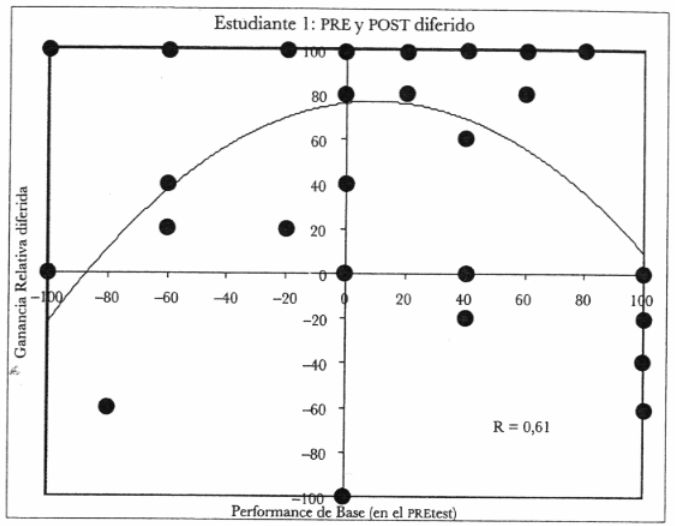


Figura 28: Correlación con GR en el POST diferido

El análisis de estos gráficos permite concluir una confirmación de la hipótesis de una relación curvilínea entre los conocimientos de base y la Ganancia Relativa (GR) generada por un proceso de aprendizaje.

Los conocimientos de base van, de izquierda a derecha, desde los errores con alto grado de certeza hasta las respuestas correctas con alto grado de certeza, y la Ganancia Relativa es medida a través de un test *a posteriori* (POST-test). En términos concretos, eso significa que:

- Cuando el estudiante "no sabe nada" previamente (posición central en el eje horizontal) su Ganancia Relativa es más amplia.
- Cuando en sus conocimientos previos el estudiante se equivoca con alto grado de certeza (en el extremo izquierdo del eje horizontal) su Ganancia Relativa es baja. Esto puede explicarse por la siguiente hipótesis, aplicable *en este caso específico*: en caso de *misconception* (concepto erróneo que se cree correcto) la "remediación" que se autoadministró este estudiante (por ejemplo, consultando algunas fuentes) no fue lo suficientemente fuerte como para producir una modificación amplia (considerando que la Ganancia Relativa posible es de 200%). Una confrontación con la Ganancia (simple) permite, en cada caso, reforzar o debilitar este tipo de hipótesis.
- Cuando al inicio el estudiante ya domina muy bien el contenido (posición en el extremo derecho del eje horizontal), su Ganancia Relativa también es baja, e incluso puede ser negativa: la segunda vez que se mide la misma competencia (POST-test) no puede aparecer mejor que la primera vez (pues ya había alcanzado el máximo). Solo puede tener el mismo puntaje o un puntaje ligeramente menor.

Esta relación es la que se dibuja en las "U" invertidas que muestran las figuras 27 y 28. Es interesante notar que dicha relación no solo se confirma en el POST-test diferido (Figura 28), sino que aparece con mayor fuerza. En definitiva, según este análisis, los estudiantes con menos conocimientos y con baja certeza en relación a estos, tiene mayor probabilidad de lograr una mayor Ganancia Relativa (GR).

Conclusión general del capítulo

La medición es una herramienta al servicio de la acción y de la reflexión. Por lo mismo, las medidas deben ser válidas, fiables, finas, fecundas... para poder aplacar nuestra sed de actuar y de comprender.

Nuestras horas son minutos cuando esperamos saber
Y siglos cuando sabemos lo que se puede aprender.

Más importantes que las medidas son los objetivos de nuestras acciones y las hipótesis de nuestras investigaciones, así como los retos sociales e individuales que ellas nos plantean.

Sabemos que los vasos sirven para beber.
Lo malo es que no sabemos para qué sirve la sed.

Proverbios y Cantares, Antonio Machado (1875-1939)

Referencias

Bangert-Drowns, R.L., Kulik, J.A., Kulik, C.L.(1983). Effects of coaching Programs on Achievement Test performance. A meta-analysis. Review of Educational research. 53, N°4, 571-585.

Bangert-Drowns, R.L., Kulik, J. A., yKulik, C.-L.C. (1985). Effectiveness of computer-based education in secondary schools. Journal of Computer-Based Instruction, 12, 59-68.

Binet A. y Simon Th. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. In: L'année psychologique. Vol. 11. pp. 191-244

Camus, A. (1944) Sur une philosophie de l'expression. Essai in Poésie 44. , et Œuvres complètes, tome 1, Paris, La Pléiade, p.908)

Carraher Nunnes, T., Carraher, D.W., Schliemann, A.D. (1985).Mathematics in the streets and in schools. British Journal of Developmental Psychology, 3, 21-29.

Carver, R. P. (1974). Two dimensions of tests: Psychometric and Edumetric. American Psychologist, 29, 152-518.

Cohen, J. (1994). The earth is round (p<.05), American Psychologist, 49, 997-1003.
Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, vol. 16, no 3, 1951, p. 297-334.
Cross, L. yFrary, (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple choice tests, Journal of Educational Measurement, vol. 14, 313-321.
Doane, D. y Seward, L. (2011). Measuring skewness: a forgotten statistic? Journal of Statistics Education. Vol. 19 (2), pp. 1-18.
Ebel, R. (1969). Expected reliability as a function of choices per item, Ed. yPsych. Meas., 29, 565-570.
Glass, G.V. (1976). Primary, secondary, and meta-analysis of research, Educational researcher, 5, 3-8.
Glass, G., McGaw, B. y Smith, M. (1981). Meta-analysis in social research, Beverly Hills : Sage Publ.
Hartley, S. (1977).Meta-analysis of the effects of individually paced instruction in mathematics. Doctoral Dissertation. University of Colorado.
Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators.Journal of Educational Statistics, 6, 107-128.
Hovland, C., Lumsdaine A., y Sheffield, F.D. Experiments on Mass Communication. Princeton: Princeton University Press. 1949
Jans, V. (1998), Self-Directed Learning of University Students Using a Hypermedia on English : Spectral Analysis of Their Performances, in T. Ottmann y I. Tomek (Eds), ed-media y ed-telecom 98, 10th World Conf. on Educ. mm, aace, Freiburg, june 20-25, pp. 688-693.
Kirk, R. (1996), Practical significance : a concept whose time has come, EdyPsychMeas.,56,5,746-759.
Kulik, J.A., Kulik, L.C. y Cohen, P.A. (1979). A meta-analysis of outcome studies of Keller's personalized system of instruction, American Psychologist, 1979, 38, 307-318.
Kulik, J. (1981). Integrating findings for different levels of instruction. Paperat the aera. Los Angeles.
Lardenoye, F. (1994), Expérimentation d'olafo, un hypermédia d'entraînement à l'espagnol, mémoire de licence en Philosophie et Lettres, Université de Liège.
Leclercq, D y Poumay, M. (1999). De l'apprentissage par les médias à l'éducation aux médias. Colloque du Conseil d'Education aux Médias. Bruxelles Décembre 1999
Leclercq, D. (2003). Connaissance partielle, analyse spectrale et métacognition. Chap 1 de D. Leclercq, (Ed). Diagnostic cognitif et métacognitif au seuil de l'université. Ed. Univ. de Liège. p. 33-49.
Leclercq y Poumay (2007). La métacognition. In D. Leclercq Méthodes de Formation et Théories de l'Apprentissage. Editions de l'université de Liège.
Leclercq, D. (2012). Validación de instrumentos y levantamiento de sistema de evaluación de aprendizajes progresivos. Informe de la asesoría en la Facultad de Educación de la U. C. Temuco (Chile) en octubre 2012.
Leclercq, D. Laguesse, C. y Henrotay, P. (2012).). Metacomprehension and cognitive vigilance. A multi-teachers experiment in last year of secondary school. Conference of the earli sig "Metacognition". UCSC Milano.
Lord, F. y Novick, M. (1968). Statistical theories of Mental Tests Scores. Reading (Mass.).Addison-Wesley.
Ndabawarukanye, C. (2004). Evaluation des connaissances sur la «chaîne de survie» dans un établissement de prise en charge de pathologies neurologiques. Mémoire de Licence en Santé publique. Université de Liège.
Perry, W.G. (1970). Forms of intellectual and ethical development in the college years: A scheme. New York: Holt, Rinehart y Winston
Perry, W.G. (1985). Different worlds in the same classroom: Students' evolution in their vision of knowledge and their expectations of teachers. In Gullette, M.M. (Ed.), On teaching and learning. Volume 1, 1-17. Cambridge, MA: Harvard-Danforth ctyl.
Pikler, E. (1979). Se mouvoir en liberté dès le premier âge, Paris : PUF.
Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests, Copenhagen :Danmark Paedagogik Institut.

Rosnow, R y Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science, American Psychologist, 44, 1276-1284.
Rotter, J. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 80 (1, whole, n° 609).
Ters, F., Mayer, G. y Reichenbach, D. (1973). L'Echelle Dubois-Buysse d'orthographe usuelle française. Paris : UCDL.
Thomas F. (1995). Télécole, Bruxelles: Fondation Roi Baudouin.
Tukey, J.W. (1991). The philosophy of multiple comparisons. Statistical Science, 6, 100-116.
Verstegen, C. (2009). Miroir, mon beau miroir, dis-moi si j'étudie bien. Impact subjectif d'un entraînement à la metacognition par tests spectraux dans un cours du premier semestre de la première année à l'université. Mémoire de Master en Sciences de l'Education. Université de Liège.

IDEAS E INNOVACIONES Dispositivos de Evaluación de los Aprendizajes en la educación

Dieudonné LECLERCQ y Álvaro CABRERA MARAY 2014

Resumen de cada capitulo

Parte 2: Dispositivos de Evaluación de los Aprendizajes (DEA- para la evaluación de desempeños complejos

Parte 3: Evaluación de saberes o recursos aislables

p. 259

Parte 4: Principios estratégicos en evaluación

Se puede bajar gratuitamente via http://hdl.handle.net/2268/180060