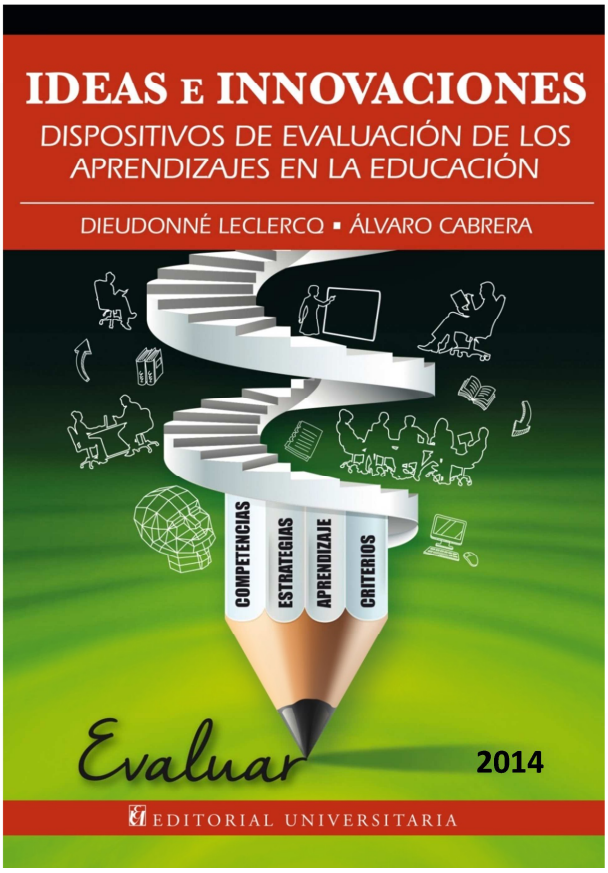

Dieudo LECLERCQ

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.

Capítulo IV

ETICPRAD. Ocho criterios de validez de un Dispositivo de Evaluación de los Aprendizajes (DEA)

Dieudonné Leclercq

Las secciones A y B de este texto fueron publicadas en Leclercq D. (2006) "L'évolution des QCM". En Figari, G. y Mottier-Lopez, L. *Recherches sur l'évaluation en Education.* Paris: L'Harmattan, 139-146.

Este capítulo tiene una doble intención. En primer lugar, presentar el resumen que consideramos más completo de los criterios de calidad de un Dispositivo de Evaluación de los Aprendizajes (DEA): el modelo ETICPRAD (Leclercq, 2006). Esta sigla reúne las ocho dimensiones de la validez de un DEA, según presentamos más adelante en la sección A. La segunda intención es hacer perder su inocencia[19] al lector respecto a las Preguntas de Selección Múltiple (PSM): no existe una única forma de plantear PSM. Lo anterior será ilustrado al mostrar que la evolución de las PSM es una sucesión de modificaciones que han intentado satisfacer los criterios de calidad de un DEA (y lo siguen haciendo). A continuación se presentan los 8 criterios, enumerados, definidos, marcados con un asterisco (*) y con la primera letra en mayúscula (formando el acrónimo ETICPRAD).

A. Criterios de validez de un Dispositivo de Evaluación de los Aprendizajes

A.1. La validez *Ecológica** (Brunswick, 1943) de un DEA, o "validez aparente" (en inglés *face validity*), es mayor en la medida en que la situación de evaluación corresponde a una situación de la vida real. Se supone que la evaluación es capaz de representar o predecir las condiciones de una situación real.

A.2. La validez *Teórica** (Cronbach y Meehl, 1955) se descompone en validez de contenido (o de "cobertura": ¿se ha testeado todo lo que debía evaluarse?) y en validez de *constructo* (¿está el DEA basado en un modelo creíble, científicamente fundado, de, por ejemplo, los procesos mentales?).

A.3. La validez *Informativa** (diagnóstica, en lo posible) es la multiplicidad de informaciones que resultan de la evaluación, sus especificidades (apuntar sobre una

[19] En el sentido que B. Bloom (1972) daba a esta expresión: nunca más poder decir "no sabía". Una persona formada debe declararse culpable.

capacidad y no sobre la vecina), sus niveles de precisión (sensibilidad), su valor explicativo.

A.4. La validez *Consecuencial** (Green, 1998) se aprecia en las consecuencias del DEA sobre las representaciones, sobre los actos (por ejemplo: revisar o no el contenido, cambiar o no su método de estudio) de los aprendices, de los docentes y/o del sistema.

A.5. La validez *Predictiva** de las medidas obtenidas es su capacidad de predecir eficazmente (es decir, con precisión y exactitud) otras medidas, a menudo posteriores, como por ejemplo el éxito escolar o profesional, el nivel de logro en otra prueba, etc.

A.6. La *Replicabilidad** o *Fiabilidad* (en inglés *reliability*, en francés *fidélité*) de una medida es su estabilidad en el tiempo o entre varios correctores. Una fórmula (Ebel, 1969) precisa el número de preguntas de una prueba, y, con Preguntas de Selección Múltiple, el número de soluciones propuestas (correctas + distractores) necesario para obtener un nivel de replicabilidad dado de la medida (0,8 por ejemplo). Otras fórmulas responden a lo mismo de la manera inversa: ¿Cuál debe ser el coeficiente *n* de longitud (extensión) de la prueba para alcanzar una replicabilidad deseada (por ejemplo 0,80 o 0,90)?

A.7. La *Aceptabilidad** o praticabilidad del DEA, para el docente, se refiere a la adhesión a los principios en los que se basa la evaluación y la aplicabilidad (viabilidad) de los métodos (duración, materiales y lugares requeridos, precauciones antifraude, etc.).

Para el estudiante, la Aceptabilidad* concierne a la adhesión y/o a la familiaridad con el tipo de prueba. Ha sido demostrado (Leclercq, 1986) que mientras más familiarizado está el/la estudiante con los procesos de *testing* y con las escalas de calificación, más avezado/a es en los tests (en inglés *test wiseness*), y más altas son sus probabilidades de éxito, especialmente con las Preguntas de Selección Múltiple.

A.8. La validez *Deontológica** (o ética) se presenta en varias formas, siendo la equidad la más conocida. Desde hace tiempo, la docimología[20] crítica (Piéron, 1963) ha demostrado que la calificación (por parte de jueces) de hojas de respuesta no solo son no-concordantes inter-jueces, sino que son no-concordantes intra-juez. Las pruebas con corrección objetiva evitan otros efectos deplorables (de contraste, de severidad de juez, de halo, el efecto Posthumus, etc.), pero presentan otras limitaciones. Los derechos de los estudiantes son cada vez más y más reconocidos, y los sistemas de evaluación deben garantizar más y más la transparencia del DEA, por ejemplo en términos de re-calculabilidad de los puntos desde la hoja "en bruto".

[20] Ciencia de los exámenes.

B. Algunos momentos claves en la historia de las Preguntas de Selección Múltiple interpretados en términos de ETICPRAD

B.1. La gloria de la consigna clásica desde su nacimiento

Frederick J. Kelly (1915) desarrolló la primera versión de un test constituido por Preguntas de Selección Múltiple (PSM): el Kansas Silent Reading Test. Apurados por seleccionar a los oficiales entre los reclutas para la guerra de 1914-1918, EE.UU. depositó su confianza en los *Army tests* concebidos por Otis. Esos tests de inteligencia[21], administrados a 1,7 millones de personas, eran en forma de PSM operando con la consigna clásica: "Una sola de las soluciones propuestas es correcta y se puede dar una sola respuesta". El que EE.UU. resultara estar entre los vencedores de la guerra dio una gran credibilidad a este tipo de test (Validez Predictiva*). Durante los años que siguieron las modalidades de test sistemático aumentaron aún más la exigencia por eficiencia (razón costo/eficacia) tan apreciada en EE.UU. (validez de Aceptabilidad*). Después de la guerra de 1940-1945 todavía en EE.UU., la exigencia creciente por no discriminación racial (*Civil Rights*) en la calificación hizo que se valorizara la corrección neutral de lo que llamaron los "*objective tests*" (validez Deontológica*).

El acto de añadir criterios de análisis *a posteriori* de las respuestas, para calcular los índices de discriminación (por ejemplo, vía correlaciones *point biseriales*, como Davis en 1946) dio a las PSM una validez de Replicabilidad* vía la psicometría. Esos cuatro tipos de validez explican, desde nuestro punto de vista, la mayor adhesión a las PSM que existe en EE.UU. por sobre Francia, siendo los franceses pioneros de los test de coeficiente intelectual (Binet y Simon, 1905). Aunque desde 1963, en su libro *La Docimología*, Piéron viene evidenciando las discordancias importantes que pueden existir entre las calificaciones de diferentes jueces frente a una misma hoja "redactada", e incluso la inestabilidad de la calificación de un mismo juez para la misma hoja, en momentos diferentes, las autoridades francesas, estando al tanto de este problema, han mantenido la notación subjetiva con la justificación de que es ciega (los jueces no conocen la identidad del estudiante, por lo tanto las injusticias se reparten al azar según un viejo principio, francés también, de igualdad). Puede ser que la práctica más y más corriente de interponer recursos judiciales frente a decisiones de calificación (moda originada también en EE.UU.) exija que esta situación sea reconsiderada.

B.2. Un ataque teórico sobre el azar resulta en la *correction for guessing* clásica

Tversky (1964) definió el "poder" de un test ("*the power of a test*") como "1 – la probabilidad de obtener el resultado perfecto gracias al azar". Ahora bien, en cada PSM clásica que comporta *k* soluciones el estudiante tiene una probabilidad de *1/k* de dar la solución correcta por azar. Varias técnicas han sido desarrolladas para paliar este defecto.

[21] Alfa tests para los que sabían leer y Beta tests para los iletrados.

Desde 1920 Mc Call recurrió a la *correction for guessing* clásica, que consiste en establecer las tarifas como sigue: la Tarifa en caso de Respuesta Correcta (TC) vale +1 punto, la Tarifa en caso de Omisión (TOM) vale 0 y la Tarifa en caso de Respuesta Incorrecta (TI) vale −1/(k−1).

Este proceso fue criticado muy poco tiempo después de su creación (West, 1923). Es la posición del autor, aun hoy, que este proceso es inadecuado, en primer lugar porque se basa en un modelo teórico falso de la actividad mental de un estudiante al contestar una PSM clásica: el primero de los tres modelos descritos por Bruce Choppin[22] (1975).

En este modelo 1, cuando el estudiante "sabe" elige la respuesta correcta, y cuando "no sabe" elige al azar entre las soluciones propuestas. De donde resulta la *correction for guessing* clásica.

El modelo 2 empieza como el primero, pero en lugar de contestar al azar cuando no sabe, el estudiante empieza eliminando las soluciones que sabe que son falsas y elige entre las que quedan. Este modelo 2, del cual el 1 es una variante extrema, reconoce la noción de conocimiento parcial promocionada por De Finetti (1965). Este modelo ha inspirado consignas de tipo PSRM (Preguntas con Soluciones y Respuestas Múltiples), que consisten en invitar al estudiante a eliminar las soluciones incorrectas, resultando en una PSM clásica (una sola solución correcta), con notas variando de −(k −1) a +(k −1). Esto produce una medida más sutil, más Diagnóstica*.

El modelo 3 de Choppin ahonda en el concepto de conocimiento parcial y dice que cuando una persona se enfrenta a una pregunta (abierta o PSM), empieza imaginando soluciones, después las ordena según un orden de probabilidad decreciente y si la consigna le fuerza a dar solo una, elige la que tiene (a sus ojos) la probabilidad más alta. Este modelo inspira a recurrir a los grados de certeza porque, como lo decía De Finetti (1965), "Solo la probabilidad subjetiva puede dar una significación objetiva a cada respuesta y a cualquier método de medición y de puntaje" (p. 111).

Al tener como referencia los modelos de Choppin es cuando aparecen las debilidades de la *correction for guessing* clásica: (1) Le falta validez Deontológica* porque es injusta: penaliza ciegamente a las personas a quienes se ha prohibido expresar su grado de duda. Además (2) le falta validez Informativa* para los docentes, porque no les muestra nada más que el número de respuestas correctas. Por fin (3), y por las mismas razones, le falta validez Consecuencial* para los estudiantes, porque excepto "omitir más frecuentemente" no tiene efecto alguno sobre su comportamiento. Cross y Frary (1977) demostraron (ver detalles en Leclercq, 1986) por qué este proceso tiene poco impacto en disuadir la "adivinación" de la respuesta por parte de los estudiantes.

[22] Ver Capítulo 13 sección C.

Quienes promueven el uso de los grados de certeza[23] sostienen que esta técnica entrega respuesta a las tres fallas señaladas antes.

B.3. Un vendaval de críticas teóricas sobre los procesos mentales evaluados y no evaluados

a) *Las Preguntas de Selección Múltiple clásicas no evalúan el recuerdo (evocación) de memoria_*

Es evidente que las PSM no pueden pretender evaluar la capacidad de evocar conocimiento, sino de reconocerlo[24]. En efecto, desde hace mucho tiempo se sabe (Luh, 1922) que la performance de reconocimiento tiene una taza de éxito más alta que la de evocación. Esta observación ha sido muchas veces confirmada en contextos tan diferentes como el aprendizaje de idiomas extranjeros (Bahrick, 1984) o de medicina (Schuwirth, 1998). Añadir la solución "Ninguna" (u "Otra") a las soluciones posibles mejora la validez Teórica*.

b) *Las Preguntas de Selección Múltiple clásicas invitan a un razonamiento al revés*

Aun con la solución "Ninguna" u "Otra", las PSM inducen un proceso mental que no corresponde a lo que los estudiantes tienen que practicar en la vida normal. Con una PSM clásica, el estudiante tiene la tendencia a considerar (y eliminar) las soluciones propuestas, y solo después a elegir una solución, en lugar de primero evocar una personal. Esto corresponde al Modelo 2 de la actividad mental descrita por Choppin. Ahora bien, lo que es deseable evaluar es la capacidad del estudiante de evocar / concebir la solución, y solo después confrontarla a soluciones propuestas. Es el principio de las PSN (Leclercq, 2005) o Preguntas con Soluciones Numerosas, que permiten utilizar hojas para la lectura óptica de las respuestas en forma de marcas: el estudiante recibe una lista de centenas de soluciones arregladas en orden alfabético (como el índex de un libro) entre las cuales tiene que elegir. Cada solución posible tiene un número en el orden del listado (por ejemplo de 001 hasta 700) y es con este número de tres cifras (leíble por el lector óptico de marcas -LOM) que el estudiante contesta. En esta modalidad se mantienen las ventajas de la automatización de la corrección (validez de Aceptabilidad*), y al mismo tiempo se aumenta validez Teórica* (de cobertura o de contenido) a la dimensión de los conocimientos. La automatización de la corrección permite administrar muchas preguntas de este tipo (por ejemplo, 100 en una hora).

Las PSN no se justifican cuando las respuestas pueden hacerse con un teclado de computador. Hay incluso sistemas donde basta introducir las primeras letras de la palabra y el sistema propone las soluciones numerosas. Schuwirth (1998) ha aplicado este principio en tests de diagnóstico médico: si el estudiante introduce "diab", el sistema le propone "diabetes de tipo 1, diabetes de tipo 2", etc. Ha

[23] Ver sección B.3d y Capítulo 16.
[24] Ver detalles en el Capítulo 15, sección B.2.

llamado a esto *Long Menu Questions*, un sistema que mejora la validez de Aceptabilidad* para el estudiante.

c) *Las Preguntas de Selección Múltiple clásicas refuerzan el currículum oculto de la escuela.*
El currículum oculto (de la escuela) es lo que nadie enseña pero que todo el mundo aprende. Por ejemplo, que cuando hay una pregunta es obligatorio contestarla. Ahora bien, algunas preguntas, porque son absurdas o excesivamente intrusivas, ¡no deben o no pueden recibir respuesta alguna! Se aprende entonces que cuando la autoridad hace una pregunta, esta es naturalmente pertinente y bien formulada. Se aprende así que toda pregunta tiene una respuesta (correcta) y que, si uno no la sabe, no se debe reflexionar para encontrarla con razonamiento. En suma, el currículum habitual (aunque afortunadamente hay más y más excepciones) no ejercita en la vigilancia cognitiva, en la detección de las anomalías, de las incoherencias, etc., especialmente con sus modalidades de *testing*, siendo las PSM clásicas una de las más representativas. ¡Esta es una grave laguna en la validez Teórica* de esta técnica! Por todas estas razones es que hemos desarrollado (Leclercq, 1986) las PSM con Soluciones Generales Implícitas o PSM SGI. Estas soluciones son cuatro: Ninguna, Todas, Faltan datos en el enunciado de la pregunta y Absurdo en el enunciado. Son Generales porque valen (son idénticas) para todas las preguntas de un test PSM SGI. Son implícitas porque son presentadas solo una vez (al comienzo del test) y no se repiten en cada pregunta: el estudiante debe pensar en ellas solo, por su cuenta. En consecuencia, este proceso impacta positivamente en la validez Informativa* (o diagnóstica) porque permite distinguir dos niveles de la taxonomía de Bloom: la comprensión (sin trampa) y el análisis (con trampas en el enunciado de la pregunta). Gilles (1999) ha mostrado, en el ámbito de asignaturas de medicina, que las PSM SGI en las cuales la respuesta correcta es una de las cuatro SGI tienen una validez Predictiva* superior a las preguntas en las cuales la respuesta correcta es una solución simple (visible).

d) *El encuentro de las Preguntas de Selección Múltiple y los grados de certeza.*
Recurrir a grados de certeza es independiente de las PSM: aplican también para preguntas "abiertas", que exigen respuestas redactadas. Shuford *et al* (1966), Van Naerssen (1962) y De Finetti (1965) demostraron que la consigna para indicar la seguridad en la respuesta no debe ser verbal ("poco seguro", "medianamente seguro", "muy seguro") sino probabilística (en porcentajes de probabilidad). Además, hemos demostrado (Leclercq, 1982, 1993) que una precisión –un grado de discriminación sobre la propia seguridad– más alta que 20% es ilusoria, resultando nuestra consigna en 6 grados: 0%, 20%, 40%, 60%, 80%, 100% seguro/a de la respuesta.

En la línea de los autores ya citados, pensamos que este proceso tiene una mayor validez Ecológica* que los tests habituales que impiden a los estudiantes expresar sus dudas. Choppin (1975) ha descrito este problema en sus modelos 1, 2 y 3. Denunció la visión maniquea (todo o nada) de frases como "Contesten solo cuando saben, y omitan si no saben", pues con frecuencia estamos en un estado mental de conocimiento parcial (De Finetti, 1965), en particular en situaciones de aprendizaje. El grado de duda explica comportamientos de verificación (en el diccionario, por ejemplo) como lo hemos demostrado experimentalmente (Leclercq y Gilles, 1993, p. 45).

Con las PSM, los grados de certeza resuelven (pero es un efecto secundario, no el principal) el problema del *guessing (adivinación)*, resultando en una contribución a la validez de Aceptabilidad* (para los docentes) de las PSM.

Al final, los grados de certeza muestran su contribución más importante en la validez Informativa* de las PSM, cuando las soluciones erróneas son elegidas más frecuentemente y con un grado de certeza más alto que la (o las) solución(es) correcta(s), lo que es anormal. Esta situación es reveladora de concepciones erróneas (en inglés *misconceptions*).

Dejamos hasta aquí esta dialéctica entre los mejoramientos sucesivos de las PSM y las críticas que continúan, ambas contribuyendo a mejorar varios aspectos de validez de las medidas. La historia no está acabada. Invitamos a quienes lo deseen a escribir unas de estas páginas en la historia de las PSM.

C. Aplicación de los criterios ETICPRAD a un DEA

C.1. Razones para concebir Dispositivos de Evaluación de los Aprendizajes con varios componentes

a) Cada tipo de evaluación no puede, en solitario, evaluar todos los objetivos a un nivel aceptable, de acuerdo con los criterios de calidad o tipos de validez de ETICPRAD. Entonces, un DEA a menudo está constituido por varios eventos de evaluación, que ocurren en varios momentos, con varias finalidades, midiendo varios objetivos, con varias técnicas e instrumentos. El resultado es que el conjunto de todo ello constituye un DEA que satisface a la gran mayoría de los criterios ETICPRAD.

b) Utilizar varios métodos de evaluación tiene ventajas, siendo la más importante la *validez consecuencial* del conjunto. Ha sido demostrado que variar los modos de evaluación es formativo. La tabla que sigue es un resumen de la tabla original del informe de Chouinard *et al.* (2005) sobre una investigación en 106 clases (1.844 estudiantes) en escuelas primarias y 61 clases (1.719 estudiantes) en escuelas secundarias. La evaluación unimodal era utilizada en 62% de las clases de primaria y en 72% de las clases de secundaria. Los signos "–" y "+" indican cuál de los enfoques obtiene mejores resultados por cada criterio (o si se obtienen resultados similares: "=").

Tabla 1: Comparación de enfoques unimodales y multimodales. Adaptado de Chouinard *et al.* (2005)

	Evaluación en 2003	
	Unimodal	Multimodal
Interpretación de las exigencias de la tarea	–	+
Implicación general en el trabajo	=	=
Implicación en tareas específicas	–	+
Planificación	–	+
Control	–	+
Rectificación–Ajuste	–	+
Recuerdo de lo que ha sido leído	–	+

La razón de este impacto parece evidente: hay varias evaluaciones, cada una con sus propios objetivos, apuntando a varios procesos mentales; como los alumnos estudian según las evaluaciones (lo que se llama *test driven curriculum*), esos alumnos deben entender las diferencias entre los varios tipos de evaluación y prepararse diferentemente para cada uno. Eso ejerce (y consecuentemente desarrolla) la capacidad de manejar su aprendizaje.

C.2. Cómo aplicar ETICPRAD en un DEA

La siguiente pauta se presenta vacía. Una pauta completa se encontrará en el Capítulo 5 sección D.

Tabla 2: Pauta para evaluar un Dispositivo de Evaluación de los Aprendizajes (DEA) con los ocho criterios de calidad ETICPRAD

Criterios	Componentes del DEA		
	Componente 1	Componente 2	Componente 3
Descripción:			
E			
T			
I			
C			
P			
R			
A			
D			

Este tipo de descripción se puede hacer antes de la aplicación del DEA (en la fase PRE), durante (fase PER) y/o después (fase POST). En los dos últimos casos la reacción de los estudiantes puede ser conocida.

Se puede aplicar a nivel de un curso y también a nivel de un programa.

Referencias

Bahrick, H.P. (1984). Semantic memory content in permastore: 50 years of memory for Spanish learned in school. Journal of Experimental Psychology: General, 120, 1-29.

Bloom, B.S. (1972). L'innocence en pédagogie. Education - Tribune Libre, 135, 14-20.

Brunswick, E. (1943). Organismic achievement and Environment Probability. Psychological Review, 50, 255-272.

Choppin, B.H. (1975). Guessing the answer on objective tests. British Journal of Educational Psychology, 45, 206-213.

Chouinard, Bowen, Cartier, Desbiens, Laurier y Plante (2005). Effets de différentes approches évaluatives sur l'engagement et la persévérance scolaires dans le contexte du passage du primaire au secondaire. FQRSC, 2003-PS-)4321.

Cronbach, L. y Meehl, P. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cross, L. y Frary, R. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple choice tests. Journal of Educational Measurement, vol. 14, 313-321.

Davis, F.B. (1946). Analyse des items. Louvain: Nauwelaerts (1966).

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 18, 87-123.

Ebel, R.L. (1969). Expected reliability as a function of choices per item. Educational and Psychological Measurement, 29, 565-570.

Gilles, J.L. (1999). Apports des mesures métacognitives lors d'un test de compréhension d'un article scientifique, in C. Depover y B. Noël (Eds). Approches plurielles de l'évaluation des compétences et des processus cognitifs. Actes de la 12e Conférence de l'ADMEE Mons: UMH-FUGAM, 19-30.

Green, D.R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. Educational Measurement, 17, 16-19, 34.

Leclercq, D. (1982), Confidence marking, its use in testing,. in Postlethwaite y Choppin, Evaluation in Education, vol. 6, 161-287, Oxford: Pergamon Press.

Leclercq, D. (1993). Validity, Reliability and Acuity of Self-Assessment in Educational Testing, in Leclercq D. y Bruno J. (1993), Item Banking: Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin: Springer Verlag, 114-131.

Leclercq, D. (1986). La conception des QCM. Bruxelles: Labor.

Leclercq, D. y Gilles, J.L. (1993). Hypermedia: Teaching Through Assessment, in D. Leclercq y J. Bruno (1993). Item Banking, Interactive Testing and Self Assessment. NATO ASI Series F112. Heidelberg: Springer Verlag, 31-48.

Leclercq, D. (Ed.) (2003). Diagnostic cognitif et métacognitif au seuil de l'université. Le projet MOHICAN mené par les 9 universités de la Communauté française Wallonie Bruxelles. Liège: Editions de l'Université de Liège.

Leclercq, D. (2005). Edumétrie et docimologie pour praticiens chercheurs. Editions de l'université de Liège.

Leclercq, D. (2006) L'évolution des QCM. In G. Figari et L. Mottier-Lopez. Recherches sur l'évaluation en Education. Paris : L'Harmattan, 139-146.

Luh, C.W. (1922). The conditions of retention. Psychol. Monograph, 31, 142, 401-410.

Mc Call, W.A. (1920). A new kind of school examination. Journal of Educational Research, 1, 33-46.

Piéron, H. (1963). Examens et docimologie, Paris: Presses Universitaires de France.

Schuwirth, L. (1998). An approach to the assessment of medical problem solving: Computerised Case-based Testing, Ph. D., Rijksuniversiteit Maastricht: Datawyse Universitaire Press.

Shuford, E., Albert, A. y Massengil, N.E. (1966). Admissible probability measurement procedures, Psychometrika, 31, 125-145.

Tversky, A. (1964). On the Optimal Number of Alternatives at a Choice Point'. Journal of Mathematical Psychology 1(2): 386-391.

Van Naerssen, R.F (1962). A scale for the measurement of subjective probability. Acta Psychologica, 20, 2, 159-166.

West, P.V. (1923). A critical study of the right minus wrong method. Journal of Educational Research, 8, 1-9.

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