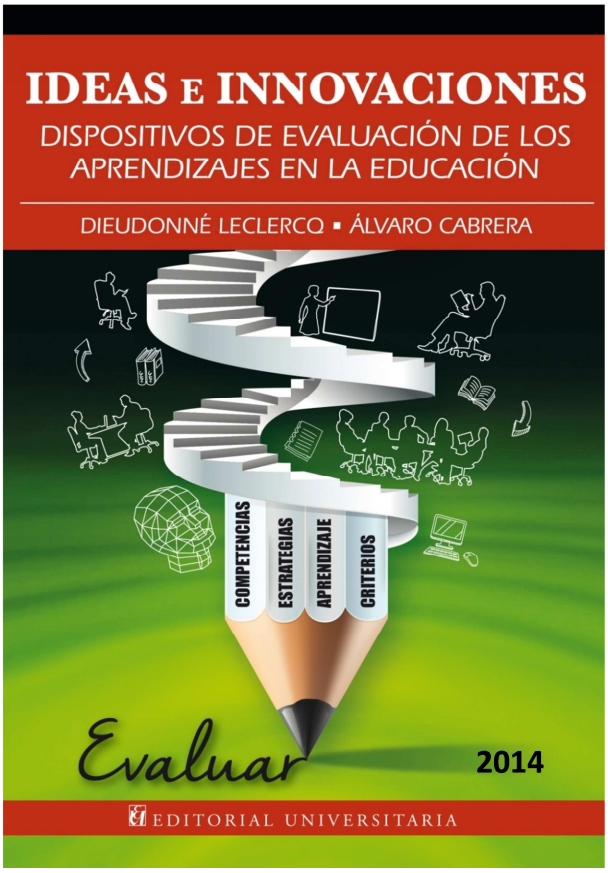

Dieudo LECLERCQ

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.

PARTE 2

Dispositivos de Evaluación de los Aprendizajes (DEA) para la evaluación de desempeños complejos

Capítulo VI

La calificación subjetiva de los desempeños complejos

Dieudonné Leclercq y Álvaro Cabrera

A. La Docimología

A.1. Una ciencia nueva y un método experimental

En 1934 Laugier, Piéron y Weinberg publicaron *Estudios docimológicos*, donde estudiaron varios fenómenos vinculados a la corrección de las hojas del examen externo que se llama, en Francia, el "Baccalauréat"[30]: una prueba externa (con las mismas preguntas en todo el país) impuesta a todos los estudiantes del fin de la escuela secundaria y corregidas de modo anónimo. La guerra interrumpió estas investigaciones, hasta que en 1963 Piéron publicó su libro *La docimología* (ciencia de los exámenes).

La palabra *"docimología"* viene del griego antiguo: δοκιμαςω (examinar), δοκιμη o δοκιμσια (prueba), δοκιμαστης (examinador) y δοκιμαστικος (capacitado a examinar). Algunos dispositivos experimentales (ver secciones siguientes) fueron imaginados por Piéron y sus colaboradores (Laugier y Weinberg) y sus sucesores (Noizet, Bonniol, Caverni). Gilbert de Landsheere ha realizado síntesis de esas obras.

A.2. El peligroso mito de la curva de Gauss

Esta expresión de De Landsheere (1992, p. 36) evoca el hecho de que muchos jueces de pruebas piensan que las calidades de estas se distribuyen conforme a una ley de Gauss o del azar, es decir, que muchas tendrán una calidad "media" (en Gauss 68%), habrá menos "buenos" (13,5%) o "mediocres" (13,5%) y pocos "excelentes" (2,5%) o "muy débiles" (2,5%). Esta representación está influenciada por el hecho de que muchos fenómenos se distribuyen así (Figura 1), como resultado del azar, incluidos algunos tests psicológicos que miden aptitudes (que se distribuyen al azar).

En educación, combatimos el azar e intentamos obtener la perfección. Y al no ser la perfección de este mundo, intentaremos obtener una curva en forma de "J" (Jota mayúscula), lo más pronunciada posible.

[30] El principio de este Baccalauréat fue instituido por Napoleón en Francia. En Chile existió una versión de esta prueba (el "Bachillerato"), que se aplicó desde mediados de la década de 1840 (pocos años después de la fundación de la Universidad de Chile en 1842) hasta 1966. Para esa fecha había derivado en una prueba con Preguntas de Selección Múltiple (PSM) de corrección automática, y fue precursora de la Prueba de Aptitud Académica, PAA (1967 – 2002) y la actual Prueba de Selección Universitaria, PSU.

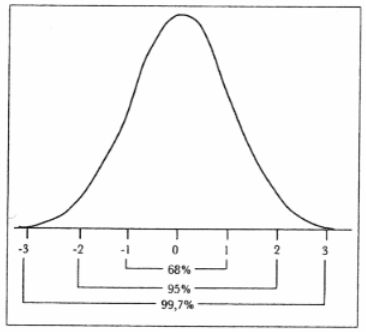

Figura 1: Curva en forma de campana de Gauss

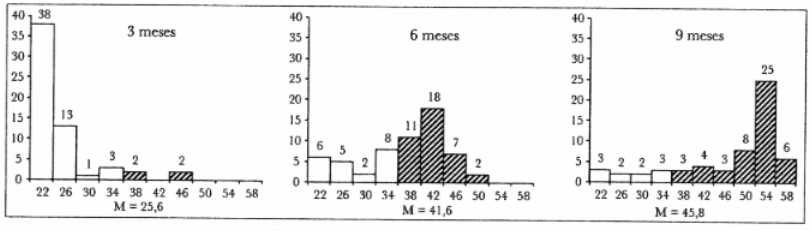

Figura 2: Curvas en i, Gauss, y Jota

En la Figura 2 se ve la evolución de los resultados de los mismos alumnos (de primer año de educación primaria) tomando el mismo test (de lectura) 3, 6 y 9 meses después de su ingreso a la escuela. Al comienzo los resultados se presentan según una distribución que tiene una forma en "i" (curva de la izquierda): la mayoría de los resultados son débiles, es decir, están cerca del mínimo posible. Durante el año la curva evoluciona pasando por una forma cercana a la de Gauss (curva del centro). Al final del año, la curva evoluciona otra vez, tomando una forma de jota (curva de la derecha), donde la mayoría de los estudiantes se acercan al máximo puntaje posible de este test (63 puntos).

A.3. La ley de Posthumus[31]

A) LA LEY

La idea (falsa) de que los resultados deben presentarse en una forma de campana de Gauss es apoyada por la convicción de que los exámenes deben ser selectivos y, para eso, discriminatorios. En consecuencia, algunos docentes tienen la tentación de eliminar las preguntas que tienen una tasa de éxito alto (más del 80%) o bajo (menos del

[31] Docente holandés en Indonesia durante la segunda guerra mundial.

20%)[32], como lo hacen los constructores de tests de inteligencia, porque estas preguntas no diferencian bastante (dejan al 80% de los respondientes indiferenciados). Antes hemos visto que si bien una curva de Gauss puede ocurrir durante el aprendizaje, lograr tal resultado no es el objetivo.

La ley de Posthumus fue formulada así por De Landsheere (1992, p. 242):

> "El/la docente tiende a ajustar el nivel de su enseñanza y el nivel de sus evaluaciones de las performances de los estudiantes, de modo tal que año a año se conserva aproximadamente la misma distribución de puntajes (en forma de Gauss)".

B) LAS DOS MALAS CONSECUENCIAS DE LA LEY

1. Algunos docentes construyen sus calificaciones de modo que obtienen una curva de Gauss en vez de reflexionar sobre el real valor de la respuesta o producto de cada estudiante.
2. Algunos docentes además consideran que hay un porcentaje fijo de estudiantes que pueden (deben) aprobar el curso, y por lo tanto también una cantidad que debe reprobar. Eso explica que a menudo las tasas de éxito sean las mismas cada año, cualquiera sea el esfuerzo de los docentes o de los estudiantes por mejorar: si mejoran los resultados promedio, el umbral de éxito (el resultado que ha logrado el 60%) sube. Este fenómeno, con referencia normativa (ver Capítulo 3), ha sido llamado "hipótesis socio-aritmética" por Hutmacher (1993). Él ha preguntado a docentes de primaria: "A su parecer, sobre los 20-24 alumnos de una clase, ¿cuál es el número normal de fracasos?", y ha recibido como respuesta más frecuente "2". Cuando el Cantón de Ginebra decidió disminuir el número de alumnos por clase a no más de 15, la tasa de fracaso, en lugar de bajar como se había anticipado, aumentó. Hutmacher explica con su hipótesis esta observación contraintuitiva: como el número (2) de fracaso por clase no había cambiado, un fracaso de 2 sobre 20 es 10%, pero de 2 sobre 15 es de 13%.

C) UNA DEMOSTRACIÓN EXPERIMENTAL DE LA LEY DE POSTHUMUS

Gjorgjesvki[33] invitó a cinco profesores de una misma asignatura de la enseñanza secundaria a calificar (independientemente de los otros jueces) 100 pruebas sobre una escala de 5 grados: (1) Insuficiente, (2) Mediocre, (3) Bien, (4) Muy bien, (5) Excelente.

Después dio a otros cuatro jueces 15 copias calificadas como "bien" por los cinco primeros. Estos cuatro jueces entregaron la repartición que muestra la Figura 3:

[32] Estas preguntas pueden llegar a ser eliminadas exclusivamente por su bajo valor discriminatorio, sin importar si pretenden evaluar un saber importante de lograr de acuerdo con los propósitos del curso; dicho de otra forma, las preguntas de las evaluaciones no necesariamente se relacionan con los logros de aprendizaje propuestos por el curso.

[33] Descrito por Rot y Butas, 1959.

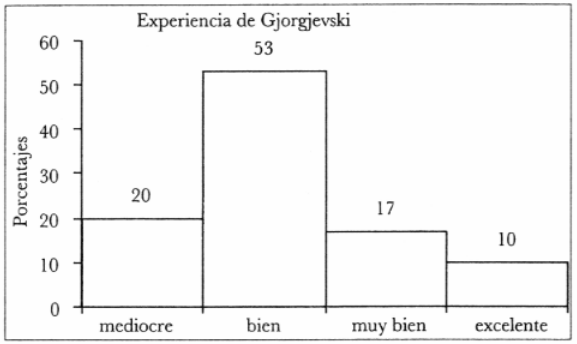

Figura 3: Distribución (por cuatro jueces) de trabajos calificados previamente como "bien" por otros cinco jueces

B. La docimología crítica constata los riesgos de la calificación subjetiva

B.1. La ausencia de concordancia inter-jueces con respecto al orden

Esta ausencia de concordancia ha sido medida entregando la misma hoja de respuesta de un estudiante (fotocopiada) a varios jueces.

La distribución de notas

Por ejemplo, Piéron (1963, p. 123) muestra un caso en que ha entregado a 76 docentes la misma hoja de "composición francesa" para ser calificada, resultando la siguiente distribución de puntajes:

Tabla 1: Distribución de los puntajes de 76 docentes para el mismo trabajo de redacción

Nota	0-1	2-3	4-5	6-7	8-9	10-11	12-13
Número de Docentes	1	6	20	34	10	3	2

B.2. La ausencia de concordancia inter-jueces con respecto al éxito/fracaso (umbral o corte)

Agazzi (1967, p. 119) menciona una investigación[34] en la cual, para 6 pruebas de contenidos diferentes, 6 jueces han calificado hojas de respuesta del Bac (Bachillerato francés) sobre un máximo de 20 puntos, siendo 10 el umbral de éxito. Los resultados se muestran en la Tabla 2:

[34] De Laugier y Weinberg, colaboradores de Piéron.

Tabla 2: Distribución de umbral de éxito (o corte) de seis jueces para seis pruebas del Bac Francés

	Fracaso para los seis jueces	Éxito para unos y fracaso para otros	Éxito para los seis jueces
Latín	40%	50%	10%
Composición francesa	21%	70%	9%
Inglés	37%	47%	16%
Matemática	44%	36%	20%
Filosofía	9%	81%	10%
Física	37%	50%	13%

B.3. El caso de los exámenes orales

Hartog y Rodes (1936, p. 35-41) han observado que la correlación de la clasificación (es decir, el orden en el cual los candidatos son ubicados por los jueces) de 16 candidatos, por parte de dos tribunales diferentes (de cinco miembros cada uno) era de 0,41 entre las calificaciones de ambos tribunales. Trimble (1934), que había interrogado oralmente a 25 estudiantes de psicología, pidió a dos colegas (A y B) que los re-interrogaran (oralmente también) con los mismos criterios (vocabulario, conocimiento del contenido, apreciación general). Las tres correlaciones fueron muy bajas, con una media de 0,32:

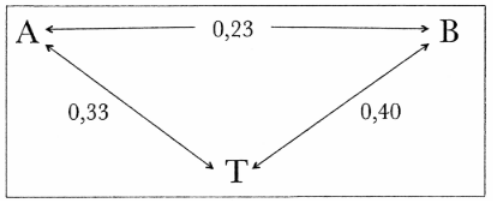

Figura 4: Correlación de los resultados de 3 diferentes jueces

Piéron *et al.* (1962, p. 150) observaron que, en el Bac de 1955, las amplitudes de variación (av o diferencia entre las dos notas extremas) de los exámenes orales eran más amplias que las de los exámenes escritos.

Tabla 3: Amplitud de variación exámenes escritos y orales en Bac 1955

	Escrito	Oral
Filosofía (13 jueces)	3,92	7,18
Matemáticas (17 jueces)	5,00	8,54

B.4. *La inconsistencia (o inestabilidad) intra-juez*

De Landsheere (1992, p. 45) da un ejemplo: "Hartog y Rhodes (1936, p. 15) pidieron a 14 docentes de historia calificar 15 hojas de respuesta por segunda vez, habiendo transcurrido 12 y hasta 19 meses desde la primera calificación, y borraron todas las trazas (huellas) de la primera corrección. Los docentes tenían que otorgar puntajes y además decidir sobre el éxito global o el fracaso. Para 92 casos sobre 210 el veredicto fue diferente entre las dos ocasiones".

B.5. La falta de concordancia entre la calificación verbal y la numérica

Algunos evaluadores pensaron que el uso de expresiones verbales aumentaría la concordancia interjueces, lo que resultó en la adopción de calificaciones del tipo "Muy bien, Bien, Satisfactorio, Débil, Insuficiente". Las cuatro curvas de la Figura 5 presentan los resultados de una investigación hecha en 1958 por Reuchlin en Francia (relativa al cálculo) en clases de alumnos de 11 años. Los profesores de estos alumnos fueron invitados a atribuir a cada uno de ellos uno de los cuatro adjetivos "Muy bueno, Bueno, Medio, Mediocre". 654 estudiantes fueron juzgados "Muy buenos" por sus profesores, 1.303 "Buenos", 1.551 "Medios" y 1.300 "Mediocres". No es sorprendente constatar que la categoría "Medio" ha sido elegida con la mayor frecuencia (¡Posthumus una vez más!).

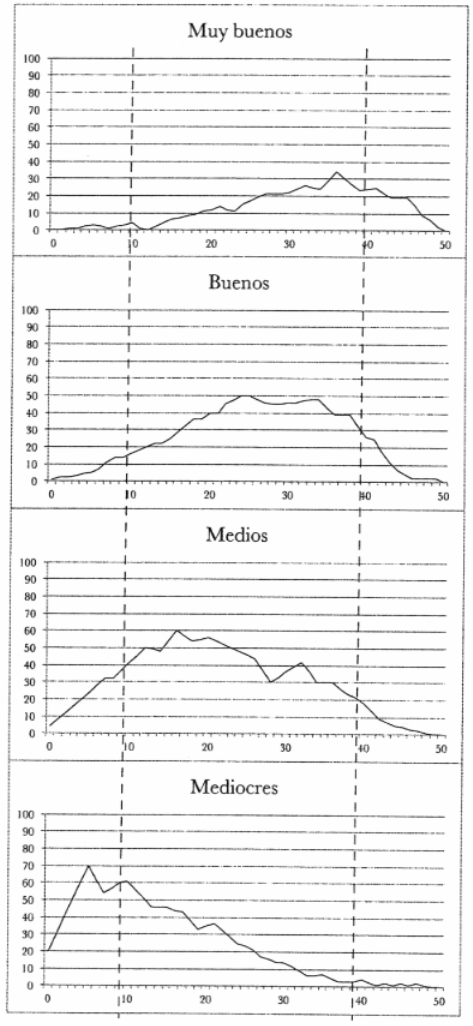

Figura 5: Comparación entre resultados de test de cálculo (corrección objetiva) y evaluación de jueces con categorías verbales (de mediocre a muy bueno) de los mismos estudiantes

En paralelo, estos 4.808 alumnos contestaron un test de cálculo calificado objetivamente sobre un máximo de 50 puntos. Se ve en la Figura 5 que las 4 distribuciones se superponen mucho y que, en la zona que va de 10 a 40 puntos, el mismo nivel (objetivo) de desempeño puede corresponder a cualquiera de los 4 niveles subjetivos (verbales).

Reuchlin (1959) comenta: *Un docente (de primaria o de secundaria) conoce más que otros los contenidos del programa que son dominados por cada uno de sus alumnos. Lo que ignora es la gravedad de cada debilidad cuando se compara, no al interior de una clase que puede ser fuerte o débil, sino que con los resultados de todo un país.*

C. La docimología crítica explica las razones y mide los efectos "de corrección" o "de evaluación subjetiva"

Experiencias de la docimología crítica han demostrado cuáles son los mecanismos que distorsionan la evaluación correcta de un desempeño.

C.1. El efecto de halo

> *El efecto de halo presenta un carácter afectivo... A menudo se sobrestiman las respuestas de un estudiante que tiene un aspecto bonito, con una mirada franca, con una dicción agradable...la escritura también puede afectar al juez* (De Landsheere, 1992, p. 49).

Chase (1968) demostró que una escritura fea hace bajar el puntaje.

Para verificar este fenómeno Weiss (1969) realizó la siguiente experiencia[35]:

Dos hojas (redacción en francés) de alumnos de 10 años fueron re-dactilografiadas y entregadas a dos grupos de 23 profesores de educación primaria cada uno.

Al grupo n°1 se le dijo lo siguiente:

"La hoja 1 ha sido escrita por un alumno medio que le gusta leer tiras cómicas; sus padres son empleados".	"La hoja 2 ha sido escrita por un alumno dotado; su padre es redactor de un conocido periódico".

y al grupo 2 se le dijo exactamente lo contrario.

Además de una calificación "global", tres aspectos (ortografía, estilo, fondo) debían ser juzgados sobre una escala de 5 niveles (1 = Insuficiente y 5 = Muy Bien). Las notas de los dos grupos de correctores han sido las siguientes:

[35] Descrita por De Landsheere, 1992, p. 50.

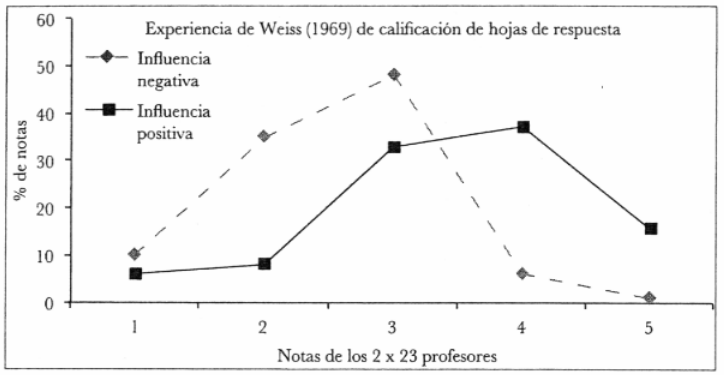

Figura 6: Comparación de las distribuciones de notas luego de influencia positiva o negativa

C.2. El efecto de estereotipo

> *Conocer los resultados anteriores de un alumno –incluso desconocido– puede influenciar al juez…resultando en una inmutabilidad en el juicio* (De Landsheere, 1992, p. 47-48).

Caverni, Fabre y Noizet (1975) invitaron a profesores de ciencias de educación secundaria a calificar (sobre 20 puntos) 4 hojas de respuesta, las que iban acompañadas de 5 notas (inventadas) que el estudiante autor de la respuesta había recibido con anterioridad.

Las respuestas acompañadas de notas anteriores con media alta y dispersión estrecha recibieron la nota media 9,69 y las hojas acompañadas de notas anteriores con media baja y dispersión amplia recibieron la nota media 6,69.

De Landsheere comenta:

> Se podría haber utilizado otro descriptor: la progresión (subiendo) o la regresión (bajando) de las 5 notas anteriores. …Este fenómeno (de estereotipo) también se aplica a performances donde existen criterios de corrección objetivos, como la ortografía. La experiencia siguiente lo demuestra: un profesor de lenguaje realiza regularmente dictados, de modo que conoce a los alumnos que obtienen los mejores resultados …para ellos, el número de errores no detectados es más grande que para los alumnos más débiles en ortografía. En el caso de los primeros, el profesor presume que no hay error alguno; a los segundos les acecha.

C.3. El efecto de orden o de contraste

> *Los estudiantes hábiles han descubierto hace mucho tiempo la importancia de los contrastes: en un examen oral, pasar inmediatamente después de un candidato brillante es desfavorable, y suceder a uno más débil que uno mismo puede ser una ventaja… excepto que la debilidad de las respuestas que el interrogador ha obtenido recién lo haya puesto de mal humor* (De Landsheere, 1992, p. 52).

Bonniol (1965) ha verificado esta hipótesis utilizando lo que ha llamado anclas altas (hojas de respuesta con buenas notas antes) y anclas bajas (lo contrario). Dio la misma hoja (H) a varios jueces, pero con anclas diferentes. Colocar un ancla alta (A) es ubicar una hoja excelente exactamente antes de la hoja H, de modo que la corrección de A desfavorezca la de H que sigue. Por supuesto, colocar un ancla baja (B) en lugar de una alta (A) resulta en el efecto inverso. Bonniol llama "ancla grave" a la sucesión de tres anclas altas.

C.4. El sesgo de confirmación

Noizet y Caverni (1978, p. 141) explican el impacto de las anclas por el mecanismo del sesgo de confirmación:

> *Es probable que los primeros indicios –ya sean positivos o negativos– van a guiar la búsqueda de indicios …con el evaluador intentando encontrar indicios que confirmen sus primeras inferencias, más que indicios que pudiesen ponerlas en duda.*

De Landsheere (1992, p. 54) comenta: *Al parecer, si debe cometer errores, al estudiante le conviene cometerlos en la segunda parte del examen.*

El efecto de estereotipo puede explicarse por el sesgo de confirmación.

C.5. El sesgo de confirmación y el examen oral

Con frecuencia los docentes universitarios descubren que los resultados de sus estudiantes en los exámenes escritos se distribuyen según una curva de Gauss, aunque los resultados del examen oral se distribuyen en una curva en forma de U (muchos muy débiles, y otros muchos muy buenos).

La distribución de Gauss es previsible si las preguntas son elegidas sobre la base

- de su dificultad con una tasa media de éxito cercana al 50%, y
- de sus índices de discriminación: con una alta correlación *point biserial* (ver Capítulo 23) de la respuesta correcta.

Una posible explicación de la curva en forma de U en los resultados del examen oral es el sesgo de confirmación: para tomar una decisión sobre el éxito o fracaso de un estudiante (y quedar tranquilo con la decisión tomada) el profesor prefiere disponer de datos sin ambigüedad, siendo la mejor situación posible el que todas las repuestas vayan en la misma dirección: todas falsas o todas correctas. Por lo tanto, se produce una tendencia, después de solo algunas respuestas, a hacer preguntas para confirmar lo que ya se ha observado y, consecuentemente, concluir el diálogo.

D. La docimología positiva propone una Solución n° 1: varios jueces

D.1. El número necesario de jueces (si es que no se han puesto de acuerdo previamente)

Piéron reporta una investigación instructiva: "Queriendo obtener un coeficiente de fiabilidad alto (0,99), y basándose en las medias de los índices de correlación obtenidos para cada categoría de prueba, Laugier y Weinberg concluyeron que el número de jueces necesario, para cada categoría, es el siguiente:"

Tabla 4: Cantidad de jueces necesaria para alcanzar una correlación 0,99 entre las calificaciones

Composición francesa	Latín	Inglés	Matemáticas	Disertación filosófica	Física
78	19	28	13	127	16

Se debe comentar que un índice de 0,99 es muy alto, y en la mayoría de los casos sería normal contentarse con una correlación de 0,90.

D.2. El peligro de un acuerdo ilusorio

Cuando se pregunta a profesores de una disciplina (Lenguaje, por ejemplo) cuáles son sus criterios para evaluar la calidad de una respuesta o realización (una composición, por ejemplo), a menudo dan la misma lista de criterios (por ejemplo, la pertinencia y la originalidad de las ideas, el estilo, el vocabulario, la ortografía, etc.), de modo que se puede pensar que deberían otorgar la misma calificación a la misma hoja de respuesta. Sin embargo, sabemos que no ocurre así. Un punto de acuerdo que no se debe olvidar es el peso relativo que cada docente asigna a cada criterio, y estos varían de un docente a otro. Así, no es suficiente que acuerden los criterios: debe también existir acuerdo sobre sus pesos relativos.

D.3. Un ejemplo de doble corrección

Leclercq y Hubert (1990) corrigieron las hojas de respuesta de 33 estudiantes que habían recibido las siguientes instrucciones:

> Considerando como objetivo general que los alumnos de 14 a 16 años "consuman la TV de una manera activa e inteligente", detalla los sub-objetivos en los ámbitos Afectivos, Sensorio-motores, Cognitivos, de Identidad, y de Decisión (ASCID). Los criterios de evaluación de tu respuesta serán:
>
> - la pertinencia de la clasificación (en términos ASCID)
> - la variedad (en términos de flexibilidad, no de fluidez –conceptos de Torrance)
> - la precisión (grado de detalle)
> - la calidad técnica de la redacción de los objetivos

Las instrucciones eran, al mismo tiempo, los criterios de evaluación. Pero los pesos específicos de los criterios no fueron anunciados.

En la *etapa 1* los dos jueces (A el profesor y B la asistente) corrigieron, sin comunicarse, con el método de la *impresión general* (es decir, sin asignar puntajes a cada uno de los criterios), y otorgaron un puntaje global, de entre 0 a 20, a cada hoja de respuesta.

En la *etapa 2* profesor y asistente propusieron pesos relativos para los criterios. El resultado fue el siguiente:

Tabla 5: Pesos relativos para cada criterio según el profesor (A) y la asistente (B)

	A (Profesor)	B (Asistente)
Pertinencia de la clasificación (en términos ASCID)	6	7
La variedad (en términos de flexibilidad)	6	6
La precisión (grado de detalle)	4	5
La calidad técnica de la redacción de los objetivos	4	2

En la *etapa 3* adoptaron una sola ponderación para los criterios (la de A).

En la *etapa 4* cada corrector dio un puntaje a *cada criterio* (evaluación *analítica*) para cada una de las hojas de respuesta.

En la *etapa 5* se establecieron las correlaciones y los correctores confrontaron sus puntajes. En las figuras 7 y 8 se ve que B ha sido un poco más severa que A (la recta de tendencia está debajo de la diagonal), tanto en la evaluación analítica como en la evaluación global. La correlación inter-correctores es mejor cuando la evaluación es analítica (0,88) por sobre la evaluación global (0,85); esta última correlación se acerca a la que Britton (1963) observó cuando se practica una evaluación "rápida".

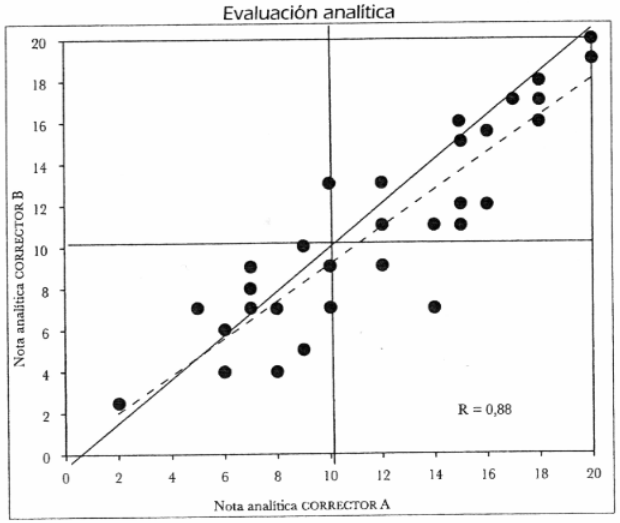

Figura 7: Correlación entre la evaluación analítica de A y B

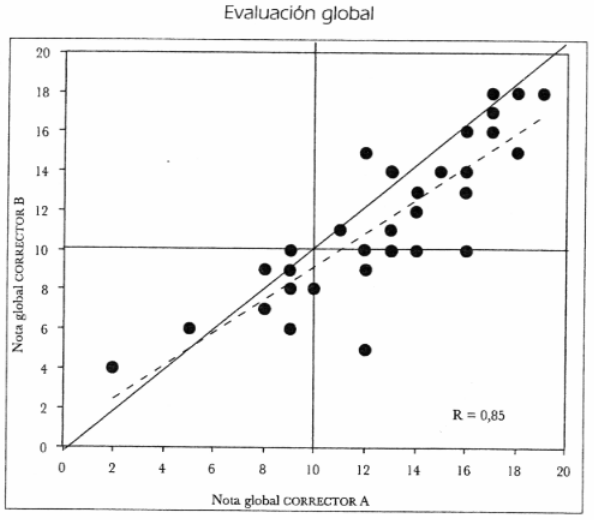

Figura 8: Correlación entre la evaluación global de A y B

En las figuras 9 y 10 se ve que los dos correctores fueron más severos en sus evaluaciones analíticas que en sus evaluaciones globales (lo que ocurre frecuentemente).

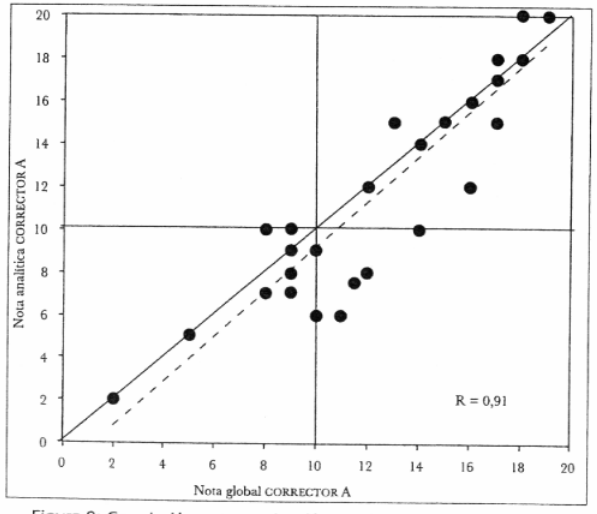

Figura 9: Correlación entre evaluación analítica y global Corrector A

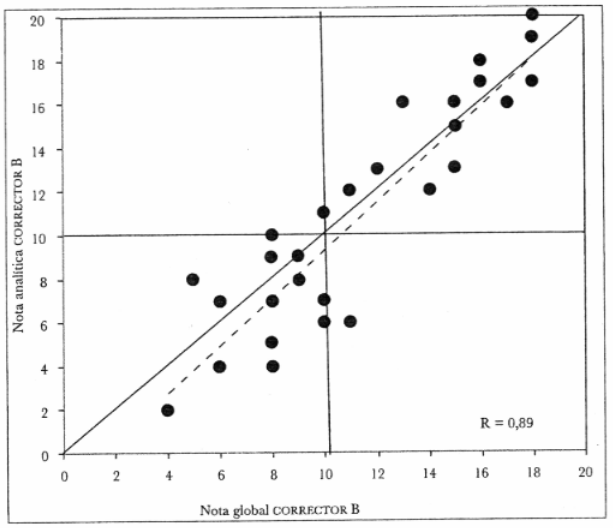

Figura 10: Correlación entre evaluación analítica y global Corrector B

En la *etapa 6* los correctores re-examinaron aquellas hojas de respuesta donde las calificaciones globales divergían en más de un punto sobre 20, entre un corrector y otro, y luego las calificaciones analíticas que presentaban estas discordancias. En base a este análisis modificaron algunas de sus notas, de forma que las correlaciones mejoraron. La Tabla 6 muestra la evolución de las correcciones, tanto inter-jueces como intra-juez, antes y despúes de la concertación entre jueces de la etapa 6.

Tabla 6: Evolución de las correlaciones inter–jueces e intra–juez antes y después de un proceso de concertación

Antes de la concertación entre jueces				Después de la concertación entre jueces			
		0,91				0,93	
		A: global	A: analítico			A: global	A: analítico
0,89	B: global	0,85		0,90	B: global	0,89	
	B. analítico		0,88		B. analítico		0,92

La *etapa 7* consistió en colocar la nota final, resultando de la media de las calificaciones de los dos correctores, después de los ajustes realizados en la etapa 6.

Además, Leclercq se preguntó si sus notas globales habían sido influenciadas por lo que ya sabía de los estudiantes (*efecto de halo*). Para determinarlo calculó la correlación entre la diferencia de sus notas globales y analíticas con la apreciación global que tenía de cada estudiante (en el curso en general, no por esta producción). La correlación fue de –0,05 (cerca de 0). ¡No hubo efecto de halo!

E. La docimología positiva propone una solución nº 2: la estandarización

La estandarización consiste en intentar aplicar a las condiciones de la evaluación algunas normas comunes con el propósito de que todos los estudiantes sean tratados de la misma forma en los momentos de evaluación. En específico, se podría traducir en que todos reciban las mismas preguntas, en la misma situación, con el mismo tiempo, y sean evaluados por los mismos jueces usando los mismos criterios. La estandarización está íntimamente ligada a la dimensión de validez Deontológica (ver Capítulo 4) de un DEA.

E.1. El examen oral es el prototipo del examen no estandarizado

Piéron, Reuchlin y Bacher (1962) describen una situación donde tres profesores, A, B y C, interrogaron a los mismos 40 estudiantes, y los tres exámenes de cada estudiante fueron grabados (sonido). Se evidenció que la duración de las intervenciones de los jueces durante el examen oral difería mucho de un juez a otro (ver Tabla 7).

Tabla 7: Duración de las intervenciones de 3 jueces en un examen oral

	MIN	MEDIA	MÁX
A	32%	46%	59%
B	47%	57%	73%
C	41%	63%	77%

E.2. ¿Es posible estandarizar el examen oral?

Ilustraremos el punto con un ejemplo de Leclercq y su colega Denis, quienes durante 15 años interrogaron oralmente a centenares de estudiantes (aproximadamente 200 cada año) en el mes de junio (fin del año académico en Europa). Estos estudiantes habían sido testeados en enero y mayo con Preguntas de Selección Múltiple (PSM) y Soluciones Generales Implícitas (SGI: ver Capítulo 13) acerca de 5 de los 7 capítulos del libro base del curso. Los otros dos capítulos (3 y 6) eran reservados para el examen oral. En este examen oral el estudiante:

- sortea una pregunta del Capítulo 3 y otra del Capítulo 6 desde dos paquetes de preguntas;
- prepara durante 10 minutos las dos preguntas, con ayuda del libro y todos los demás recursos que quiera;
- se presenta al interrogador A (Leclercq) para entregar, sin ayuda alguna, sus respuestas a la pregunta del Capítulo 3; el interrogador interviene con preguntas adicionales según las respuestas anteriores;
- hace lo mismo con el interrogador B (Denis) para presentar su respuesta a la pregunta del Capítulo 6.

En cada caso, los examinadores comunican al estudiante la calificación que ha obtenido, con un máximo de 20 puntos.

En 1995 Leclercq y Denis, buscando verificar si eran capaces de evaluar las mismas habilidades transversales en los mismos estudiantes, tuvieron la idea de evaluar no solo la performance vinculada a la pregunta, sino que además buscaron evaluar *habilidades transversales, comunes a todas las preguntas, y que no pueden ser evaluadas con PSMs*:

- estructura de la respuesta
- vocabulario técnico
- originalidad de los ejemplos (que no fueran los que el docente entregó durante el curso)

Para eso acordaron una escala de seis grados para cada uno de esos tres criterios transversales:

Tabla 8: Escala de seis grados para tres criterios de evaluación de un examen oral

1. Estructura	1	2	3	4	5	6
2. Vocabulario	1	2	3	4	5	6
3. Ejemplos	1	2	3	4	5	6

Desde los primeros intentos con los primeros estudiantes se dieron cuenta de que era una misión imposible. Ello se refleja en que sus notaciones eran del tipo que se muestra a continuación, donde cada uno de los criterios tiene la misma evaluación:

Tabla 9: Ejemplos de evaluación de los tres criterios para dos estudiantes (27 y 182)

Estudiante 27

1. Estructura	1	(2)	3	4	5	6
2. Vocabulario	1	(2)	3	4	5	6
3. Ejemplos	1	(2)	3	4	5	6

Estudiante 182

1. Estructura	1	2	3	(4)	5	6
2. Vocabulario	1	2	3	(4)	5	6
3. Ejemplos	1	2	3	(4)	5	6

Eso significa que fueron capaces de hacer diferencias entre los estudiantes, influenciados por los criterios, pero *no fueron capaces de evaluarlos distintamente (por separado)*. Esta imposibilidad se produce pues ocurre un importante efecto "de halo". Como descubrieron luego, este tipo de halo había sido descrito poco antes por Engelhard (1994, p. 99): "*Cuando una notación analítica es requerida sobre escalas diferentes, pero en realidad el juez utiliza una notación global (holistic scoring) … Lo que es evidente cuando el evaluador produce una gran taza de configuraciones idénticas, por ejemplo 222222 o 333333*".

Además de estas dificultades para la estandarización, Leclercq y Denis se dieron cuenta de que su dispositivo experimental tenía otras debilidades. Para comentarlas las planteamos en una lista de calidades que los exámenes orales debiesen tener si queremos que sean estandarizados (y veremos que es una "misión casi imposible").

E.3. Condiciones teóricas ideales de estandarización de un examen oral... y la realidad

Tabla 10: Condiciones ideales de estandarización de un examen oral y comentarios de Leclercq y Denis

Calidades teóricas ideales que deberían tener los exámenes orales	El dispositivo de examen oral de Leclercq y Denis (LyD)...
1. Todos los estudiantes deberían contestar a las mismas preguntas	...no respeta este criterio. LyD lo sabían.
2. ...en las mismas condiciones (momento del día, porcentaje de intervención del juez)	...no respeta este criterio. LyD lo sabían.
3. Si hay varios jueces, todos deberían juzgar la misma performance del estudiante, ya sea directamente o vía una grabación.	...no respeta este criterio. LyD lo sabían.
4....y en las mismas condiciones (de fatiga, de orden, de anonimato)	...no respeta este criterio. LyD lo sabían.
5. Los correctores deberían juzgar las performances de todos los estudiantes según los mismos criterios.	...intentaron hacerlo, pero fue un fracaso.
6. Los criterios de evaluación deberían haber sido anunciados, y de la misma forma (con el mismo tiempo dedicado a ello, por ejemplo)	... respeta este criterio
7. Si diferentes preguntas son utilizadas con varios estudiantes, tendrían que evidenciar y evaluar del mismo modo los mismos aspectos de la competencia.	...intentaron hacerlo, pero fue un fracaso.
8. Los evaluadores deben ser capaces de evaluar la performance que se pide al estudiante	...no respeta este criterio. LyD descubrieron que no eran capaces de hacerlo
9.....de la misma manera.	...no respeta este criterio. LyD descubrieron que no eran capaces de hacerlo

E.4. Las verdaderas razones del fracaso

(1) Los examinadores, limitados como todos los seres humanos en sus memorias de trabajo, no fueron capaces de evaluar durante el mismo tiempo de discusión sobre un tema (10 minutos) la comprensión, el conocimiento, la capacidad de aplicar, de juzgar, y de expresarse de un estudiante, *además* de los tres criterios añadidos. Por supuesto, esto puede mejorar si la interacción es grabada en video y si el interrogador puede verla muchas veces, deteniéndose cuando quiere, anotando las palabras exactas del estudiante, etc. Es evidente que este tipo de dispositivo sería muy costoso en términos de tiempo del corrector. Hubo varias otras razones para no aplicar el sistema de grabación:

– la mayoría de los estudiantes se impresionan cuando son grabados (y por razones de deontología lo deben saber de antemano), y el examen no es un buen lugar para añadir más estrés al que ya está presente.
– "*Reuchlin grabó (el sonido) las respuestas de 20 exámenes de física durante el Bac y las hizo evaluar por 16 docentes... La media de las notas presentó una variación de 8 a 13 sobre 20*"[36].

Así, Leclercq y Denis no aplicaron la grabación... pero por una razón diferente y que se expone a continuación.

(2) Rápidamente se dieron cuenta de que las preguntas no habían sido concebidas para evaluar, al mismo tiempo, los tres criterios transversales añadidos. Algunas preguntas se limitaban a evaluar la capacidad de estructurar la respuesta, pues no demandaban el uso de vocabulario técnico o de ejemplos. Otras podían ser respondidas satisfactoriamente dando ejemplos originales, pero sin necesariamente utilizar el vocabulario técnico apropiado. Casi ninguna pregunta exigía que una respuesta satisfactoria incluyera necesariamente las tres características.

E.5. ¿Hay esperanza de estandarizar exámenes para evaluar desempeños complejos?

Veremos, en el Capítulo 8 dedicado a los ECOES (Exámenes Clínicos Objetivos Estructurados), que sí es posible concebir (y realizar) exámenes orales estandarizados... pero con otros costos.

Para asegurar la estandarización, los ECOE deben concebirse y diseñarse utilizando escalas y rúbricas detalladas, cuestión que se aborda en la siguiente sección.

F. La docimología positiva propone una solución n° 3: escalas y rúbricas

F.1. Tipos de escalas insuficientes

Una escala de evaluación sirve para posicionar la calidad de un desempeño en un *continuum* que va de los grados más bajos a los más altos. Algunas escalas, por sí solas, son insuficientes para asumir la complejidad de la evaluación de los aprendizajes en el ámbito de la educación superior. A continuación presentamos tres ejemplos de este tipo de tablas.

[36] Piéron *et al.* (1962, p. 51) citado por De Landsheere (1992, p. 41).

F.1. A) LAS ESCALAS GRÁFICAS

Estas escalas van de "Muy bajo" hasta "Muy alto", con variaciones en la denominación según el contenido. Por ejemplo, una escala de dolor puede ser:

No perceptible_____Insoportable

Estas escalas, llamadas *Visual analog scales* (VAS) son materializadas en medicina por una regla de madera con la cual el paciente expresa la intensidad de su dolor. En la cara de la regla que ve el paciente no hay ninguna dimensión, salvo los dos extremos, y hay un cursor que él/ella debe manipular. Al reverso de la regla, en la cara que ve el médico, hay referencias numéricas. El médico ve los extremos de forma inversa a como los ve el paciente:

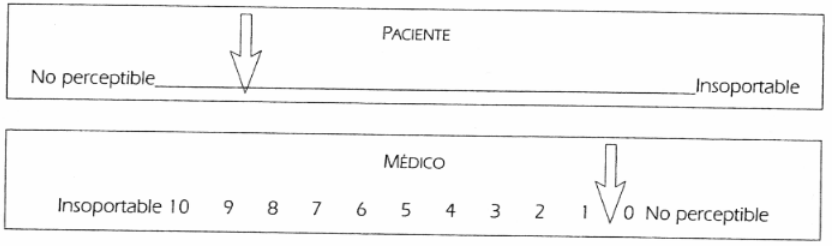

Figura 11: Escala visual del dolor para medicina.

F.1. B) LAS ESCALAS MUDAS

Son aquellas escalas donde los números son vistos y utilizados tanto por los observadores (por ejemplo, los médicos en la VAS del dolor) como por quienes responden. Esto no debe impresionar pues los números no tienen referencia alguna, y por tanto no se puede realizar operaciones matemáticas como adicionar, sustraer, dividir y multiplicar (por ejemplo, en estas escalas, 3 no es dos veces menos que 6). Los números pueden ser códigos ordinales (como A, B, C, D...), e incluso cuando son la serie de los números ordinarios, no garantizan que la distancia entre 3 y 4 es de la misma amplitud que la distancia entre 6 y 7: *no son medidas métricas*.

F.1. C) LAS ESCALAS DE ESPECÍMENES

Estas escalas listan, en orden creciente, ejemplos famosos de objetos representativos de una cualidad.

Ejemplos:

Para posicionar *el tamaño del bigote*, pueden ser especímenes

(1) el de Charles Chaplin (2) el de Albert Einstein (3) el de Salvador Dalí.

Para posicionar *la magnitud de terremotos* se puede evocar

(1) Como el de Santiago en marzo de 1985
(2) Como el de Talca-Constitución en febrero de 2010
(3) Como el de Valdivia en 1960

Este tipo de escala puede servir cuando se trata de juzgar características como la limpieza de una escritura manual, o la granularidad de una fotografía, o la magnitud de un terremoto. En estos casos, donde no hay una unidad de medición o, si la hay, no es familiar a los utilizadores, se entregan ejemplos concretos (fotografías o descripciones) tanto a los estudiantes como a los jueces.

F.2. Las escalas descriptivas

Son escalas que proveen una descripción para cada uno de sus escalones. En cada nivel se detalla una situación o un desempeño, y usualmente siguen una lógica incremental: el escalón superior incluye los elementos del anterior. El objetivo de definir escalas de este tipo es que las calificaciones de un mismo desempeño, que ha sido evaluado por varios jueces, tengan una menor dispersión.

F.2. A) UNA EXPERIENCIA CONFORTANTE

De Bal, Beckers y De Landsheere (1977, pp. 111 y 212) realizaron una experiencia en Liège, donde tres grupos de jueces calificaron una Composición Francesa y una de Ciencias, hechas por estudiantes de 15 años. Se evaluó de tres modos, distintos para cada grupo: (1) apreciación global, (2) apreciación analítica con escalas, y (3) apreciación analítica con rúbricas con escalones descritos.

Las diferencias (Desviaciones Estándar) entre los jueces de los grupos 1 y 3 disminuyen en los dos ámbitos. Es confortante la hipótesis de que *rúbricas con escalones descritos mejoran la concordancia inter-jueces*.

Se ve también que *las medias mejoran*. No debemos generalizar esta observación, pues el contenido de la prueba influencia el proceso de evaluación. Para algunos contenidos, por ejemplo el dictado para evaluar la ortografía, los correctores tienen un modelo de referencia ÚNICO, y a menudo adoptan una corrección *por disminución o detrimento* (por cada error). En otros contenidos, y es el caso de las competencias, cuando hay VARIAS respuestas aceptables y muchos criterios han sido definidos (como para una composición francesa o un informe sobre una experiencia en un laboratorio de ciencias), se practica más frecuentemente una corrección *por aumentación o añadido*.

Tabla 11: Dispositivo experimental de tres grupos de Nj[1] jueces juzgando dos producciones complejas, una en Francés y una en Ciencias, y los impactos en términos de variaciones (Desviaciones Estándares) entre las notas de los jueces

	Grupo 1 Apreciación global	Grupo 2 Apreciación analítica con escalas	Grupo 3 Apreciación analítica con rúbricas con escalones descritos
Nj en Comp. Francesa (252)	101	70	81
Nj en Ciencias (346)	122	111	113
Criterios		Nombre (1 palabra) de los 3 criterios Aaaaaaa 1 2 3 4 5 Bbbbbb 1 2 3 4 5 Ccccccc 1 2 3 4 5	Nombre (1 palabra) de los 3 criterios Aaaaaa 1 2 3 4 5 Definición de cada escalón Bbbbbbb 1 2 3 4 5 Definición de cada escalón Ccccccc 1 2 3 4 5 Definición de cada escalón
Juicio final idéntico :	Posición sobre una escala de 5 escalones: 1 2 3 4 5		
Comp. Francesa: Media	3,05	3,38	3,50
Dev Estándar	*0,78*	*0,64*	*0,53*
Ciencias: Media	3,02	3,22	3,44
Dev Estándar	*0,68*	*0,74*	*0,59*

[1] Nj = número de jueces.

F.2. B) LA ESCALA DE MERCALLI (DEMOULIN, 1992, P. 63–65) SOBRE LOS TERREMOTOS ES PROBABLEMENTE EL EJEMPLO MÁS FAMOSO DE ESCALA DESCRIPTIVA:

(1) No percibido, salvo por escasas personas en circunstancias favorables (ej: aquellos que poseen un sismógrafo de gran sensibilidad)
(2) Percibido por personas descansando, especialmente en los pisos superiores de las construcciones. Objetos suspendidos pueden oscilar.
(3) Percibido claramente dentro de los inmuebles, pero muchas personas no reconocen que es un terremoto. Automóviles inmóviles pueden oscilar un poco.
(4) Durante el día, muchas personas lo notan dentro de las habitaciones. Platos, ventanas, puertas pueden ser sacudidos, aunque pocas personas se dan cuenta en la calle y espacios exteriores. La sensación parece como cuando un camión pesado hace temblar un inmueble.
(5) Sentido por todo el mundo: a muchas personas los despierta. Algunos platos y algunas ventanas se rompen; yesos de paredes se fisuran; se vuelcan objetos inestables. Se notan modificaciones en la posición (verticalidad) de algunos árboles o palos altos.
(6) Sentido por todo el mundo; muchos se espantan y corren fuera de los inmuebles. Algunos muebles pesados se desplazan; algunos yesos y cornisas caen, y algunas chimeneas son levemente estropeadas.
(7) Todo el mundo se precipita fuera. Los daños son considerables en los inmuebles mal construidos. Es sentido por personas conduciendo automóviles.
(8) Los daños son leves en los inmuebles especialmente construidos para resistir, pero más importantes en inmuebles ordinarios. Otros se desploman. Chimeneas, paredes, columnas, monumentos caen.
(9) Aun construcciones especialmente bien construidas no mantienen la vertical; importantes inmuebles caen. El suelo se fisura. Canalizaciones subterráneas se rompen.
(10) La mayoría de las casas son destruidas desde sus fundaciones; el suelo se fisura con violencia; rieles se curvan y tuercen. Desplazamientos de terreno son considerables.
(11) Casi ninguna construcción resiste. Puentes son destruidos; las fisuras en el suelo son anchas y extensas.
(12) Destrucción total. Olas de tierra son visibles. Objetos son proyectados en el aire.

F.2. C) ¿POR QUÉ USAR RÚBRICAS (ESCALAS CON ESCALONES DESCRITOS) EN EL ÁMBITO EDUCACIONAL?

Las rúbricas son escalas descriptivas que se utilizan en la evaluación de desempeños cuya complejidad demanda más que una corrección "objetiva" (incluso automática) que determine si el estudiante ha dado, o no, LA (única) respuesta correcta posible. Desempeños más complejos (que involucran actuaciones observables, exigen movilización de saberes, y suponen ciertas actitudes) demandan una evaluación subjetiva donde el o los jueces deben observar, estimar y emitir un juicio en tiempos usualmente estrechos. En estos casos *las rúbricas son un apoyo para objetivar la evaluación y calificación que se hace de cada estudiante.*

Algunas rúbricas pueden estar asociadas a una producción en particular (una realización específica en un curso), y por lo tanto detallar en las descripciones de sus peldaños las características esperadas de esa realización. Otras pueden ser concebidas para competencias genéricas, comunes a muchas evaluaciones y transversales al plan de formación, como los ejemplos que se proveen en el libro de Villa y Poblete (2007) para el listado de competencias genéricas del proyecto Alfa-TUNING Europa.

La evaluación de algunos desempeños o producciones complejas requiere el uso de varias rúbricas combinadas. En estos casos se define una escala descriptiva para cada uno de los indicadores que en conjunto dan cuenta de un determinado Resultado de Aprendizaje.

Invitamos al lector a ver algunos ejemplos de rúbricas utilizadas en evaluaciones en farmacia (Capítulo 8, sección C3, tablas 2 y 3) y en producciones artísticas (Capítulo 10, secciones C3.2 y C3.3, tablas 3 y 4).

F.2. D) UNA LÓGICA INCREMENTAL

La mayoría de las rúbricas sigue una lógica incremental, donde el escalón superior supone el logro del anterior, lo contiene y supera, lleva más allá. En este sentido, *la escala completa representa los distintos estadios de desarrollo de un determinado aprendizaje*, mostrando el itinerario esquemático (de un extremo a otro) que puede recorrer un estudiante en el proceso de lograr el resultado y desarrollar la competencia.

Los *párrafos descriptivos* de cada nivel *delimitan y ejemplifican* lo que hace un estudiante hipotético en ese punto de desarrollo de la competencia. Es decir, *describen aquello que constituye evidencias de logro del aprendizaje esperado*[37], o *evidencias de dificultades*[38] en el proceso.

F.2. E) UN PISO MÍNIMO EN EL ITINERARIO (ESQUEMÁTICO) DE DESARROLLO

Esta lógica incremental puede a veces estar ausente en los niveles más bajos de la rúbrica. En algunos casos la escala se inicia describiendo aspectos que no son logrados ("no cumple…", "no logra…", "no expresa…"), hasta que en determinado nivel la si-

[37] Ver los niveles 3, 4 y 5 de la rúbrica que se presenta en la Tabla 12, sección F.2.e.
[38] Ver los niveles 1 y 2 de la rúbrica que se presenta en la Tabla 13, sección F.2.e.

tuación se invierte ("sí logra"). La Tabla 12 presenta un ejemplo de esto: los peldaños 1 y 2 de estas escalas se centran en lo que el estudiante aún no logra, lo que se revierte en el nivel 3; los escalones 4 y 5 describen desempeños destacados, siguiendo una lógica incremental a partir del nivel 3.

Tabla 12: Rúbrica para el indicador n°1 de la competencia "trabajar en equipo" para ALLO-evaluación (del profesor, por un jurado, de pares), con peldaños 1 y 2 descritos "en negativo"

A) INDICADOR 1: REALIZÓ LAS TAREAS ASIGNADAS POR EL GRUPO DENTRO DE LOS PLAZOS REQUERIDOS				
1	2	3	4	5
No cumplió con las tareas asignadas	Cumplió parcialmente las tareas asignadas y/o...	Cumplió las tareas asignadas	→	→
	No cumplió los plazos requeridos	En los plazos requeridos	→	→
			La calidad de su tarea supuso un notable aporte al equipo	→
				Su trabajo orientó y facilitó el del resto de los miembros del equipo

F.2. F) UN PRINCIPIO DE POSITIVIDAD

Frente a la posibilidad descrita en el punto anterior, *es necesario resguardar que en general las descripciones hagan hincapié en lo que el estudiante muestra/hace/realiza antes que en lo que está ausente*. En consecuencia, los peldaños anteriores al piso mínimo deberán entonces concentrarse en aquello que constituye evidencia de debilidades o aprendizajes aún en desarrollo, como lo muestran los escalones 1 y 2 de la Tabla 13.

Tabla 13: Rúbrica para el indicador n° 2 de la competencia "trabajar en equipo" para ALLO-evaluación, con peldaños 1 y 2 descritos "en positivo"

B) INDICADOR 2: PARTICIPA DE FORMA ACTIVA EN LOS ESPACIOS DE ENCUENTRO DEL EQUIPO, COMPARTIENDO LA INFORMACIÓN, LOS CONOCIMIENTOS Y LAS EXPERIENCIAS				
1	2	3	4	5
Se ausenta con frecuencia y su presencia es irrelevante	Interviene poco en el debate, principalmente a requerimiento de los demás	Se muestra activo/a y participativo/a en los encuentros del grupo	→	→
			Con sus intervenciones fomenta la participación y mejora la calidad de los resultados del equipo	→
				Sus contribuciones son fundamentales tanto para el proceso grupal como para la calidad del resultado

F.2. G) RÚBRICAS Y UMBRAL DE ÉXITO

Así, las rúbricas están vinculadas al concepto de umbral de éxito. Uno de los escalones medios de la rúbrica describe lo que se considera como el piso mínimo de logro del aprendizaje involucrado, asociado a la calificación mínima para aprobar, y *constituye así un estándar* (mínimo) de logro o de desarrollo de la competencia. Bajo el estándar los niveles describen evidencias de dificultades en el aprendizaje, y sobre el estándar se ubican evidencias de un logro destacado, que muestran profundización, consolidación y proyección de aspectos del desarrollo de la competencia.

F.2. H) RÚBRICAS Y LA PROFESIÓN

Los niveles superiores de la rúbrica (de todas las rúbricas que se utilizan en una carrera o programa determinado) pueden llegar a representar una parte del *conjunto de criterios de calidad* de la reflexión y la actuación que definen el *racional de la profesión y/o la disciplina*, de acuerdo con la visión que promueve esa comunidad y esa institución. De esta forma, las rúbricas se transforman en un instrumento que contribuye a lograr uno de los objetivos relevantes de cualquier proceso formativo en la educación superior: compartir con el estudiante un *conjunto de criterios de calidad de la reflexión y la actuación profesional y ciudadana*, y a la vez *hacerlo partícipe de una comunidad* que comparte estos estándares y criterios.

F.2. I) IMPACTOS DE LAS RÚBRICAS

Las rúbricas pueden tener *impactos positivos* tanto para los profesores como para los estudiantes. A los primeros les permite *objetivar la evaluación de desempeños complejos*, permitiendo el *acuerdo entre jueces* y *disminuyendo la dispersión* de las calificaciones. Para los estudiantes significa *mayor transparencia* en la evaluación (mayor validez Deontológica) y la oportunidad de una *retroalimentación de calidad* que favorezca su aprendizaje (siempre que el docente decida tomar esta oportunidad). El hecho de que la calificación se explique mediante descripciones de desempeños favorece la detección de fortalezas y debilidades, y permite el autodiagnóstico (validez Informativa) y la autorregulación (validez Consecuencial). Al proveer niveles inferiores y superiores al estándar mínimo entrega contraejemplos (lo que no hacer) y muestra lo que es esperable en un nivel superior de desarrollo de la competencia y logro del aprendizaje; muestra al estudiante un camino a través del cual mejorar.

F.2. J) ¿CUÁNTOS ESCALONES DEBE TENER UNA RÚBRICA?

En muchos casos, más de 5 escalones superan la capacidad que tenemos para discriminar, tanto para el caso de concebir la rúbrica como para utilizarla en un momento de evaluación (mientras se observa al estudiante, por ejemplo). En caso de ser necesario, se

puede tener la opción de ubicar el desempeño de un estudiante determinado en un punto intermedio entre dos escalones, y asignar puntaje de forma acorde. Tres escalones o menos parecen poco para describir una evolución esquemática de un logro de aprendizaje (considerando desde un estado de ausencia de aprendizaje hasta un logro destacado), haciendo las distinciones muy gruesas y poco diagnósticas (para profesores y estudiantes). Por estas razones nos inclinamos por considerar óptimas rúbricas de 4 o 5 escalones.

F.2. k) ¿Quién utiliza las rúbricas?

Una misma rúbrica puede ser adaptada para la *ALLO-evaluación* (ya sea *por el profesor, por un jurado externo o por los pares*), como los ejemplos de las tablas 12 y 13, y para la *IPSO-evaluación* (por el *propio estudiante*), como el ejemplo de la Tabla 14.

Tabla 14: Rúbrica para el indicador n° 1 de la competencia "trabajar en equipo" para IPSO-evaluación (autoevaluación por el propio estudiante)

A) Indicador 1: Realicé las tareas que me fueron asignadas por el grupo dentro de los plazos requeridos				
1	2	3	4	5
No cumplí con las tareas asignadas	Cumplí parcialmente las tareas asignadas y/o...	Cumplí las tareas asignadas	→	→
	No cumplí los plazos requeridos	En los plazos requeridos	→	→
			La calidad de mi tarea supuso un notable aporte al equipo	→
				Mi trabajo orientó y facilitó el del resto de los miembros del equipo

G. Conclusiones

La docimología crítica nos muestra la gran cantidad de sesgos y efectos distorsionadores de la evaluación a los que estamos expuestos como profesores. Estas distorsiones pueden afectar la exactitud y la ecuanimidad de la evaluación. Tomar conciencia de ellos es el primer paso para lograr evitarlos o minimizarlos. Los pasos siguientes implican (i) reflexionar sobre nuestros instrumentos de evaluación a la luz de los criterios de calidad (ETICPRAD), (ii) concertar con nuestros colegas los criterios con que evaluaremos a los estudiantes, sus definiciones y pesos relativos, y (iii) la elaboración y uso de rúbricas cada vez que sea posible y pertinente.

Referencias

Agazzi, A. (1967). *Les aspects pédagogiques des examens.* Strasbourg: Conseil de l'Europe.

Bonniol, J.-J. (1965). Les divergences de notation tenant aux effets d'ordre de la correction. Cahiers de Psychologie, 8, 181-188.

Britton, J. (1963). Experimental Marking of English Composition Written by Fifteen-Year-Olds. In Educational Review, Birmingham, vol. 16, 1, 17-23.

Caverni, J.-P., Fabre, J.-M., Noizet, G. (1975). Dépendance des évaluations scolaires par rapport à des évaluations antérieures: études en situation simulée. *Le Travail Humain*, 38(2), 213-222.

Chase, C. (1968). The impact of some obvious variables on essay test scores. In Journal of Educational Measurement, 5, p. 315-318.

De Landsheere, G. (1971, 1992). *Evaluation continue et examens. Précis de Docimologie.* Bruxelles: Editions Labor et Paris: Fernand Nathan.

De Bal, R., Beckers, J. y De Landsheere, G. (1977). Construire des échelles d'évaluation descriptives, Bruxelles: Direction Générale des Etudes, Collection Pédagogie et Recherche, n° 6.

Demoulin, A. (1992). Les tremblements de terre. Parcs nationaux, 47 (3/4), p. 56-76.

Englehard, G. JR. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. Journal of Educational Measurement, summer, vol. 31, 2, 93-113.

Hartog, P. y Rhodes, E.C. (1936). An examination of Examinations. Londres : Mac Millan, 2e édition.

Hutmacher, W. (1993). Quand la réalité résiste à la lutte contre l'échec scolaire. Analyse du redoublement dans l'enseignement primaire genevois. Genève: Service de la Recherche sociologique, *Cahier n°36.*

Laugier, H., Piéron, H., Toulouse, E. y Weinberg, D. (1934). *Etudes docimologiquessur le perfectionnement des examens et concours.* Paris: Conservatoire national des arts et métiers, Publications du *Travail humain*, Série A, n°3.

Leclercq, D. y Hubert, S. (1990). L'évaluation des productions. Chap 3 de D. Leclercq. Evaluation et docimologie. Editions de l'Université de Liège.

Noizet, G. y Caverni, J.-P. (1978). Psychologie de l'évaluation scolaire. Paris: Presses universitaires de France.

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

Piéron, H., Reuchlin, M. y Bacher, F. (mars-juin 1962). Une recherche expérimentale de docimologie sur les examens oraux de physique au niveau du baccalauréat de mathématique, in Biotypologie.

Reuchlin, M. (1959). Recherches docimologiques du service de recherches de l'Institut National d'Etude du Travail et d'Orientation Professionnelle (INETOP). Le Travail Humain, 22, p. 11-18.

Rot, N. y Butas, Z. (1959). Les distributions des notes scolaires comparées eux distributions des résultats obtenus aux tests de connaissances. *Le travail humain, XXII,* 1-2.

Trimble, O. (1934). The Oral Examination: Its Validity and Reliability, School and Society, vol. 39, no. 1009, April 28, 1934, pp. 1-3.

Villa, A. y Poblete, M. (2007). Aprendizaje basado en competencias: una propuesta para la evaluación de las competencias genéricas. U. de Deusto, Bilbao.

Weiss, R. (1969). The reliability of the number marking system: An Australian Study. In J. Lauwerijs y D. Scalon, Examinations. London: Evans Brothers. p. 101-107.

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