Capítulo XVI

Autoevaluación con grados de certeza: Un microscopio para la evaluación de los aprendizajes

Dieudonné Leclercq

"*El problema en este mundo es que los necios están seguros de todo y los sabios llenos de duda*".
Bertrand Russel

"*Doloroso no es lo que ignoramos; es lo que sabemos... pero es falso.*"
Mark Twain

Este capítulo se inspira mucho en el artículo publicado por Leclercq en la revista ETP 2009 con el título: *La connaissance partielle chez le patient: Pourquoi et comment la mesurer.*

A. Los seis desafíos en la utilización de los grados de certeza

Al solicitar a un estudiante que indique cuán seguro está de la respuesta entregada, es decir, que añada, a cada una de sus respuestas en una evaluación, un grado de certeza sobre la exactitud de la misma, se abordan varios desafíos simultáneos. A continuación desarrollaremos seis de ellos.

A.1. Desafío epistemológico: la definición de "dominio"

Los grados de certeza permiten al estudiante expresar sus dudas, tal como ocurre en el mundo profesional. Muchas veces el estudiante está consciente de su grado de certeza en un conocimiento (por ejemplo, "poco seguro"), pero el sistema de evaluación habitual le prohíbe expresarlo. Con demasiada frecuencia se prohíbe, en el momento de la evaluación, expresar una duda, aun cuando sea fácil verificar que la toma de conciencia acerca de una duda inicia comportamientos de búsqueda de información, y por lo tanto debiese ser fomentada y no inhibida en el proceso formativo.

Por ejemplo, Leclercq y Boskin (1990) han permitido a 50 estudiantes aprender un contenido con la ayuda de una híper-media en un computador, y tomar notas personales (informatizadas) para cada pantalla. Después esos estudiantes rindieron, en dos ocasiones –a modo de pre y post-test– una prueba de 15 preguntas con grados de certeza[119].

[119] En una escala asimétrica donde los grados posibles eran (*0*) 0-25%; (*1*) 25-50%; (*2*) 50-70%; (*3*) 70-85%; (*4*) 85-95%; (*5*) 95-100%.

Entre el pre-test y el post-test se permitió a los estudiantes (que no habían sido informados de las respuestas correctas) consultar sus apuntes, proceso que fue monitoreado *on line*. Esta consulta mejoró las tasas de éxito y los grados de certeza:

Tabla 1: Mejoras objetivas y subjetivas que resultan de la consulta de apuntes

	PRE	POST	Mejora	Ganancia Relativa
Tasa media de éxito	34%	58%	+24%	36%
Certeza media	49%	71%	+22%	43%

Pero lo más interesante es que se confirma la hipótesis según la cual la tasa de consulta de los apuntes depende de:

(1) la exactitud (o no) de las respuestas en el pre-test. En este caso, 35% de las consultas se vincularon con respuestas correctas en el pre-test, y 65% con respuestas incorrectas.

(2) el grado de certeza en el pre-test. En la Figura 1 se ve que bajo 80% de certeza la tasa de consulta llega al límite de 60% (¡es enorme!). Pero cuando los estudiantes están más seguros, se observa una función inversa: mientras menos seguros, más consultan. La tasa de consulta baja hasta 26% en las respuestas que tenían un grado de certeza máxima (aquí 97,5%). Estos datos evidencian que los estudiantes saben distinguir entre respuestas correctas e incorrectas, y que eso afecta sus comportamientos.

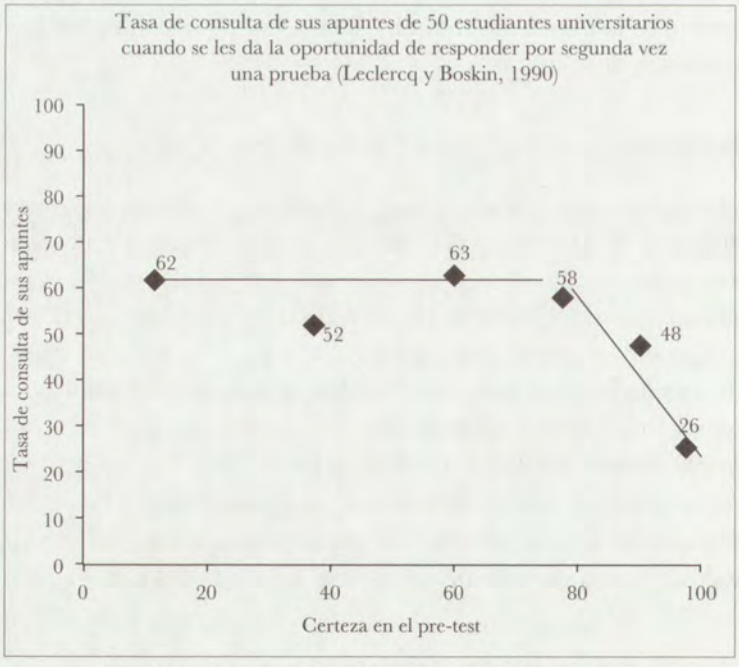


Figura 1: Relación entre certeza y consulta de apuntes

A pesar de esta evidencia, los estudiantes se quejan de la tradición impuesta que los fuerza o bien a contestar callando su grado de certeza (es decir, el corrector considera cada respuesta como si fuera dada con el máximo grado de certeza), o bien a omitir (es decir, el corrector considera que el estudiante es completamente ignorante).

Mi postulado epistemológico es el de Bruno De Finetti[120] (1965, p. 109):

El conocimiento parcial existe. Detectarlo es necesario y factible.
Además, preparar a los estudiantes (para esta detección) tiene un gran valor educativo.

Propongo la siguiente formulación:

> A menudo, *aprender* algo no es pasar de una ignorancia total a un dominio total, sino que es pasar de un estado de conocimiento parcial a otro estado de conocimiento menos parcial, siendo la ignorancia total y el conocimiento perfecto casos particulares (y extremos) del conocimiento parcial.

A.2. Desafío de medición en la investigación: la necesidad de un microscopio para el pensamiento

En la literatura de investigación sobre el aprendizaje, cuando se comparan diferentes métodos, con frecuencia las conclusiones son *NSD: Non Statistical Differences* (no hay diferencias estadísticamente significativas). Mi postulado es que:

> Con frecuencia existen diferencias, pero estas no han sido medidas con un instrumento lo bastante sutil. Los grados de certeza puede ser un *microscopio sobre el pensamiento.*

Durante demasiado tiempo los investigadores en educación han trabajado como biólogos que no contaran con microscopios, o como químicos que trabajaran sin balanza. *El grado de precisión del instrumento debe ser apropiado al grado de sutileza del objeto medido* (en este caso, los procesos mentales). Una experiencia ilustra esto: Leclercq *et al.* (1998), presentaron "casos" (situaciones grabadas en video) de conflictos auténticos en una clase entre el profesor y algunos estudiantes. Después de cada episodio (de 30 segundos),

(1) el video es interrumpido y la pregunta del *pre-test* es "¿Qué hizo el estudiante?" o "¿Qué hizo el docente?" (Se provee una lista de soluciones entre las cuales está la correcta)[121]
(2) los estudiantes contestan (en un computador) *con grados de certeza*,
(3) hay un *debate* (de 3 a 5 minutos) sobre el asunto
(4) los estudiantes contestan otra vez (*post-test*) la misma pregunta que en el pre-test.

[120] Y de varios otros autores como Van Naerssen (1962); Shuford *et al.* (1966) y Leclercq (1982, 1993).

[121] Es el sistema de los "casos programados" inventado por De Waele (1975). Ver Leclercq y Vanden Brande (1997).

Para uno de estos episodios, 22 de los 23 estudiantes contestaron correctamente en el pre-test, con un grado de certeza promedio de 58,6%, y los mismos lo hicieron en el post-test con un grado de certeza promedio de 75%, lo que es una mejora importante (40% de ganancia relativa). El único estudiante que se equivocó lo hizo en el pre-test con 77% de certeza y en el post-test con 12% de certeza, que es también una *ganancia relativa* importante: 38% (ganó 65% sobre los 177% de *ganancia posible*). Lo anterior se muestra en la Figura 2.

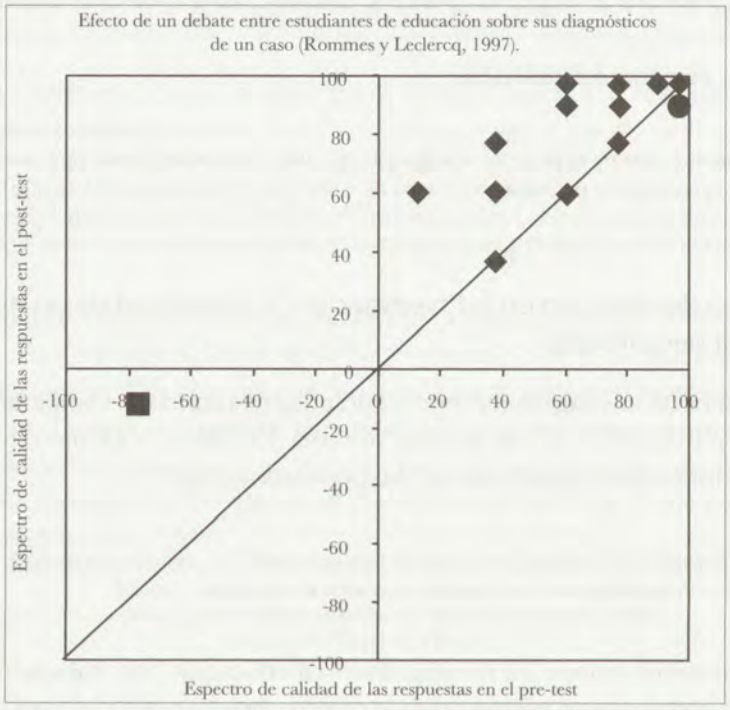


Figura 2: Ganancias en seguridad en las respuestas después de un debate

Estos cambios son *invisibles para un observador que no utiliza el microscopio que constituyen los grados de certeza.* Quien no los utiliza solo constata que no hay ningún cambio en las respuestas (se mantienen correctas), diciendo "No hay modificación alguna. El debate no ha servido de nada" (*el fenómeno "NSD" otra vez*) y concluye: "Se puede (debe) suprimir el debate".

A.3. Desafío de denominación de las mejorías (o empeoramientos): las tasas de Confianzas utilizables y de Imprudencias peligrosas

Autoevaluarse ayuda en el autodiagnóstico de los procesos mentales, como lo demuestra el Capítulo 9 dedicado a la metacognición. Sin embargo, no todos los conocimientos son útiles para la actuación. Al igual que Hunt (1993), distingo entre los conocimientos

- utilizables: correctos con alta certeza (en mis cursos[122], sobre 50%).
- peligrosos: incorrectos con alta certeza (en mis cursos, sobre 50%).
- inutilizables: correctos o incorrectos, pero no lo bastante seguros como para que la persona los use (en mis cursos, bajo 50%).

Llamo *% de Confianza utilizable* a la tasa (el %) de las respuestas correctas con una certeza >50%, y *% de Imprudencia peligrosa* a la tasa (el %) de las respuestas incorrectas con una certeza >50%. Estas distinciones permiten interpretar los resultados de una prueba en forma más sutil.

Por ejemplo, una de nuestras estudiantes (Lucas, 2001) midió las respuestas de 300 estudiantes antes (PRE) y después (POST) de un video sobre "las acciones que pueden salvar vidas" (en casos de emergencia). En la pregunta 15 (fractura de pierna), las tasas de respuestas fueron las siguientes:

Tabla 2: La evolución PRE–POST interpretada sin uso de grados de certeza en la pregunta 15 de Lucas (2001).

	incorrectas	correctas
PRE	37%	63%
POST	12%	88%

Esta mejora parece interesante: la *Ganancia Relativa* es 67% (es decir, 25% sobre un 37% de mejora posible).

Si consideramos los grados de certeza, los resultados son más sutiles:

Tabla 3: Las respuestas del pre–test interpretadas con grados de certeza

	incorrectas		correctas		
PRE	37		63		100%
	27	10	14	49	100%
Certeza	mayor a 50%	menor a 50%	menor a 50%	mayor a 50%	
	peligrosas	inútiles		útiles	

En un 78% las respuestas correctas son utilizables (49/63) ¡no el 100%! *(% de Confianza utilizable* = 77,7%).

Sobre el 37% de errores ¡73% (27/37) es peligroso! *(% de Imprudencia peligrosa* = 72,9%).

[122] Digo "en mis cursos" porque los umbrales de satisfacción (de exigencia) pueden variar de acuerdo con los ámbitos (ver sección A.4).

Tabla 4: Las respuestas del post-test interpretadas con grados de certeza

	incorrectas		correctas		
POST	12		88		100%
	7	5	4	84	100%
Certeza	mayor a 50%	menor a 50%	menor a 50%	mayor a 50%	
	peligrosas	inútiles		útiles	

Ahora 95% (84/88) de las respuestas correctas es utilizable *(% de Confianza utilizable* = 95,4%).

Sobre el 12% de errores ¡solo 58% (7/12) es peligroso! *(% de Imprudencia peligrosa* = 58,3%).

Estas nuevas estadísticas proveen una imagen diferente de los progresos y una base diferente para interpretar los resultados. Entregan información más sutil sobre las fortalezas y debilidades de cada estudiante, lo que permite aconsejarle de forma más pertinente.

A.4. Desafío de definición de los umbrales de satisfacción y de éxito

En algunas profesiones la gestión del conocimiento es de importancia vital. Debido a esta condición, cuando se trata de evaluar el grado de dominio, se pide a los estudiantes y a los profesionales que añadan un grado de certeza a sus respuestas. Lo anterior ha sido implementado en la formación para varias profesiones, observándose que el *umbral de satisfacción* (o de "utilizabilidad del conocimiento") varía entre ellas.

Con pilotos y mecánicos de aviones[123] el umbral se fijó en 100%: sus respuestas eran consideradas como satisfactorias si eran 100% correctas, cada una con *100%* de certeza.

En medicina de urgencia[124] y con enfermeras (incluidas las que practican las transfusiones de sangre[125]), el umbral también se fijó en *100%*.

En un programa de formación de pacientes diabéticos[126] para el autocuidado, el umbral fue fijado en *80%* por los especialistas.

Aunque este capítulo está dedicado a la utilización de grados de certeza con estudiantes, se proveerán algunos ejemplos de otros ámbitos para ilustrar algunos conceptos.

El umbral de satisfacción es arbitrario y depende del contenido y de las exigencias. Por ejemplo, en mis cursos de primer año de pregrado en la universidad lo he fijado en *50%*. Pero puede ser modificado en el caso de que sean estudiantes de posgrado quienes contesten a esta misma prueba en los mismos saberes.

123 Leclercq (1975, 1982).
124 Leclercq y Micheels (2010).
125 Colaboración Leclercq - Hospital Tenon en París.
126 Brutomesso *et al.* (2003), Leclercq *et al.* (2003), Reach *et al.* (2005).

La decisión del *umbral de éxito* puede ser apoyada por procesos como el de Angoff (1971, p. 515), quien pide a expertos decidir, para cada pregunta, cuál es la probabilidad de éxito de un estudiante "mínimamente competente"[127], y definir el umbral de éxito en la prueba como la suma de las probabilidades promedio de los jueces. "Para ayudar a los jueces a imaginar lo que es un *estudiante mínimamente competente*, se les invita a imaginar los errores que tolerarían en personas reputadas como competentes, o a pensar en los errores que cometen ellos mismos sin por eso estimarse incompetentes" (V. de Landsheere, 1988, p. 148).

A.5. Desafío de medición: umbrales personales de respuesta

Más y más docentes piensan que una respuesta correcta con un alto grado de certeza debe recibir una nota mejor que una respuesta correcta con un grado de certeza bajo, y que una respuesta incorrecta con un grado alto de certeza debe ser sancionada más que una respuesta incorrecta con un grado de certeza bajo.

Es el credo de De Finetti (1965, p. 111), que dice:

Solo la probabilidad subjetiva puede dar una significación objetiva
(1) a cada respuesta
(2) y a cada método de puntuación.

El punto (1) se basa en el razonamiento presentado en la sección A.1: la interpretación de una respuesta depende de la consigna de la pregunta, de acuerdo a si prohíben, permiten o imponen los grados de certeza. Cuando se prohíben los grados de certeza, *cada estudiante aplica su propio umbral de respuesta*, decidiendo contestar u omitir. Este umbral, que varía entre los estudiantes, no es conocido por el profesor.

El punto (2) se ha presentado en esta sección A5, y continúa con la concepción de un sistema de asignación de puntaje que incluye no solo la exactitud de la respuesta, sino también el grado de certeza asociado. A este tipo de respuesta, que incluye tanto la respuesta a la pregunta como la seguridad en la misma, le llamaremos respuesta *acertinada*[128].

A.6. Desafío de docimología: cómo asignar puntaje

A continuación, un ejemplo. La Tabla 5 muestra los resultados en una prueba de cuatro estudiantes ficticios, que con el método clásico obtienen la *misma nota:* 10 puntos sobre un máximo de 20, pues todos han dado 12 respuestas correctas en una prueba de 20 preguntas (respuesta correcta = +1; respuesta incorrecta = –0,25). Sin embargo,

[127] O el porcentaje de los estudiantes mínimamente competentes que obtendría un éxito.

[128] ¿Por qué un neologismo (también en francés)? Porque lo necesitamos. ¡Inventémoslo antes de que tengamos que traducirlo del inglés!

sus distribuciones de certezas no son las mismas. Si se considera este factor, de acuerdo con los principios para asignar puntaje que se desarrollarán en el Capítulo 17, los cuatro estudiantes reciben nuevas notas, diferentes entre ellas.

Tabla 5: Las calidades de las respuestas y las notas de cuatro estudiantes ficticios

	incorrectas		correctas		nota clásica	nota nueva
	8		12			
Estudiante A	5	3	6	6	10	10
Estudiante B	2	6	3	9	10	12
Estudiante C	6	2	2	10	10	11
Estudiante D	2	6	10	2	10	11
Certeza	mayor a 50%	menor a 50%	menor a 50%	mayor a 50%		
	peligrosas	inútiles		útiles		

Se observa que, en este ejemplo

- la nota o puntaje clásico (10) es inferior al número de respuestas correctas (12) pues se aplicó una penalización por los errores (–0,25 por cada respuesta incorrecta).
- las nuevas notas son siempre iguales o más altas que las clásicas, porque se aplicó un principio de "plus metacognitivo", que el estudiante debe "ganar" por la calidad subjetiva de su conocimiento (ver Capítulo 17).
- el estudiante A no ganó ningún "plus metacognitivo" porque su uso de los grados de certeza no fue satisfactorio: su *% de Confianzas utilizables* no estuvo sobre el 50% (6/12) y su *% de Imprudencias peligrosas* fue mayor que 50% (5/8).
- el estudiante B obtuvo 2 puntos de "plus metacognitivo" pues se los ganó con justo merecimiento: su *% de Confianzas utilizables* fue mayor que 50% (9/12) y su *% de Imprudencias peligrosas* estuvo por debajo de 50% (2/8).
- el estudiante C ganó solo un punto de "plus metacognitivo", porque si bien su *% de Confianzas utilizables* fue mayor que 50% (10/12) sus *Imprudencias peligrosas* fueron mayores que 50% (6/8).
- el estudiante D también ganó un punto de "plus metacognitivo", pues su *% de Confianzas utilizables* no estuvo por sobre 50% (2/12), pero su *% de Imprudencias peligrosas* fue menor que 50% (2/8).

Este ejemplo constituye UNA forma de cambiar el principio de asignación de puntaje, tomando en cuenta la calidad subjetiva del conocimiento junto a la exactitud de la respuesta, pero puede haber otras. Esta propuesta específica será explicada en detalle en el Capítulo 17.

B. ¿Cómo recolectar los grados de certeza de los estudiantes? Cinco principios para asegurar la Validez Teórica

B.1. Principio 1: Probabilidades (porcentajes) y no escalones verbales

Varios autores[129] han demostrado que los grados de certeza deben ser expresados en términos de probabilidades (de 0 hasta 1) o de porcentajes (de 0 hasta 100), y no en expresiones verbales del tipo "nada seguro", "poco seguro", "medianamente seguro", "muy seguro", etc.[130], debido a los múltiples significados que las personas asignan a las mismas palabras. Es fácil reproducir la experiencia de Fabre (1993), quien demuestra que las mismas palabras son significadas de forma diferente por personas variadas. Para esto pidió a 143 personas que atribuyeran una nota entre 0 y 10 a cinco expresiones verbales, obteniendo los siguientes resultados:

Tabla 6: Repartición de los valores numéricos para una misma expresión verbal

Grados de certeza——>		0	1	2	3	4	5	6	7	8	9	10	M	Ec-T
Afirmo que	143			1			2	2		11	25	102	9,49	1,12
Estoy seguro que	143					1	1		3	17	44	77	9,31	0,97
Pienso que	143		1	2	4	3	31	39	30	21	11	1	6,33	1,57
Me parece que	143		7	19	17	25	39	23	9	3	1		4,37	1,72
Supongo que	119	3	7	17	26	32	2	16	11	4	1		4,13	1,84
		3	15	39	47	61	75	80	53	56	82	180		

Al no estar expresadas en una escala numérica, las consignas que usan palabras no permiten medir el *realismo* (ver Capítulo 17), la sobrestimación y la subestimación (o al menos dificulta enormemente esta medición); tampoco permite posicionar la Respuesta-acompañada-de-grado-de-certeza (o mi neologismo: *"respuesta acertinada"*) en el espectro de calidad de las respuestas (ver sección C).

Leclercq ha observado que incluso con alumnos de 12 años es fácil utilizar la expresión "1 posibilidad sobre 5", probablemente porque hay 5 dedos. Sin embargo, la facilidad de uso no implica el realismo.

B.2. Principio 2: Evaluar el conocimiento parcial escondido en las omisiones y los umbrales individuales de respuesta

Con frecuencia los evaluadores introducen la posibilidad de contestar "No sé"[131]. Por el contrario, nuestra recomendación es introducir en la consigna: "Cuando Ud. no

[129] Van Naerssen (1962), Shuford *et al.* (1966), De Finetti (1965), Leclercq (1975, 1982).
[130] Forma que, lamentablemente, utilizaron pioneros como Cooke (1906), en meteorología.
[131] En inglés, *I don't know*.

sepa, indíquelo eligiendo la certeza 0%, y conteste a pesar de que Ud. piense que no sabe. Esto permitirá medir su conocimiento parcial, y Ud. no será penalizado de ninguna forma" (ver Capítulo 17).

Lo anterior parece complicado, pero se basa en los mismos principios expuestos previamente, sobre tener en cuenta el conocimiento parcial y concebir un modo de respuesta que sea conforme a la teoría, es decir, la visión epistemológica que tenemos del saber.

A continuación, dos ejemplos donde los evaluadores han permitido la expresión del conocimiento parcial.

A) LA EXPERIENCIA DE "LAS X Y LOS 0" DE SANDERSON (1973)

Este autor testeó a 120 estudiantes que finalizaban su formación en medicina mediante tres pruebas, cada una compuesta de 100 Preguntas Verdadero-o-Falso (resultando en 36.000 respuestas). Les pidió contestar sobre una hoja con dos celdas para cada pregunta: la celda "Verdadero" y la celda "Falso".

En una primera fase, los estudiantes tenían que escribir una cruz (x) para sus respuestas "definitivas", y dejar vacías las dos celdas cuando "no saben". La consigna precisaba que obtendrían +1 punto por cada respuesta correcta y una penalización de −1 punto por cada error.

En una segunda fase, sobre las hojas de respuesta ya contestadas en forma "definitiva", los estudiantes fueron invitados a escribir un 0 en las celdas vacías, indicando la respuesta que hubiesen dado de haber sido obligados a contestar todas las preguntas. La consigna precisaba que esos 0 no serían considerados en el cálculo de la calificación.

De las 5.367 respuestas 0 que se obtuvieron como resultado, 53,7% eran correctas (¡sobre 50%!). Los mejores estudiantes obtuvieron una tasa de respuestas correctas de 55,3%, y los peores una tasa de 52,5%. Sanderson observa que:

(1) estas diferencias positivas, sobre el 50%, son un indicador de un "residuo de conocimiento correcto" (lo que preferimos llamar *conocimiento parcial*) y mientras más competentes los estudiantes, más los desfavorece la consigna "No conteste cuando no sabe".
(2) si los 0 hubieran sido considerados en el cálculo del puntaje, el orden de los resultados de los estudiantes no hubiera cambiado mucho en general, pero para algunos estudiantes la diferencia hubiera sido amplia.
(3) la personalidad del estudiante (tomar o no tomar riesgos) juega un rol importante en la decisión de contestar (marcar una x) u omitir, porque si se equivoca va a recibir un castigo de −1 punto, lo que es enorme. Mientras más se reduce esta penalización, menos interviene la personalidad.

b) Los "bolígrafos de tinta azul y de tinta roja" de Cross y Frary (1977)[132]

Estos investigadores aplicaron el mismo dispositivo experimental del ejemplo anterior, pero con PSM clásicas (una sola respuesta correcta) de 4 soluciones, en lugar de PFV-no sé, de modo que existe un 25% de probabilidad de obtener una respuesta correcta al azar cuando el estudiante "No sabe".

En la fase 1 los estudiantes debían contestar (con un bolígrafo de tinta azul) y dejar sin respuesta las preguntas en las cuales "No saben". En la fase 2 fueron invitados a contestar con un bolígrafo de tinta roja a las preguntas que no habían contestado con el bolígrafo de tinta azul. La tasa de respuestas correctas con la tinta roja fue de 30%, lo que es mucho más que el 25% esperado. Estos resultados son otra demostración de:

(1) la existencia del conocimiento parcial
(2) la inadecuación de la consigna "Conteste solo si Ud. sabe, y no conteste si Ud. no sabe"
(3) la influencia perturbadora de la *correction-for-guessing* clásica, que fomenta la inhibición de los estudiantes, al hacer diferencias entre omitir (no se pierde ningún punto) y equivocarse (se pierden puntos).

Cross y Frary presentaron a estos mismos estudiantes preguntas de *meta-metacognición*[133]:

(1) "¿Qué hace Ud. cuando no está seguro de su repuesta?" y
(2) "Durante este test, ¿qué ha decidido Ud.?"

En base a las respuestas, distinguieron entre:

- "*santos*": los que en la pregunta (1) eligieron la respuesta "Adivino solo si tengo una preferencia por una solución o si puedo eliminar al menos una solución" (60%) y en la pregunta (2) "He seguido las instrucciones de las consignas" (43%),
- "*pecadores*": los que, en la pregunta (1) eligieron "Doy una respuesta, al azar si es necesario" (34%) y en la pregunta (2) "He adivinado más frecuentemente que lo que las consignas recomiendan" (42%).

El 30% de éxito en las respuestas con tinta roja se distribuía de forma diferente entre los estudiantes: los "santos" tuvieron un éxito promedio de 31,8% y los "pecadores" de 28,5%. Las omisiones son puntos perdidos, y el ejemplo evidencia que los estudiantes que siguen la consigna de omitir se "castigan" (se privan de más puntos) que aquellos que no las siguen.

[132] En Leclercq (1982), pp. 129-130.
[133] Expresión de Buratti y Allwood (2012).

B.3. Principio 3: No más de 7 escalones

Los seres humanos somos limitados en nuestra capacidad de distinguir fiablemente entre varios escalones del *continuum* que va de 0% hasta 100%. Leclercq (1982, pp. 241-256) estudió este asunto pidiendo a 300 docentes de secundaria contestar[134] 100 preguntas[135], usando tres escalas de grados de certeza (4 escalones[136], 10 escalones[137], 40 escalones[138]). Dos meses después les entregó otra vez las 100 preguntas y sus 100 respuestas de dos meses atrás, pero sin sus grados de certeza, pidiéndoles producir otra vez esos grados de certeza. Estudió los "gráficos de replicabilidad" o "de fiabilidad", y ha descubierto que la fiabilidad es buena para 4 escalones, pero no lo suficiente para 10. Demostró que no sirve de nada el permitir al estudiante contestar con grados de certeza más detallados que 6 o 7 niveles en el *continuum*.

En ciencias, *la precisión no puede ser más grande que la exactitud (o error de medición)*. Después de haber utilizado muchas consignas diferentes (Leclercq, 2003) recomiendo utilizar la siguiente consigna:

Con PRBs y PSMs[139], "Elige un Grado de Certeza (GdC) entre: *5% 20% 40% 60% 80% 95%.* Si Ud. ignora completamente la respuesta, indíquelo con la certeza 5%, y después conteste para que se pueda medir su conocimiento parcial.

Con Preguntas Verdadero-Falso, la escala se limita a: *50% 60% 80% 100%.* Si Ud. ignora la respuesta, indique 50%, y conteste.

Figura 3: La consigna Leclercq (2014).

Esta consigna (o escala) corresponde a la que De Finetti (1965, pp. 102-109) llama *five stars system* (*sistema de las 5 estrellas*) donde el estudiante acompaña su respuesta con 0 hasta 5 estrellas, cada estrella pesando 20% de certeza.

B.4. Principio 4: Un sistema de cotejo apropiado (proper scoring rule)

Este asunto es la segunda razón del fracaso de las investigaciones sobre grados de certeza en educación durante los años 1960-1970 en EE.UU. El problema es sutil, y en los últimos 50 años ha recibido varias soluciones inadecuadas, que los docentes deben conocer para que no re-inventen algunas de estas soluciones que, aunque atractivas, tie-

[134] En un ámbito de conocimiento donde no se puede olvidar y tampoco aprender en dos meses: las estadísticas de la lengua francesa (la probabilidad de que una letra siga a otra en una oración).

[135] Adivinar (con un grado de certeza) la letra que sigue en una oración truncada al azar (el Leclercq Guessing Game).

[136] 0-25%, 25-50%, 50-75%, 75-100%.

[137] 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%.

[138] 0%, 2,5% ; 5% ; 7,5% ; 10% ; 12,5%, etc.

[139] PRBs=Pregunta de respuesta breve; PSMs= Pregunta de selección múltiple.

nen graves debilidades escondidas. Por estas razones, un capítulo completo (el 17) está dedicado a las formas de calificar (asignar puntajes a) las respuestas acertinadas (es decir, respuestas con grado de certeza).

B.5. Principio 5: Entrenar a los estudiantes

Si los grados de certeza desean ser utilizados al final del curso, como parte de las evaluaciones con función certificadora, es necesario crear las condiciones para que los estudiantes se ejerciten sistemáticamente en el uso de los grados de certeza. El Capítulo 9, dedicado a los TEMS (Tests Espectrales Metacognitivos), muestra un ejemplo de esto. En otras situaciones, donde no existe la preocupación porque la evaluación cumpla una función sancionadora, los grados de certeza con función formativa han sido utilizados logrando amplios beneficios. Fue el caso del proyecto MOHICAN (ver sección C de este capítulo), de las pruebas utilizadas con pacientes diabéticos (ver sección C de este capítulo) y de varias otras innovaciones.

C. El espectro de calidades de las respuestas

C.1. Los dos hemiespectros

El espectro completo permite posicionar cada respuesta acertinada según sus dos cualidades: la exactitud y la certeza. El hemiespectro de la izquierda (que va de –100 hasta –0) permite posicionar las respuestas erróneas. El hemiespectro de la derecha permite posicionar las respuestas correctas. En el centro están ubicadas las omisiones (ausencia de respuesta), aunque no se recomiendan, pues son la mancha ciega de la evaluación: no permiten medir el grado de conocimiento parcial, tampoco conocer el umbral de respuesta personal del estudiante (ver sección A5). Un ejemplo de este espectro se muestra en la Figura 4.

Se presenta en una tabla "plegada" y con los datos en cifras destacadas:

Tabla 7: Hemiespectros de calidades "plegados" de 175.725 respuestas acertinadas en el test MOHICAN[140] de vocabulario: 45 PSMs (de 5 soluciones + Ninguna + Todas) por 3.801 estudiantes (Leclercq, 2003, p. 72)

	0	20	40	60	80	100	OM
Porcentaje de Respuestas Incorrectas (NRI)	4	6	10	9	8	8	5
Porcentaje de Respuestas Correctas (NRC)	4	5	7	10	10	13	

[140] En este proyecto (Leclercq, 2003) fueron testeados 4.000 estudiantes ingresando a las universidades francófonas de Bélgica, en octubre de 1999, con 10 pruebas constituidas de PSMs (con 5 soluciones + Ninguna + Todas) y grados de certeza.

Desplegando esta tabla se obtiene el espectro completo con 13 posiciones que van de –100 hasta +100, y que resulta en el siguiente gráfico:

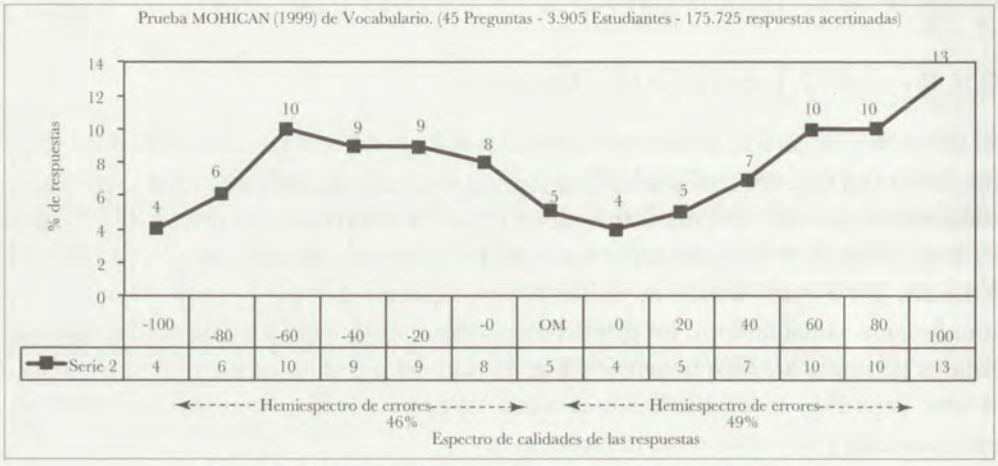


Figura 4: Espectro de calidades de las respuestas acertinadas

Un gráfico de este tipo se puede interpretar analizando:

1. Las omisiones: son 5%, lo que indica que en este test hubo muchos estudiantes que no lograron contestar a todas las preguntas (probablemente por falta de tiempo);
2. Las respuestas correctas: su tasa (aquí 49%) y su distribución, en este caso en *forma de Jota*, que es lo que se espera (con pocas respuestas correctas con dudas y muchas muy seguras).
3. Las respuestas incorrectas: su tasa (46%) y su distribución: *en forma de campana (de Gauss), pero con una zona de pendiente poco pronunciada (casi horizontal) a la derecha* (lo que indica que una mayoría de los errores viene con grados de seguridad bajos, lo que es deseable). La buena noticia es que la tasa de las peores respuestas (incorrectas con 100% de certeza) es baja (4%).
4. Otros índices, como la Confianza promedio, la Imprudencia promedio, la Prudencia, el Matiz, la Lucidez, la tasa de certeza alta entre las respuestas correctas, o "Firmeza" (ver Capítulo 17, sección C).

Este tipo de gráfico puede ser establecido para cada estudiante y/o para cada pregunta.

C.2. Analizar la distribución espectral usando palabras

Es cómodo utilizar palabras al analizar la distribución espectral de las respuestas y sus certezas. La propuesta (arbitraria) que sigue está inspirada en términos que aparecen en la literatura[141]:

[141] De Finetti (1965), Shuford *et al.* (1966), Hunt (1993).

Tabla 8: Posiciones del Espectro de calidades expresadas en estados de los recursos cognitivos

Confusión o mal conocimiento					Ignorancia reconocida			Dominio o buen conocimiento				
Imprudente o ignorada (idea falsa)			Prudente o reconocida		?			Con duda		Confiado		
-100	-80	-60	-40	-20	-0	OM	0	20	40	60	80	100
total	alta, temeraria	baja	alta	máxima				máxima	alta	baja	alta	total
peligrosas			inutilizables							utilizables		
% de Imprudencias peligrosas (Concepciones erróneas)			% de errores prudentes = Prudencia		OM					% de Confianzas Utilizables = Firmeza		
% promedio de certeza = Imprudencia (media)								% promedio de certeza = Confianza o Firmeza (media)				

Estas palabras pueden ser utilizadas para comunicarse entre el docente y los estudiantes, y para tomar decisiones como:

- dar prioridad a la remediación de las confusiones ignoradas (o imprudencias peligrosas), y/o
- calcular los progresos en varias posiciones. Por ejemplo, en las posiciones de buen conocimiento, calcular la evolución de la tasa del conocimiento con alta (80%) o total (100%) confianza.

El mismo tipo de tabla puede ser concebida para las mediciones con pruebas de Verdadero-o-Falso:

Tabla 9: Posiciones del Espectro de calidades expresadas en estados de los recursos cognitivos evaluados en pruebas Verdadero–o–Falso

Confusión o mal conocimiento				?			Dominio o buen conocimiento		
							Confiado		
Imprudente o ignorada (idea falsa)		Prudente o reconocida	Ignorancia reconocida				con duda	confiado	
-100	-80	-60	-50	OM		50	60	80	100
total	alta temeraria	baja					baja	alta	total
peligrosas		inutilizables						utilizables	
% de Imprudencias peligrosas		% de errores prudentes = Prudencia						% de Confianzas Utilizables = Firmeza	
% promedio de certeza = Imprudencia				OM			% promedio de certeza = Confianza		

También se puede calcular:

- La *Confianza Promedio* (o ConfM): el promedio de las certezas acompañando las respuestas correctas.

- La *Imprudencia Promedio* (o ImpM): el promedio de las certezas con las respuestas incorrectas.
- El *Matiz (o Nuancia o Resolución)*, es decir, la diferencia entre confianza e imprudencia (ConfM – ImpM), ausente en las tablas 8 y 9, y que vale 18% (62%-44%) en el ejemplo de la Figura 4.

C.3. Espectros individuales de calidades de las respuestas: ilustración con un TE[142] oral con PVFs[143]

A-M. Rinaldi[144] evaluó a 38 pacientes diabéticos usando 35 preguntas referidas al autocuidado. Como los pacientes debieron ser interrogados en forma oral, se prefirió utilizar PVFs en lugar de PSMs. Durante la evaluación de los conocimientos, la enfermera que realizó la interrogación tenía en su mano el cuestionario y leía las preguntas en voz alta, sin dejar que la persona interrogada viera el formulario. Este también contenía las respuestas correctas (V o F), de modo que la enfermera podía anotar directamente la posición espectral de la repuesta, acompañada de uno de los cuatro grados de certeza entre los que debía elegir el paciente: 50%, 60%, 80%, o 100% de seguridad en su respuesta.

La Figura 5 muestra un ejemplo con dos preguntas de un cuestionario de este tipo (TEM-PVF), donde aparecen pre-indicadas las respuestas correctas (a la derecha) y las incorrectas (a la izquierda) de la pregunta con las letras V o F. Las calidades espectrales de las dos respuestas de un paciente fueron indicadas por la enfermera-interrogadora encerrándolas en círculos gruesos:

Respuestas Incorrectas (RI)							Respuestas Correctas (RC)			
100%	80%	60%	50%	V	P30. En caso de pérdida súbita de conciencia, la gente cerca debe inyectarle inmediatamente de 2 a 4 unidades de insulina.	F	50%	(60%)	80%	100%
100%	80%	60%	(50%)	F	P31. Un esfuerzo físico prolongado puede provocar una hipoglicemia	V	50%	60%	80%	100%
Hemiespectro incorrectas							Hemiespectro correctas			

Figura 5: Ejemplo de un Test Espectral Metacognitivo (ver Capítulo 9) Verdadero-o-Falso para 2 preguntas

Durante la interrogación, sobre la mesa está siempre visible la consigna:

Verdadero – *Falso*	*50%*	*60%*	*80%*	*100%*

Figura 6: Consigna para expresar los grados de seguridad en preguntas Verdadero-o-Falso

[142] Es un Test Espectral (ver Capítulo 9) sin el componente de Metacognición.
[143] Preguntas Verdadero-o-Falso.
[144] Leclercq, Rinaldi y Ernould (2003).

La enfermera encerró en un círculo la posición espectral de cada respuesta. En la pregunta 30 (P30), el paciente contestó "Falso" (la respuesta correcta) con grado de certeza 60%. La enfermera marcó 60% a la derecha. En la P31, el paciente contestó "Falso" con 50% de certeza, y, como no era la respuesta esperada (correcta), la enfermera marcó 50% a la izquierda. A continuación se muestran las posiciones espectrales de las 6 repuestas de un paciente a las seis últimas preguntas de la interrogación.

Respuestas Incorrectas (RI)							Respuestas Correctas (RC)			
100%	80%	60%	50%	V	P30.	F	50%	(60%)	80%	100%
100%	80%	60%	(50%)	F	P31.	V	50%	60%	80%	100%
100%	80%	60%	50%	F	P32.	V	50%	(60%)	80%	100%
100%	80%	60%	50%	F	P33.	V	50%	60%	80%	(100%)
100%	80%	60%	50%	F	P34.	V	50%	60%	(80%)	100%
100%	80%	(60%)	50%	F	P35.	V	50%	60%	80%	100%
Hemiespectro incorrectas							Hemiespectro correctas			
Imprudencia promedio=110/2 = 55%					Matiz promedio=75–55 = 20		Confianza promedio=300/4 = 75%			

Figura 7: Posiciones espectrales de las respuestas de un paciente a las seis últimas preguntas del test

Los contenidos de las P31 y P35 deben ser revisados por este paciente-estudiante, pero también deben serlo los de las P30 y P32, porque la certeza es insuficiente, considerando que el umbral de satisfacción con pacientes diabéticos fue fijado en "Correcta con 80%" (Brutomesso *et al.*, 2003).

Recordemos que el umbral de satisfacción es fijado en un nivel que puede variar dependiendo del ámbito en que se ubiquen los contenidos. Por ejemplo, en +100% con pilotos de avión, medicina de urgencia y transfusiones de sangre, y en +50% con estudiantes en un curso de Introducción a las Ciencias de la Educación (Leclercq, 2009).

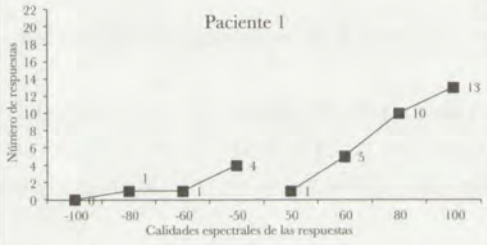
C.4. Interpretar las formas de las distribuciones espectrales individuales

En la Figura 8 aparecen las posiciones espectrales de las respuestas acertinadas de 5 de los 38 pacientes que fueron evaluados por Rinaldi con su TEM V-o-F de 35 preguntas. Abajo, en porcentajes aparecen las respuestas del grupo de 38 pacientes.

RESPUESTAS INCORRECTAS (RI)							RESPUESTAS CORRECTAS (RC)					
100%	80%	60%	50%	NRI	%RI		%RC	NRC	50%	60%	80%	100%
0	1	1	4	6	18%	Paciente 1	82%	29	1	5	10	13
0	3	2	1	6	18%	Paciente 20	82%	29	4	1	3	21
5	0	0	2	7	20%	Paciente 26	80%	28	0	0	0	28
0	0	0	15	15	43%	Paciente 28	57%	20	1	2	4	13
6	0	0	6	12	33%	Paciente 31	67%	23	1	1	1	20
7%	3%	2%	3%		14%	Grupo de 38 pacientes (%)	86%		2%	4%	6%	73%
Hemiespectro de incorrectas							Hemiespectro de correctas					

Figura 8: Distribuciones espectrales de las 35 respuestas de cinco pacientes al Test de Rinaldi

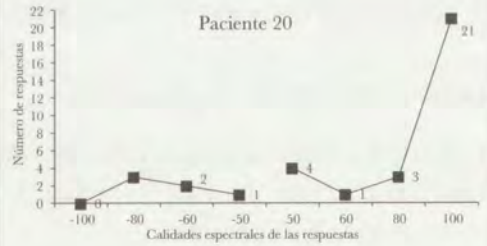
Inmediatamente, al observar el hemiespectro de la izquierda, se detectan los pacientes que presentan problemas de concepciones erradas o ideas falsas (pacientes 26 y 31). A su vez, los pacientes 1 y 28 tienen pocas respuestas correctas con certeza 100%. Para facilitar la lectura de los resultados, especialmente en el caso de evolución de la distribución de una misma persona en dos momentos diferentes en el tiempo, estas distribuciones (con los mismos datos) serán presentadas en forma gráfica[145], utilizando expresiones verbales como distribución "*en campana*", "*en i*", "*en J*", "*en U*" u "*horizontal*". Es interesante cuando dos curvas se superponen, y es posible calcular el grado de asimetría (*skewness*) de cada una de las dos distribuciones. Esto puede hacerse con una fórmula matemática que entrega un valor negativo al índice cuando el Modo (el punto más alto) se ubica a la derecha (curva en J) y positivo cuando esta cima está a la izquierda (curva en i). Aunque el examen visual de la forma de las curvas es en muchos casos suficiente para el análisis por parte de un docente, el índice de asimetría puede ser útil para concebir retroalimentaciones entregadas por computador.



Hemiespectro de errores (izquierdo):
Muy satisfactorio (en forma de Jota). Lo bueno es que no hay ningún error con 100% de certeza.

Hemiespectro de éxitos:
Muy satisfactorio (en forma de Jota).

Diagnóstico: El paciente debe preguntarse "¿Por qué dudé?" en cada una de las 16 respuestas correctas en que indicó un grado de certeza inferior a 100%.



Hemiespectro de errores: forma de campana. Lo bueno es que no hay ningún error con 100% de certeza.

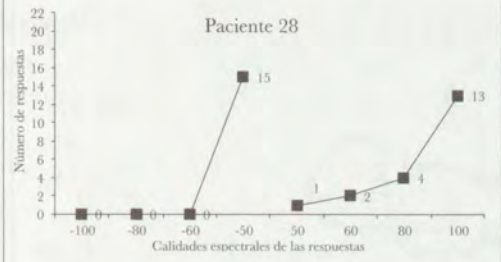
Hemiespectro de éxitos: muy satisfactorio (en forma de Jota); 4 respuestas nada seguras (50%).

Diagnóstico: El uso de la certeza 50% no es apropiado, pues sobre 5 respuestas, 4 (80%) son correctas: *falta de realismo por subestimación.*

Figura 9 parte 1: Análisis de las versiones gráficas de las distribuciones espectrales de las respuestas a prueba sobre la diabetes, de 4 de los 38 pacientes y del grupo total de la prueba (Leclercq, Rinaldi y Ernould, 2003)

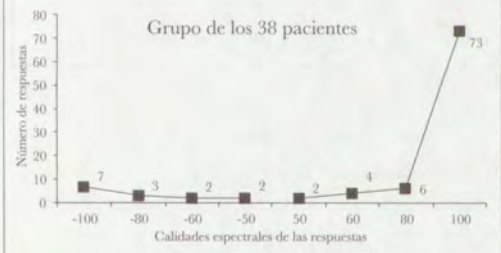
145 Aunque se debe interpretar solo los puntos y no las líneas que los vinculan.

Hemiespectro de errores: muy preocupante: curva en U, con muchos (6) errores con el mayor grado de certeza (100%). Radicaliza sus respuestas: solo 50% (ignorancia total – al azar) o 100% (seguridad total).
Hemiespectro de éxitos: casi perfecto: 20 de 23 (87%) respuestas correctas tienen 100% de certeza.
Diagnóstico: Exceso de utilización de grados de certeza 100%, pues solo 20 sobre 26 son correctas, es decir, una tasa de éxito solamente de 77%: ¡falta de realismo por sobrestimación!



Hemiespectro de errores:
Perfecto: en forma de Jota; todas las respuestas incorrectas tienen el grado de certeza mínimo (50%). Sabe que no sabe
Hemiespectro de éxitos: muy satisfactorio (en forma de Jota).

Diagnóstico: el paciente debe preguntarse "¿Por qué dudé?" en cada una de las 7 respuestas que son correctas, pero con un grado de certeza inferior a 100%.



La tasa de respuestas correctas es 85%.
La Confianza promedio es 95% (¡muy bien!)
La Imprudencia promedio es 77% (¡preocupante!)
La tasa (%) de Confianza utilizable es 86% (73/85), lo que es bueno.
La tasa (%) de Imprudencia peligrosa es 71% (10/14), lo que es preocupante.
Hemiespectro de errores: muy preocupante: el 7% de las respuestas (la mitad de los errores) tienen el grado de certeza máximo (100%).
Hemiespectro de éxitos: muy satisfactorio (porque la curva es en forma de Jota muy escarpada).
Diagnóstico: Uso excesivo del grado de certeza 100%. Urgentemente se debe corregir ese 7% de respuestas muy peligrosas.

Figura 9 parte 2: Análisis de las versiones gráficas de las distribuciones espectrales de las respuestas a prueba sobre la diabetes, de 4 de los 38 pacientes y del grupo total de la prueba (Leclercq, Rinaldi y Ernould, 2003)

Para profundizar en el autodiagnóstico sobre la base de la posición espectral de cada respuesta le sugerimos al lector ver el Capítulo 9 dedicado a los Tests Espectrales Metacognitivos (TEM).

C.5. Efectos de un proceso formativo sobre las curvas

Cuando organizamos un proceso de formación después de un pre-test, esperamos varios tipos de ganancias (mejoras), que se corroboran (o no) según los resultados de un post-test. La Figura 10 presenta una comparación de este tipo, con la superposición de curvas que representan las calidades de las respuestas acertinadas, en PRE y POST tests, en una prueba de 34 preguntas sobre el conocimiento de la "cadena de sobrevivencia" (ver dibujo), aplicada por uno de mis estudiantes (Ndabawarukanye, 2004, p. 29) a 65 profesionales.

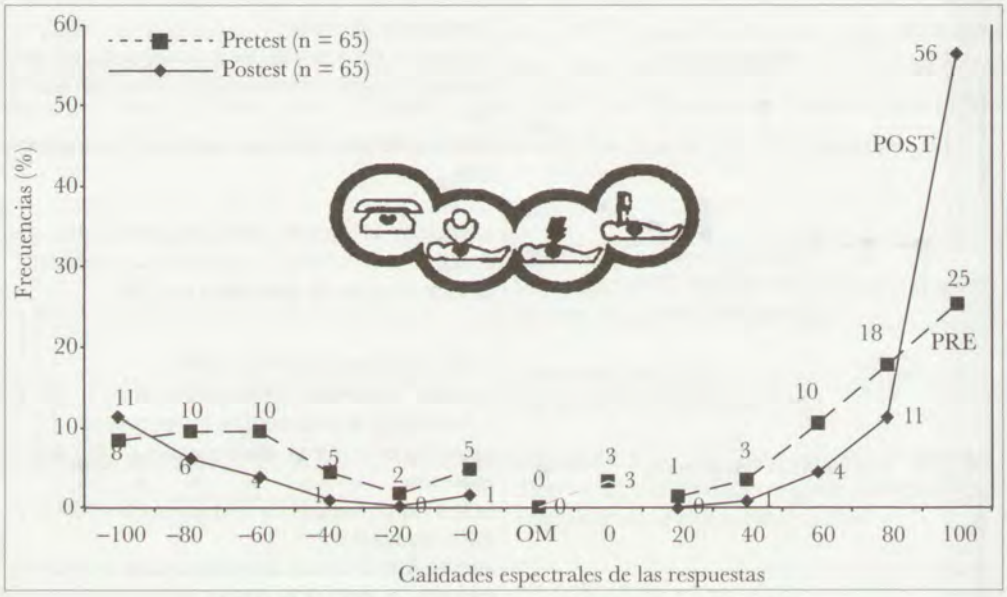


Figura 10: Calidades espectrales de las respuestas acertinadas

Algunas de las expectativas de cualquier proceso formativo son:

(1) Que mejore la tasa de respuestas correctas. En este caso, pasó de 58% en el PRE a 72% en el POST, es decir, una Ganancia Relativa de 33% (14/42).
(2) Que baje la tasa de omisión. Aquí se mantuvo igual (3%).
(3) Que la curva de las respuestas correctas se modifique para adoptar más la forma de jota. Es lo que ocurrió en este caso.
(4) Que la curva de las respuestas incorrectas también adopte una forma de jota más pronunciada. Aquí, lamentablemente, ocurrió a la inversa de las expectativas: adquirió una mayor forma de i. Es una observación frecuente que los procesos de formación aumentan la seguridad (y por ende los grados de certeza), también en algunas respuestas incorrectas, lo que mantiene y aumenta las respuestas peligrosas.

C.6. Impacto diferencial de la formación sobre las dudas y las concepciones erróneas

En Padua (Italia) Brutomesso *et al.* (2003, 2006) evaluaron a 38 pacientes diabéticos usando una prueba que consistió en 39 Preguntas de Selección Múltiple. La evaluación fue hecha antes, inmediatamente después y 3 años después de un programa de educación terapéutica del paciente. Antes, en10% las respuestas fueron erróneas con 100% de certeza, y otro 9% resultó erróneo con una certeza más baja. Inmediatamente después, estos dos porcentajes se transformaron en 7% y 3%, lo que ya indica *un impacto diferente del programa de educación sobre los dos tipos de mal conocimiento.* Tres años después los errores del segundo tipo (con grado de certeza inferior a 100%) se mantuvieron bajos, pero los del primer tipo reaparecieron. ¿Son las concepciones erróneas muy enraizadas el equivalente a "errores resistentes", que si no son erradicados por el tratamiento (el programa de educación) se vuelven inmunes a él?

C.7. El análisis de las distribuciones espectrales de las respuestas para analizar cada pregunta

Los principios de análisis son los mismos para las preguntas que para las personas. La Tabla 10 contiene las distribuciones espectrales de calidad de 10 de las 35 preguntas del test de Rinaldi: de la pregunta 16 a la 29, con las primeras palabras de la pregunta en la columna de derecha.

Tabla 10: Las distribuciones espectrales de las respuestas de los 38 pacientes a 10 de las 35 preguntas del test

	-100	-80	-60	-50	50	60	80	100	NR	INICIO DEL CUADERNILLO DE PREGUNTAS	
16	0	0	0	0	0	1	4	33	38	La insulina (producida...	
17	14	5	1	3	2	1	1	11	38	En caso de hipoglicemia...	
18	0	1	0	4	1	3	3	26	38	En caso de una infección...	
19	0	0	1	0	0	0	1	36	38	Una hiperglicemia...	
20	1	1	0	10	1	0	2	23	38	En caso de pérdida de conciencia	
21	0	0	0	0	0	0	1	37	38	En caso de hipoglicemia...	
22	4	4	1	2	4	3	1	19	38	En presencia de náuseas	
23	0	0	0	0	0	1	0	37	38	La hipoglicemia...	
24	1	0	1	2	0	1	1	32	38	En caso de hipoglicemia...	
25	10	3	0	2	1	2	2	18	38	Si usted pierde totalmente...	
26	1	0	0	0	1	0	1	35	38	En caso de hipoglicemia...	
27	5	1	2	1	1	2	4	22	38	Se habla de hiperglicemia	
28	4	0	3	1	0	1	3	26	38	En presencia de una herida	
29	0	1	1	13	2	1	2	18	38	La presencia de acetona	

Se ve que las preguntas (¿los contenidos?) que revelan un problema a remediar en forma prioritaria son:

- La pregunta 17, donde 14 pacientes sobre 38 (37%) se equivocaron con la certeza máxima (100%).
- La pregunta 25, donde ocurre el mismo problema con 10 pacientes.
- Las preguntas 20 y 29, que presentan declaraciones masivas de ignorancia reconocida (respuestas correctas o incorrectas con 50% de certeza), con 11 y 15 pacientes respectivamente que contestan así.

C.8. Síntoma y diagnóstico

Estos índices son como la temperatura corporal en medicina: indican que hay un problema, pero no constituyen un diagnóstico pues no indican el "por qué".

Las *causas deben ser buscadas clínicamente*, proponiendo, testeando, y descartando o confirmando hipótesis, tales como:

- ¿Incomprensión de algunas palabras en la pregunta?
- ¿Confusión entre conceptos? ¿Concepciones erróneas previas (falso conocimiento)?
- ¿Error de razonamiento?
- ¿Errores o ambigüedades durante el programa de educación o en algún documento de referencia entregado a los pacientes?
- ¿Interpretaciones incorrectas de observaciones cotidianas?

D. JOC: Juicio de comprensión

D.1. Metamemoria

Hasta este momento, en este capítulo no se han tratado otros métodos para fomentar sistemáticamente este aspecto de la metacognición (juicios del estudiante sobre su propio conocimiento y habilidades), aunque existen algunos muy cercanos. En esta sección presentaremos uno de ellos: la metamemoria, teorizada por Nelson y Narens (1990), quienes distinguen el "*monitoring*" (autoobservación) del "*control*" (planificación y remediación) del comportamiento (ver Figura 11). Su modelo integra conceptos tales como:

- EOL (*Ease of Learning*): el sentimiento de facilidad en el aprendizaje de un contenido o una destreza o una competencia.
- JOL (*Judgments of Learning*): la estimación que puede hacer una persona de su probabilidad de tener éxito (o de su nota más probable) en una prueba futura (Dunlovsky y Metcalfe, 2009). Este concepto ha sido utilizado mucho en el estudio de la metamemoria en laboratorio, donde se pide a la persona predecir su probabilidad

de re-evocar ítems algunas horas o algunos días después. En la vida cotidiana de un estudiante, un concepto cercano es la PRE-estimación[146] del nivel de éxito en una prueba futura (al día siguiente, por ejemplo), en un momento en que la persona todavía no conoce las preguntas.

- *FOK (Feeling of Knowing)*: medido cuando a una persona, que ha sido *incapaz de evocar* una respuesta, se le pide estimar su *probabilidad de reconocerla* (entre las opciones de una Pregunta de Selección Múltiple -PSM). Este fenómeno es comparable a la expresión: "lo tengo en la punta de la lengua" (Koriat, 2000).

Propongo el esquema que sigue, inspirado en el de Nelson y Narens (1990) donde he añadido (1) la flecha central con los procesos mentales; (2) las flechas verticales para indicar los momentos iniciales y terminales del aprendizaje (flechas subiendo) y de la evaluación (flechas bajando); (3) Las expresiones Antes (PRE), Durante (PER) y Después (POST), relativas al estudio (abajo) y relativas a la evaluación o acción (arriba); (4) la expresión JOC –*Judgement of Comprehension*– (ver sección D2); (5) las nociones de debate, con comunicación de las respuestas correctas (ver Capítulo 9 con los TEMS) y de automatismo del pensamiento (ver Capítulo 13).

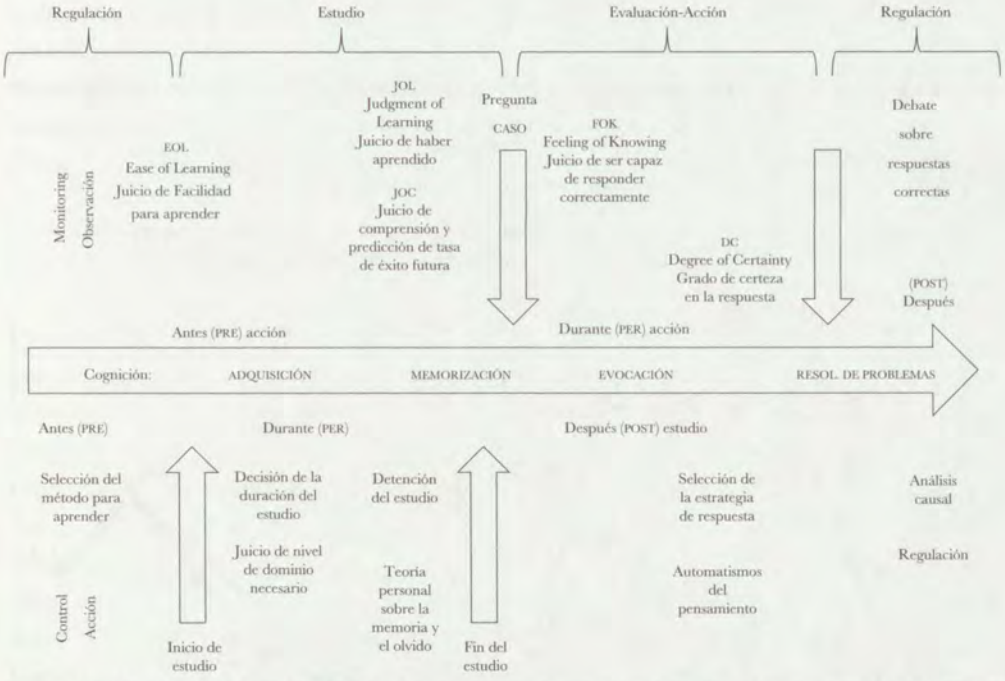


Figura 11: Modelo de *Metacognición* de Leclercq inspirado en el modelo de *metamemoria* de NyN (1990)

[146] Ver ejemplos de momentos metacognitivos PRE, PER y POST una situación de aprendizaje o de evaluación, en el Capítulo 9: TEM – Tests Espectrales Metacognitivos.

D.2. Meta-comprensión y JOC

Denomino *JOC* (*Judgment of Comprehension*), o Juicio de Comprensión, a la estimación que hace una persona sobre la probabilidad de que él o ella proveerá evidencia de la comprensión de un concepto o un mensaje, en caso de ser testeado sobre este contenido (por ejemplo, si será capaz de contestar correctamente a una PSM que trata sobre la definición de un concepto).

Es lo que hice (Leclercq *et al.*, 2002) presentando palabras (extractadas de la prueba de vocabulario de MOHICAN) a 220 estudiantes, con dos consignas sucesivas:

1. La consigna 1 era "Para cada palabra, entregue su *certeza* de que contestará correctamente a una PSM sobre la definición o sobre un sinónimo de esta palabra". Lo importante aquí es que los estudiantes están *SIN* la obligación de dar esta definición o sinónimos (modo *SIN* obligación de dar evidencias).
2. Después que han entregado sus grados de certeza, la consigna 2 era *contestar* a PSMs de comprensión, añadiendo otra vez grados de certeza, pero esta vez ya conocen las preguntas… y la dificultad real de contestarlas (modo *CON* obligación de dar evidencias), al contrario de la primera ocasión donde pudieron ilusionarse acerca de su capacidad de contestar.

La hipótesis era que cuando fueran obligados a dar evidencias (contestar a preguntas), los estudiantes serían más prudentes; es decir, los grados de certeza de los errores se ubicarían más a la derecha del hemiespectro de la izquierda. Esto fue lo que confirmaron los datos obtenidos, que se presentan en la Figura 12.

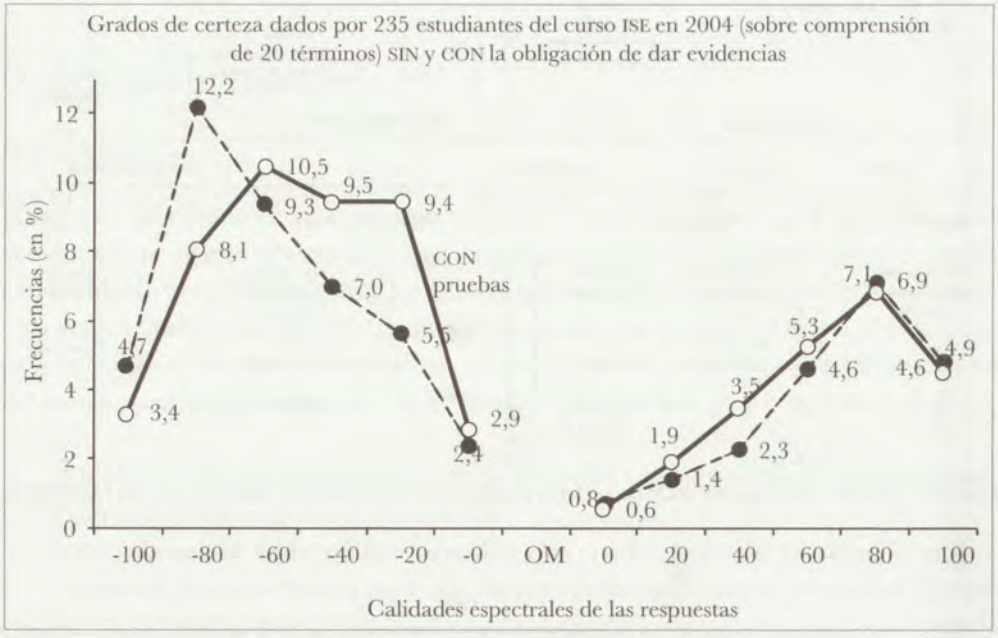


Figura 12: Grados de Certeza SIN y CON obligación de dar pruebas (las omisiones, cerca de 33%, no son representadas)

Este tipo de experimentación en tamaño real (con verdaderos estudiantes) persigue un reto formativo: demostrar a los estudiantes que a menudo, cuando leen un texto (especialmente cursos universitarios), se ilusionan sobre su comprensión de palabras o mensajes y consultan muy poco el diccionario o a expertos, porque cuando leen no están obligados a dar evidencias de que han comprendido.

E. La sobrestimación (*overconfidence*)

E.1. Un fenómeno general

La literatura científica ha producido numerosos artículos que describen la sobrestimación o Exceso de Confianza, especialmente en ámbitos como inversiones financieras (Allen y Evans, 2005), estudiantes de economía (Grimes, 2002), asuntos militares (Johnson, 2004), toma de decisiones (Sieck y Arkes, 2005), testigos en juicios (Allwood y Johansson, 2004), apuestas en casinos (Goodie, 2005), psicofísica (Soll y Klayman, 2004), programación de *spreadsheets* (Takaki, 2006), y deportes (Fogarty y Else, 2005). Existen varias explicaciones de este fenómeno (por ejemplo, la propuesta por Kruger y Dunning, 1999).

También he observado esta tendencia a nivel de estudiantes de primer año de pregrado: en promedio, la certeza media (CM) en un test es superior a la tasa de éxito (TE), lo que provoca un Error de Centración (EC), calculado por la fórmula EC = CM − TE.

Si EC < 0, hay subestimación; si EC > 0, hay sobrestimación.

E.2. Las variaciones interindividuales en la sobrestimación

En un proyecto llamado MOHICAN, cerca de 4.000 estudiantes que se encontraban ingresando a los primeros años de pregrado de las universidades francófonas de Bélgica, contestaron a 10 pruebas en ámbitos diversos: (1) vocabulario, (2) sintaxis, (3) comprensión de textos, (4) comprensión de gráficos en un ámbito de geografía, (5) matemáticas, (6) física, (7) química, (8) biología, (9) artes, y (10) historia-actualidad-economía. Se utilizaron los grados de certeza. En todas las pruebas, el EC medio (Error de Centración promedio del grupo) fue un valor positivo (> 0), indicando sobrestimación, pero con grandes variaciones entre los estudiantes. La Figura 13 presenta la distribución de los EC de los estudiantes en la prueba de Vocabulario, que contenía 45 preguntas. Fue contestada por 3.801 estudiantes, con un EC (Error de Centración) promedio de +5,5% (más a la derecha que EC = 0, representado por la línea vertical punteada), pero la distribución es amplia (la desviación estándar es 15).

En todas las 10 pruebas de MOHICAN las mujeres se sobrestiman menos que los hombres. Una observación que ha sido replicada por Figari *et al.* (2006, p. 327-331), sobre 770 estudiantes de 5to a 8vo grado en Ottawa.

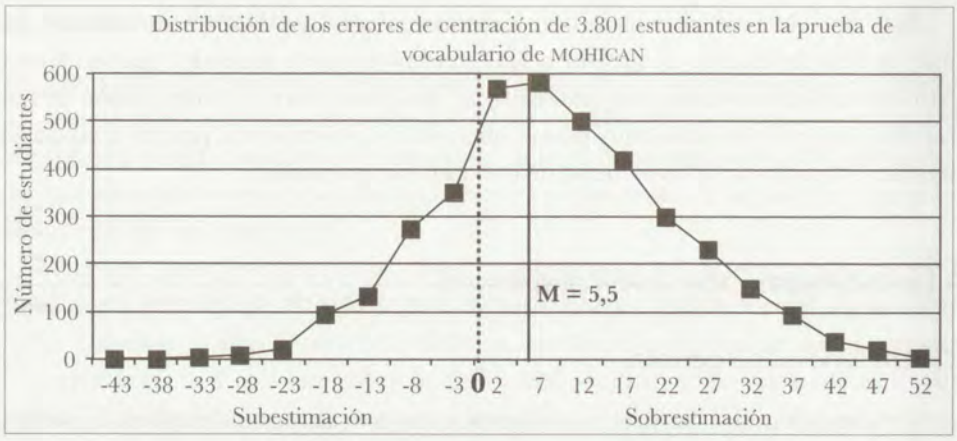


Figura 13: Distribución de los Errores de Centración en la prueba "Vocabulario" (3.801 estudiantes) de MOHICAN

F. Conclusiones

Los grados de certeza constituyen un *microscopio* para observar los procesos cognitivos y para estimular los procesos metacognitivos (ver capítulo 9). Permiten medir modificaciones (como las ilustradas en la sección D2) que son invisibles al *nivel macroscópico* o solo con las respuestas sin grados de certeza. En consecuencia, muchas investigaciones que llegaron a la conclusión "No hemos observado una diferencia (estadísticamente) significativa" pueden haber sido ciegas a modificaciones reales, porque no utilizaron un instrumento de observación suficientemente preciso.

Este capítulo ha querido demostrar la fecundidad de la validez teórica y de la validez informativa o diagnóstica de esta técnica (aspecto cualitativo), al servicio de la formación (ver Capítulo 9) y de la investigación. Los aspectos cuantitativos de docimología (cómo asignar puntajes, cómo calificar, cómo concebir listas de cotejo) son tratados en el siguiente capítulo (n° 17).

Referencias

Allen, W. y Evans, D. (2005). Bidding and Overconfidence in Experimental Financial Markets. Journal-of-Behavioral-Finance. Vol 6(3) 2005, 108-120.

Allwood, C.M. y Johansson, M. (2004). Actor-Observer differences in realism in confidence and frequency judgments. Acta-Psychologica. Vol 117(3) Nov 2004, 251-274.

Angoff, W. (1971). Scales, norms and equivalent scores. In R. Thorndike (Ed). Educational Measurement. Washington: American Council of Education, 2nd ed., 514-515.

Brutomesso, D., Gagnayre, R., Leclercq, D., Crazzola, D., Busata, E., D'Ivernois, J.F., Casiglia, E., Tiengo, A. y Baritussio, A. (2003). The use of degrees of certainty to evaluate knowledge. Patient Education and Counseling, 51, 29-37.

BRUTTOMESSO, D., COSTA, S., DAL, M., CRAZZOLA, D., REALDI, G., TIENGO, A., BARITUSSIO, A. Y GAGNAYRE, R. (2006). Educating diabetic patients about insulin use: changes over time in certainty and correctness of knowledge. Diabetes Metabolism, 32, 256-261.

BURATTI, S. Y ALLWOOD, C. M. (2012). The accuracy of meta-metacognitive judgments — Regulating the realism of confidence. Cognitive Processing, 13, 243-253. http://dx.doi.org/10.1007/s10339-012-0440-5 (consultado el 6 de enero de 2014).

COOKE, W.E. (1906). Forecasts and verifications in Western Australia. Monthly Weather Review, 34, 23-24.

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.

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.

DE LANDSHEERE, V. (1988). Faire réussir – Faire échouer. La compétence minimale et son évaluation. Paris: Presses Universitaires de France.

DE WAELE, J-P. (1975). La méthode des cas programmés en criminologie. Bruxelles: Dessart.

DUNLOVSKY, J. y METCALFE, J. (2009). Metacognition. Sage publications.

FABRE, J.M. (1993). Subjective Uncertainly and the Structure of the Set of all Possible Events. In D. Leclercq D. y J. Bruno J. (1993), Item Banking: Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin: Springer Verlag, 99-113.

FIGARI, G., RODRIGUES, P., ALVES M.-P. Y VALOIS, P. (Eds.) (2006). Evaluation des compétences et apprentissages expérientiels. Lisbonne: ADMEE Europe et Educa.

FOGARTY, G. Y ELSE, D. (2005). Performance Calibration in sport: Implications for Self-Confidence and Metacognitive Biases. International-Journal-of-Sport-and-Exercise-Psychology. Vol 3(1) Mar 2005, 41-57.

GOODIE, A. (2005). The Role of Perceived Control and Overconfidence in Pathological Gambling. Journal-of-Gambling-Studies. Vol 21(4) Dec 2005, 481-502.

GRIMES, P. (2002). The Overconfident Principles of Economics Student: An Examination of a Metacognitive Skill. Journal of Economic Education. v33 n1 p15-30 Win 2002.

HUNT, D. (1993). Human Self-Assessment: Theory and Application to Learning and Testing. In: Leclercq D. y Bruno J. (1993), Item Banking: Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin: Springer Verlag, 177-189.

JOHNSON, D. (2004). Overconfidence and war: The havoc and glory of positive illusions. Cambridge, MA, Harvard University Press 280 pp.

KORIAT, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. Consciousness and Cognition, 9, 149-171.

KRUGER, J. Y DUNNING, D. (1999). "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments". Journal of Personality and Social Psychology 77 (6): 1121-34.

LECLERCQ, D. (1975). L'évaluation subjective de la probabilité d'exactitude des réponses en situation pédagogique. Thèse de doctorat en Sciences de l'Education, Université de Liège.

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. y Boskin, A. (1990). Note taking behaviour studied with the help of hypermedia. In Estes, Heene y Leclercq (Eds), Proceedings of the 7th International Conference on Technology and Education, Bruxelles. Edimburgh: CEP Consultants, 2, 16-19.

LECLERCQ, D. Y VANDEN BRANDE, L. (1997). Une méthode pour la formation universitaire clinique en criminologie: les cas programmés, in E. Boxus, V. Jans, J.L. Gilles y D. Leclercq, Stratégies et médias pédagogiques

pour l'apprentissage et l'évaluation dans l'enseignement supérieur. Actes du 15e colloque de l'Association Internationale de Pédagogie Universitaire (AIPU), Liège : STE-Affaires Académiques, 635-644.

Leclercq, D., Georges, F., Gilles, J.-L., Reggers, Th., Rommes, O. (1998). Interactive Multimedia Programmed Biographies (IMPB): a new method for clinical training, Proceedings of at the BITE (Bringing Information Technology for Education) Conference, Maastricht, 25-27 March 1998, 406-417.

Leclercq, D., Simon, F., Marotte, P., Verschueren, A. y Lacaille, C. (2002). Former des étudiants de première candidature universitaire à des compétences transversales: Lesquelles et comment? Deuxième Congrès des chercheurs en Education de la CFWB, Louvain-La-Neuve.

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. y Poumay, M. (2003). La connaissance partielle chez l'apprenant: pourquoi et comment la mesurer, in Gagnayre *et al.* (Eds), L'évaluation de l'Education Thérapeutique du Patient, Paris: IPCEM, 27-30.

Leclercq, D., Rinaldi, A.M. y Ernould, C. (2003). Un questionnaire spectral pour l'évaluation des connaissances chez le patient diabétique. In Gagnayre *et al.* (Eds), L'évaluation de l'Education Thérapeutique du Patient, Paris: IPCEM, 31-34.

Leclercq, D. y Micheels, J. (2010). Degrés de certitude et qualités spectrales des réponses. Applications à des formations en urgence médicale. Soumis à la revue Pédagogie Médicale.

Leclercq, D. (2009). La connaissance partielle chez le patient: pourquoi et comment la mesurer. Revue d'Education Thérapeutique du Patient. 1 (2) pp. 201-212.

Lucas, C. (2001). Evaluation des connaissances d'enfants en matière d'urgence avant et après le visionnement d'une vidéo d'information. Mémoire de licence en Sciences de la santé publique. Université de Liège.

Ndabawarukanye, C. (2004). Evaluation des connaissances sur la "chaîne de survie" dans un établissement de prise en charge de pathologies neurologiques. Master tesis en « Santé publique ». Université de Liège.

Nelson, T.O. y Narens, L. (1990). Metamemory: a theoretical framework and new findings. The psychology of learning and motivation, vol 26, 125-173.

Reach, G., Zerrouki A., Leclercq , D. y d'Ivernois, JF. (2005). Adjusting insulin doses: from knowledge to decision. Patient Education and Counseling, 56, 98-103.

Sanderson, P. (1973). The "don't know" option in MCP examinations. Brit. Jour. of Medical Education, 7, 25-29.

Shuford, E., Albert, A. y Massengil, N.E. (1966). Admissible probability measurement procedures, Psychometrika, 31(2): 125-145.

Sieck, W. y Arkes, H. (2005). The Recalcitrance of Overconfidence and its Contribution to Decision Aid Neglect. Journal-of-Behavioral-Decision-Making. Vol 18(1) Jan 2005, 29-53.

Soll, J. y Klayman, J. (2004). Overconfidence in Interval Estimates. Journal of Experimental Psychology Learning Memory and Cognition. v30 n2 p299-314, Mar 2004.

Takaki, S. (2006). Self-efficacy, confidence, and overconfidence as contributing factors to spreadsheet development errors. Dissertation-Abstracts-International-Section-A:-Humanities-and-Social-Sciences. Vol 66(8-A).

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