# Capítulo XVII

# Grados de certeza y docimología: cómo calificar

Dieudonné Leclercq

A pesar de los numerosos errores metodológicos cometidos en este ámbito durante decenas, continúo creyendo en el postulado de Bruno De Finetti (1965, p. 111):

> *Solo la probabilidad subjetiva puede dar una significación objetiva a cada respuesta y a cada método de puntuación.*

## Preámbulo terminológico

En el presente capítulo varios conceptos, cercanos en sentido, tienen denominaciones que empiezan con la letra *p*, y varios otros que lo hacen con la letra *n*. Es por esto que se hace necesario explicar los diferentes significados.

*Las p:*

*p* = la probabilidad (o el riesgo) de equivocarse al tomar la decisión de rechazar la presunción de realismo (pues no debía ser rechazada). Esta probabilidad se escribe $p<0,01$ cuando el riesgo es inferior a 1 sobre 100. Se escribe $p<0,05$ cuando el riesgo es inferior a 5 sobre 100.

*ps* = la probabilidad subjetiva (de tener éxito en una evaluación, por ejemplo) expresada en valores entre 0 y 1. Es el mismo concepto de los Grados de Certeza, que a menudo se expresan en porcentajes (por ejemplo de 0% a 100%). Recomiendo que se escriba *GC%* cuando el grado de certeza se expresa en porcentajes y *ps* cuando se expresa en probabilidades. Con frecuencia las fórmulas matemáticas exigen que esté expresada en probabilidades (entre 0 y 1).

*po* = la proporción observada de respuestas correctas, expresada en valores de 0 a 1. Es el mismo concepto de la Tasa de Éxito (TE), que se expresa en porcentajes pues es el porcentaje de Respuestas Correctas (escrito *%RC*).

*pB* = la probabilidad binomial (calculada por la ley binomial) de que un evento ocurra. En nuestro caso el evento es *nc*, es decir, el número de respuestas correctas cuando se conocen los valores de *nu* (número de usos) y *ps* (la probabilidad subjetiva anunciada o el grado de certeza).

*psM* = llamado con mayor frecuencia *CM*, es la probabilidad subjetiva (o Certeza Media) en un evento de evaluación. Este índice es el promedio de todos los grados

de certeza dados por una persona, o por todas las personas de un grupo, en una prueba o momento de evaluación específico.

*Las n*

*n* = el valor por el que se multiplica el Error Estándar de Medición (*EEM*) de un porcentaje, o de una proporción observada (*po*), o de una taza de éxito (*TE*), para obtener los límites de confianza de esta *po*. Cuando *nu* > 30, *n* vale 1,96 a p<0,05 y *n* vale 2,58 a p<0,01.

*ns* = el número de sujetos (estudiantes, pacientes, profesionales) que han rendido una prueba.

*np* (o NP) = el número de preguntas en una prueba o test o, con mayor precisión, el número de respuestas esperadas. Por ejemplo, si una viñeta clínica es seguida por 5 preguntas Verdadero-Falso, *np* es 5, y no 1.

*nr* = el número de respuestas de una persona en una prueba o evaluación (máximo = *np*).

*nc* = el número de respuestas correctas.

*ni* = el número de respuestas incorrectas.

*nu* = el número de usos de un grado de certeza determinado (por ejemplo 40% u 80%). Se debe especificar que se trata de $nu_{40}$ (en cuántas preguntas respondió usando 40%), $nu_{80}$ (en cuantas usó 80%).

Los resultados descritos en este capítulo fueron obtenidos mediante el uso de varias escalas para expresar el grado de certeza:

| 0% - 25% | 25% - 50% | 50% - 70% | 70% - 85% | 85% - 95% | 95% - 100% |
|---|---|---|---|---|---|
| | 2% | 10% 25% | 50% | 75% | 90% 98% |
| | 0% | 20% 40% | 60% | 80% | 100% |

Hoy, la consigna que recomiendo es:

*5% 20% 40% 60% 80% 95%*

## A. El docente-investigador en el dilema del egiptólogo: Encontrar la luz atravesando el interior de una pirámide

La pregunta principal que motiva este capítulo es:

¿Cómo incorporar, en la calificación (asignación de puntaje), la información que entregan los grados de certeza sobre el dominio subjetivo del conocimiento que ha logrado el estudiante?

Comparamos la historia de la respuesta a esta pregunta con la exploración del interior de una pirámide, a la luz de una linterna y con el peligro de caer en fosos sin salida. Como en este tipo de exploración,

*en ciencias, cerrar puertas es tan importante como abrirlas.*

Para iluminar el camino de otros docentes-investigadores es que presento el laberinto que he descubierto, durante mis deambulaciones de 43 años por el interior de mi propia pirámide. He caído en casi todas las trampas, y me he quedado en algunas durante décadas. Sin embargo, a diferencia de muchos compañeros investigadores y usuarios, he logrado salir de ellas y seguir explorando. La suerte me ha ayudado mucho. La teoría también, porque, como decía Alexander Fleming,

*La mente no preparada no capta la mano ofrecida por la suerte.*

A continuación presentaré el camino que me parece, hoy en día, el mejor para alcanzar una posición desde la cual se pueden ver los rayos del "sol de la calidad", y que las otras fórmulas no alcanzaron. Estos rayos de sol son las dimensiones de validez teórica, informativa, consecuencial, deontológica y de aceptabilidad.


Figura 1: Los laberintos históricos de las investigaciones y las prácticas de calificación de las performances incluyendo la información entregada por los grados de certeza

Los callejones sin salida (1, 3, 3a, 3b, 4), que terminan en fosos que implican la muerte del uso de los grados de certeza, serán abordados a partir de la sección B de este capítulo. Al final, en la sección H, se encontrará una explicación muy técnica del último pasadizo vertical, que sube hasta alcanzar la verificación del realismo (6), desde donde se ve la luz. En la sección F se presenta la fórmula de cotejo más actualizada del autor, que se basa en esta verificación del realismo. Antes de partir el recorrido, haremos un resumen de los corredores y las trampas, que esperamos ayude al lector a situarse.

1. (Foso). *Definir los grados de certeza en palabras*, como "poco seguro", "seguro", "muy seguro", "totalmente seguro", etc., es un foso. En el Capítulo 16, sección B1, hemos visto por qué se trata de un callejón sin salida.
2. Solo la expresión de certeza en probabilidades (o porcentajes), es decir, en unidades numerales y métricas, permite progresar en el uso de los grados de certeza... pero es solo un umbral que debemos trasponer si queremos continuar nuestro avance.
3. (Foso). *Asignar un puntaje a cada respuesta* es el principio de cotejo más evidente, y consiste en asignar puntaje combinando la calidad (correcta o incorrecta) de la respuesta y el grado de certeza asociado. Esta forma de calificar ha dominado el campo de la pedagogía durante medio siglo, y ha producido variantes.

   La variante 3a consiste en el uso de pautas "intuitivas" (ver sección C.1).
   La variante 3b utiliza pautas conformes a la teoría de las decisiones (ver sección C.2).

   Lamentablemente, este camino no logró eliminar el que los estudiantes utilicen estrategias de respuesta que consisten en mentir cuando declaran su grados de certeza, eligiendo formas de contestar que tratan de maximizar su calificación final antes que de ser realistas en la estimación de dominio subjetivo del conocimiento.
4. Calificar las performances en una evaluación *según el dominio subjetivo* del estudiante, utilizando índices globales de la prueba, es un camino que recorrí durante muchos años, y que parecía fecundo en su búsqueda de la luz (pero he descubierto que, sin la verificación del realismo, también conduce a un foso). Consiste en:

   - medir el *dominio objetivo* a través de la *tasa de éxito*, que produce la *nota clásica*. Esto se puede calcular usando diferentes pautas. Por ejemplo, +1 punto por cada respuesta correcta y –0,25 puntos por cada error.
   - medir el nivel de *confianza* media en las respuestas correctas
   - medir el nivel de *prudencia* media en las respuestas incorrectas.
   - añadir a la nota clásica (al dominio objetivo) puntos extra según los resultados de estas dos mediciones, que combinadas constituyen el *dominio subjetivo*.

Se destaca que este modo de atribuir calificaciones tiene en cuenta los grados de certeza:

- a nivel global (de la prueba entera) y no por cada respuesta
- basándose en mediciones globales o índices de dominio subjetivo, como la Confianza y la (Im)prudencia

5. Los *puntos de la pauta de cotejo* se pueden concebir según dos lógicas. En la fórmula *5a* se trata de asignar *puntos positivos asociados a la Confianza en las respuestas correctas*, y *puntos negativos asociados a la Confianza en las respuestas incorrectas*. Por este segundo hecho es que los estudiantes pueden odiar el uso de grados de certeza, especialmente cuando la calificación final es más baja que la nota clásica calculada con pautas tradicionales. La fórmula 5b, como lo veremos en la sección E3, consiste en dar solo puntos positivos en forma de *bonus (de dominio subjetivo)*, en caso de alta Confianza (con las respuestas correctas) y alta Prudencia, o baja Imprudencia (con las respuestas incorrectas).
6. *El realismo* debe ser tenido en cuenta en el sexto nivel. Se debe verificar si las tasas de éxito con cada grado de certeza respetan la presunción de realismo (ver sección F). Si, por los datos observados, esta presunción de realismo es rechazada en un solo grado de certeza, *el estudiante pierde*, para este test en particular, el derecho a *bonus de dominio subjetivo.*

La investigación sobre los grados de certeza tuvo su periodo de mayor actividad en las décadas de 1960 y 1970, pero se mantuvo en los niveles más bajos del laberinto de la pirámide. Al ser los resultados obtenidos poco interpretables, esta línea y enfoque de investigación acabó por detenerse. Koehler (1974, p. 302) concluye así su estudio experimental acerca de la validez de varias fórmulas de cotejo[147] para asignar puntajes a la certeza declarada por los estudiantes, aplicadas a una consigna que ofrecía al estudiante 10 grados diferentes de certeza sobre el eje de las probabilidades:

> Elijo los tests sin grados de certeza, porque son mas fáciles de administrar, necesitan menos tiempo y los estudiantes no necesitan ser entrenados en ellos... (y porque, con los grados de certeza, no hemos encontrado) ninguna mejora de la validez de constructo... Sin embargo, los grados de certeza pueden ser útiles para otros desafíos, por ejemplo la técnica tutorial.

Fue una buena decisión terminar con estas investigaciones, pues la forma en que los grados de certeza fueron implementados y calificados presentaba varios errores metodológicos fatales, que conducían a fosos sin escapatoria. En las siguientes secciones se presenta cada uno de estos caminos; algunos sin salida; otros, con salidas condicionadas al hecho de adentrarse en caminos subsecuentes. Si el lector tiene prisa, le aconsejamos ir directamente a los caminos 5, 5b y 6 (secciones D, E y F).

[147] Cuadráticas, logarítmicas.

## B. Los caminos 1 y 2: el modo de expresar las certezas

El primer error metodológico cometido en el uso de los grados de certeza ya fue visto en el Capítulo 16, sección B1: utilizar una consigna con palabras en expresiones como "poco seguro", "muy seguro", etc. (pues las personas asignan diferentes significados a las mismas palabras) o, lo que es lo mismo, usar números en órdenes (0, 1, 2, 3) en lugar de porcentajes o probabilidades. Estos errores ya habían sido denunciados por Leclercq (1975). El uso de porcentajes o probabilidades es el primer paso para un uso fecundo de los grados de certeza, pero aún queda mucho camino por recorrer (y errores que evitar).

## C. El camino 3: usar escalas de puntajes (pautas de cotejo) para cada respuesta (Foso)

En esta fase de las investigaciones se cometió el segundo error metodológico: utilizar listas de cotejo que atribuyen puntos a cada respuesta. Varios modos de hacer esto se describen a continuación, pero ninguno logró impedir que *el estudiante mienta* cuando expresa su grado de certeza, con el objetivo de maximizar su calificación final en el total de la prueba. Es decir, la certeza expresada no es la estimada, debido a un interés por mejorar la nota antes que por ser realista.

### C.1. El camino 3a: las pautas intuitivas

La Tabla 1 muestra un ejemplo de estas escalas (baremos) de cotejo que favorecen la mentira:

Tabla 1: Consigna 1 (inapropiada) con grados numéricos y puntajes intuitivos

| CÓDIGO | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Significación | 0–25% | 25–50% | 50–75% | 75%–100% |
| TC (en caso de respuesta correcta) | 0 | +1 | +2 | +3 |
| TI (en caso de respuesta incorrecta) | 0 | –1 | –2 | –3 |

Esta escala es atractiva porque es *fácil de memorizar y aplicar*, siendo 3 el grado máximo de certeza y a la vez el máximo puntaje que se puede ganar o perder en cada pregunta. Sin embargo, cuando el estudiante piensa: "estoy 30% o 40% seguro de mi respuesta", su interés (maximizar su calificación) le dicta no dar el grado 1 (25-50%) –como la consigna le recomienda–, sino que grado 0 (para no afectar su puntaje, ya que su respuesta es poco segura, y puede ser errónea, lo que le restaría un punto). Del mismo modo, cuando piensa: "estoy 60% o 70% seguro", su interés le dicta no dar el grado 2

(50-75%) –como la consigna le recomienda–, sino que el grado 3, buscando una bonificación alta antes que entregar una estimación fidedigna de su certeza. Este comportamiento de los estudiantes puede ser demostrado aplicando la teoría de las decisiones, que se explica en la sección C2.

Jacobs (1971) utilizó la siguiente consigna con dos grupos (A y B), cada uno de 36 estudiantes:

Tabla 2: Consigna 2 de Jacobs (1971) con grados verbales y puntajes intuitivos

| Significación | Guess | Fairly confident | Very confident | Grupo |
|---|---|---|---|---|
| TC (en caso de respuesta correcta) | +1 | +2 | +3 | |
| TI (en caso de respuesta incorrecta) bajas penalidades | 0 | –2 | –3 | A |
| TI (en caso de respuesta incorrecta) altas penalidades | 0 | –4 | –6 | B |

Jacobs observó que la introducción de penalidades bajas no cambió significativamente la fiabilidad de las calificaciones finales en la prueba, pero las penalidades altas la hicieron bajar considerablemente. Concluye que "las penalidades altas sirvieron para aumentar las incoherencias", derivadas del uso de grados de certeza, el que está "contaminado por diferencias individuales de personalidad... (y que) necesita más tiempo y esfuerzo que las consignas convencionales".

Este es un ejemplo del uso de consignas inapropiadas y de incomprensión de los mecanismos de respuesta (de decisiones) de los estudiantes, que pese a ser erróneas tuvieron el poder de detener el uso de los grados de certeza, debido a observaciones como las de Koehler (1974) o (las mismas) de Michael (1968).

## C.2. Usar escalas de cotejo calculadas según la teoría de las decisiones

Las consignas y escalas de cotejo presentadas arriba promueven que los estudiantes escojan los grados de certeza en los extremos del rango (0 y 3), lo que puede verse en las curvas de las Notas Esperadas Estadísticamente (NEE), en la Figura 2. En este concepto (NEE) se basa la teoría de las decisiones, y su aplicación en nuevas consignas y baremos para la calificación de los grados de certeza que fueron desarrollados en una segunda fase de experimentación.

La fórmula de cálculo de la Nota Esperada Estadísticamente es:

$$NEE = (TC \,.\, ps) + (TI \,.\, qs)$$

*ps* es la probabilidad subjetiva de éxito y
*qs* es 1-ps (la probabilidad subjetiva de fracaso)
*TC* y *TI* son las tarifas (los puntajes) en caso de respuesta correcta (TC) e incorrecta (TI)

Para que los estudiantes tengan interés en dar, de manera fidedigna, el Grado de Certeza que corresponde a su *ps* (íntima convicción), y sin mentir eligiendo otro grado –motivados por maximizar su calificación–, las tarifas de puntaje (escalas de cotejo) deben ser calculadas respetando la teoría de las decisiones. Esto se ilustra en la consigna de la Tabla 3 y la Figura 3.

Tabla 3: Consigna 3 (siendo 5 el máximo por pregunta)

| CÓDIGO | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Significación | 0–25% | 25–50% | 50–75% | 75%–100% |
| TC (en caso de respuesta correcta) | 0 | +3 | +4 | +5 |
| TI (en caso de respuesta incorrecta) | 0 | –1 | –2 | –5 |

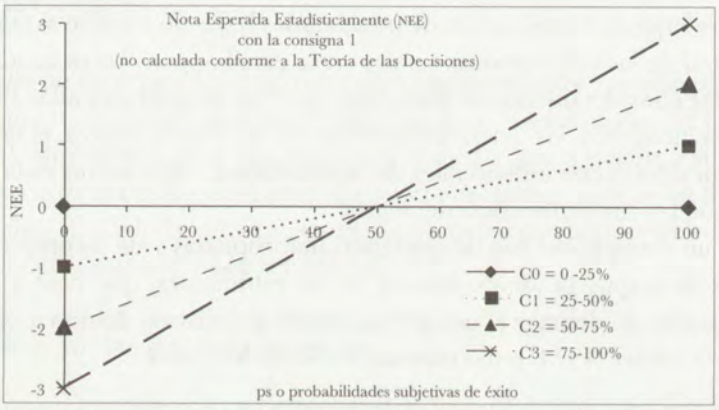


Figura 2: Las NEE con la consigna 1

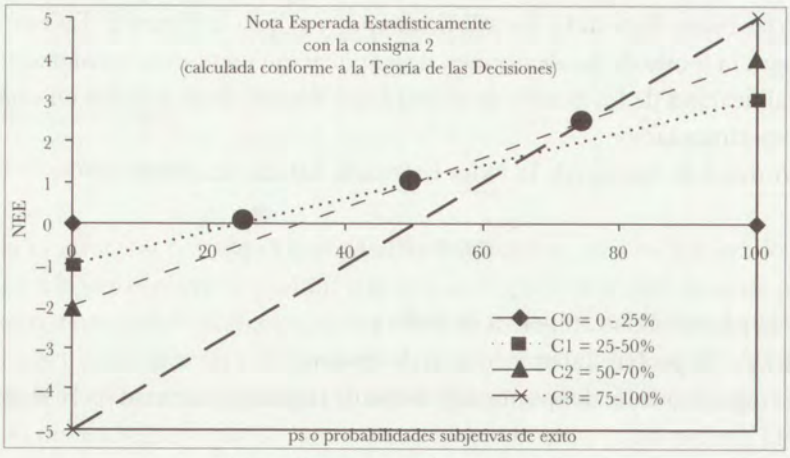


Figura 3: Las NEE con la consigna 3

En la Figura 2 se ve que, con la consigna 1, las notas esperadas (NEE) de los grados de certeza 1 y 2 nunca superan las de los grados 0 y 3. En otras palabras, para maximizar su calificación al estudiante le "conviene":

- indicar grado de certeza 0 cuando está menos de un 50% seguro de su respuesta
- indicar grado de certeza 3 cuando está más de un 50% seguro de su respuesta

Con la consigna 3 (Figura 3) cada grado de certeza es "óptimo": su curva de Notas Esperadas Estadísticamente (NEE) supera a todas las otras curvas en la zona anunciada en la consigna. De este modo el estudiante estaría interesado en decir la verdad, es decir, expresar el grado de certeza que corresponde a su probabilidad subjetiva (*ps*). El mismo principio se puede aplicar a una nueva consigna (n° 4), donde el puntaje máximo por cada pregunta es 20 puntos.

Tabla 4: Consigna 4

| Código | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Significación | 0% | 20% | 40% | 60% | 80% | 100% |
| TC (en caso de respuesta correcta) | 9 | 12 | 14 | 16 | 18 | 20 |
| TI (en caso de respuesta incorrecta) | 2,5 | 2 | 1 | -2 | -6 | -20 |

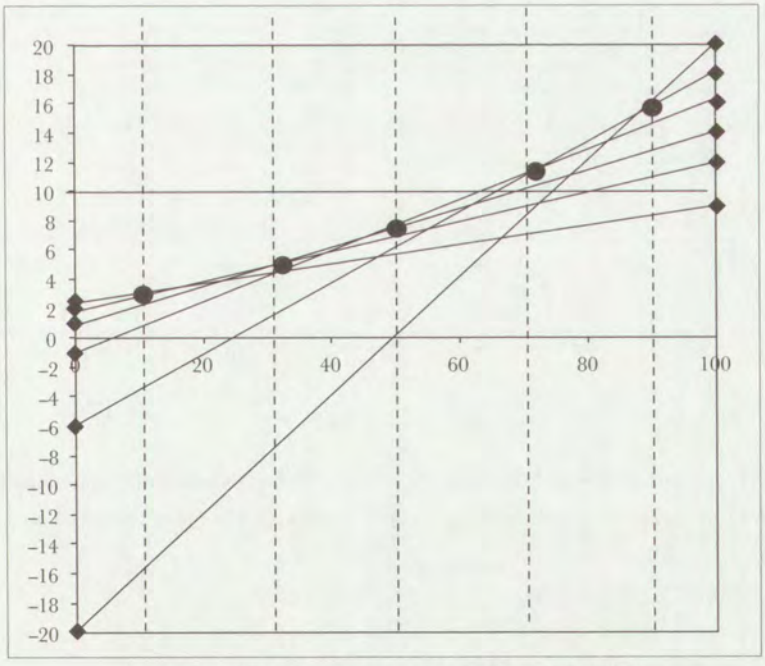


Figura 4: Gráfico de NEES según la consigna 4 (Tabla 4)

Puede sorprender, en el gráfico resultante de la consigna 4 (Figura 4), el hecho de ver puntajes positivos en caso de respuesta incorrecta (si un estudiante contesta de forma incorrecta con 0% de seguridad, obtiene 2,5 puntos en lugar de 0). Sin embargo, se debe recordar que 2,5 puntos, sobre 20 posibles, es un gran fracaso. En la Figura 4 se ve que la línea de NEEs de cada grado de certeza supera a las otras en la zona anunciada en la consigna. Por ejemplo, la línea de las NEEs del grado de certeza 60% es la más alta en la zona que va de 50% hasta 70% (siendo estos dos extremos indicados por círculos).

Con un baremo o escala de cotejo de este tipo, el máximo posible no solo consiste en dar todas las respuestas correctas, sino que además darlas todas con el grado de certeza máximo (100%). Lo anterior resulta en una exigencia extra de este tipo de evaluación, en comparación con las exigencias de las evaluaciones habituales. Es por esta razón que durante años hemos considerado como puntaje máximo distintos valores, según la *severidad* apropiada a cada contenido. Así, el máximo puede ser 16 puntos (todas las respuestas correctas, y con 60% de certeza), 18 puntos (todas correctas con 80%), o 20 puntos.

## C.3. Los seres humanos no se conforman con la teoría de las decisiones

Hubo investigadores que se contentaron con los baremos para calificar basados en la teoría de las decisiones. Sin embargo… ¡Ay! Los seres humanos no nos conformamos con esta teoría. Esto ha sido demostrado experimentalmente por Kahneman y Tversky (1979): los seres humanos prefieren lo que es cierto a lo que es probable, aunque lo probable les dé una NEE más alta que lo cierto.

Durante 20 años (de 1980 a 2000), en la Universidad de Liège, he utilizado (y varios colegas continúan haciéndolo) una lista de cotejo próxima a la consigna 4: la consigna 5 con escalones asimétricos.

Tabla 5: Consigna 5 (conforme a la teoría de las decisiones)

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| | 0–25% | 25–50% | 50–70% | 70–85% | 85–95% | 95–100% |
| TC | 13 | 16 | 17 | 18 | 19 | 20 |
| TI | 4 | 3 | 2 | 0 | –6 | –20 |

En este periodo se ha observado que los estudiantes *continúan utilizando estrategias que no son coherentes con la Teoría de las decisiones*, tanto con esta consigna como con la consigna 4 (Tabla 4).

Por ejemplo:

- evitar la certeza 100% (porque se pierde demasiado puntaje en caso de error)
- sobreutilizar la certeza 100% (olvidando que se debe contestar 100% solo si su *ps* es muy alta)

- utilizar solo dos grados (por ejemplo 20% y 80%)
- utilizar los grados del centro (40% y 60%)
- utilizar el mismo grado para todas las respuestas (por ejemplo 40%, porque en la consigna 4 no se pierde ningún punto en caso de error, y se gana 1)

Dos problemas graves resultan de este tipo de consigna:

*Problema 1*: Los estudiantes se preocupan más del puntaje (de la escala) que de la certeza que tienen en sus respuestas, de modo que el docente no puede fiarse del grado de certeza elegido. Tienen razón al hacerlo, porque se saben poco realistas: se sobrestiman, lo que confirma la literatura (ver Capítulo 16, sección E).

*Problema 2:* El uso de las certezas puede influenciar negativamente la calificación. Por eso a los estudiantes no les gustan.

Durante 10 años (2000-2010) he acumulado evidencia suficiente como para afirmar la robustez de una nueva consigna; para asegurar su resistencia a la falsificación y vincular la certeza solo con ventajas (puntos adicionales). Por lo tanto, propongo ahora abandonar un sistema de cotejo basado en la calidad *de cada respuesta*, y adoptar uno basado en *índices globales de la prueba*, como la Confianza y la Imprudencia, y reemplazar la consigna 5 por aquella que se describe en la sección E3 del presente capítulo, en la cual los grados de certeza entregan solo puntos adicionales, nunca negativos. Sin embargo, deben reunirse varias condiciones para alcanzar este puntaje adicional, *siendo central el concepto de realismo* (ver sección F).

## D. El camino 4: Calificar según la medición del dominio subjetivo (Confianza y Prudencia)

### D.1. La Confianza y la Prudencia

Estos son dos *índices del dominio* subjetivo que ha alcanzado un estudiante en una prueba, ya sea del contenido o de los procesos mentales.

La *Confianza* es el promedio de los grados de certeza que acompañan a las respuestas correctas.

La *Prudencia* es lo contrario de la *Imprudencia*, es decir, el promedio de los grados de certeza acompañando las respuestas incorrectas. Mientras más baja es la Imprudencia, más alta es la Prudencia, por lo que utilizaremos el índice matemático de Imprudencia para medir su contrario, la Prudencia.

Son llamados "Índices de resolución" (o con capacidad de diferenciar lo correcto de lo erróneo) porque se insiste sobre la diferencia entre los dos: lo mejor es tener una *Confianza alta* y una *Imprudencia baja.*

## D.2. Estadísticas de Confianza y de Imprudencia

La Tabla 6 muestra los valores promedio de estos índices (de resolución), para el conjunto de las respuestas *del grupo*, en dos pruebas de los exámenes de enero del curso ISE durante 3 años sucesivos[148].

Tabla 6: Valores promedio de 3 grupos (cohortes) de estudiantes

| CURSO ISE-ULG DE 1ER SEMESTRE DE PREGRADO | 2007 - 2008 | 2008 - 2009 | 2009 - 2010 | |
|---|---|---|---|---|
| Número de estudiantes | 180 | 224 | 168 | |
| Examen de memoria (libro cerrado) de 30 Preguntas con Soluciones Numerosas (PSN) | *83* | *80* | *81* | CONFIANZA MEDIA |
| | 35 | 26 | 30 | Imprudencia media |
| Examen de comprensión-análisis (libro abierto) de 30 Preguntas de Selección Múltiple (PSM) + SGI | 67 | 69 | 69 | CONFIANZA MEDIA |
| | 52 | 52 | 51 | Imprudencia media |

Para el examen de memoria con Libro Cerrado:

(1) La *Confianza* promedio (entre todos los estudiantes) es casi idéntica cada año (cerca de 80%).
(2) La *Imprudencia* promedio presenta mayor variación entre los años: de 26% a 35%.
(3) El *Matiz* (o *Nuancia*) promedio (diferencia entre la Confianza promedio y la Imprudencia promedio) varía de 48% (en 2007-2008) hasta 54% (en 2008-2009), es decir, se ubica alrededor de 50%. Eso constituye un gran poder de *resolución*.

Para el examen de comprensión-análisis con Libro Abierto:

(1) La *Confianza* promedio es también casi idéntica durante los tres años (cerca de 70%, es decir, 10% menos que en el otro examen).
(2) La *Imprudencia* promedio es casi idéntica cada año (cerca de 50%, muy superior al examen de memoria).
(3) El *Matiz* promedio es casi idéntico cada año (cerca de 30%).

Estas diferencias (menor confianza, mayor imprudencia, menor matiz) son otra evidencia de la mayor dificultad que tiene autoevaluarse en comprensión-análisis, comparado con autoevaluarse en memorización. Lo mismo fue observado en el proyecto MOHICAN[149] (ver Figura 5):

[148] Por supuesto, las preguntas eran diferentes cada año.

[149] Los números de estudiantes (NE) sobre los que se calcularon las estadísticas de la Figura 5 son: Vocabulario (NE = 3.905); Matemáticas (NE = 2.539); Historia (NE = 1.418); Artes (NE = 1.399). Leclercq (2003, p. 67-91).

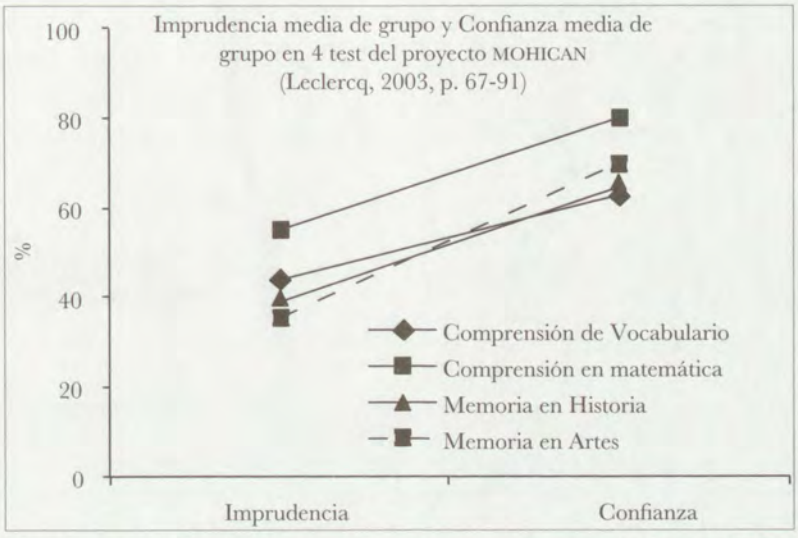


Figura 5: Diferencias entre la Confianza promedio (del grupo) y la Imprudencia promedio (del grupo) en 4 pruebas del proyecto MOHICAN

En el test de Matemáticas, *la Confianza promedio* es la más alta de los 4 tests, lo que podría ser una buena noticia. Sin embargo, el que *la Imprudencia promedio* sea también la más alta (incluso por sobre 50%) indica una sobrestimación. La tasa o *% de Confianza útil* es *87,5%* (56 / 64). La tasa o *% de Imprudencia peligrosa* es *58%* (18 / 31), lo que es enorme.

En los otros tests, la Confianza promedio es >50% y la Imprudencia promedio <50%, lo que es más satisfactorio. La tasa o *% de Confianza útil* es *71%* (34 / 48). La tasa o *% de Imprudencia peligrosa* es solo de *30%* (13 / 43).

## D.3. Índices de Resolución: Discriminación y Lucidez

*La Discriminación positiva* (o Disc +) es la diferencia entre la Confianza promedio y el umbral de satisfacción (aquí 50%).

*La Discriminación negativa* (o Disc –) es la diferencia entre la Imprudencia promedio y el umbral de satisfacción (aquí 50%):

*Disc* + = Confianza promedio – Umbral de Satisfacción (idealmente positiva)
*Disc* – = Imprudencia promedio - Umbral de Satisfacción – (idealmente negativa)

El *Matiz* (o Nuancia) es la diferencia entre la Confianza promedio y la Imprudencia promedio.

*Matiz* = ConfM - ImprM

La Figura 6 presenta el caso de una Confianza promedio de 75% (Disc + = +25%) y de una Imprudencia promedio de 55% (Disc– = +5). El hecho que Disc- tenga un valor positivo no tiene ningún impacto sobre el Matiz, que aquí vale +20 (como si la confianza promedio fuera 60% y la imprudencia promedio fuera 40%).

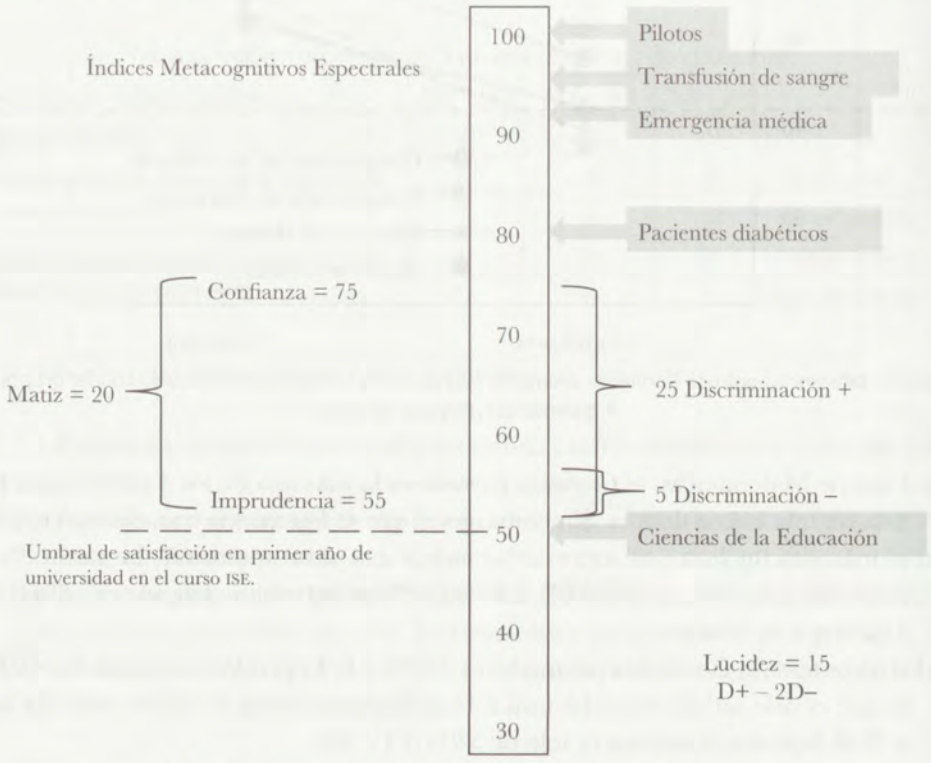


Figura 6: Un caso teórico ilustrando el cálculo de la Lucidez

## D.4. El índice de "Lucidez"

El Matiz no es lo suficientemente sensible como para detectar cuando la Disc– es mayor o menor que cero. Para corregir esto propongo calcular un índice que valoriza la *prudencia* solamente cuando la Disc– es negativa (es decir, en la dirección esperada), lo que ocurre cuando la Imprudencia promedio es inferior al "umbral de satisfacción de la Imprudencia" (por ejemplo 50%). Propongo, entonces, el concepto de *Lucidez*, y la siguiente fórmula:

$$Lucidez = (\text{Disc}+) - 2\,(\text{Disc}-)$$

En el ejemplo de la Figura 6, Disc+ = +25%, Disc– = +5%, y *Lucidez = 15%*.

En consecuencia, *la Lucidez* = 25 – (2 * +5) = 25 – 10 = *15* (aunque el Matiz, en el mismo ejemplo, sea 20).

Uno de mis estudiantes calculó los índices de Lucidez (siendo 50% el umbral de Imprudencia) en los exámenes de enero (Libro Abierto y Libro Cerrado), para tres cohortes del curso ISE:

Tabla 7: La Lucidez promedio calculada para tres cohortes que rindieron exámenes en el curso ISE[150]

| 50 ↑ Umbral (%) de satisfacción | Libro Abierto (30 PSMs + SGI) | | | | Núm. de estudiantes | Libro cerrado (30 PSNs) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Imprudencia media | Confianza media | Matiz | Lucidez | | Imprudencia media | Confianza media | Matiz | Lucidez |
| 2008 | 51,2 | 67,2 | 16 | 14,8 | 180 | 34,9 | 83 | 48,1 | 63,2 |
| 2009 | 52,4 | 68,9 | 16,5 | 14,1 | 224 | 25,6 | 79,5 | 53,9 | 78,3 |
| 2010 | 50,7 | 68,7 | 18 | 17,3 | 168 | 29,5 | 81,2 | 51,7 | 72,2 |

Hoy en día la Lucidez es el índice que mejor representa la discriminación entre la Confianza promedio y la Imprudencia promedio, incluyendo una penalización cuando el valor de cualquiera de estos índices no está ubicado en el lado correcto del umbral de satisfacción. Fue utilizado en un análisis correlacional que enseñó que la Lucidez tiene una validez de predictividad superior a la del Matiz (ver Capítulo 9, sección G2, Figura 10).

## E. El camino 5: Pautas calculadas según los índices metacognitivos de dominio

### E.1. El principio general

Leclercq y Poumay (2003) propusieron un sistema de cotejo (*a proper scoring rule*) que, al asignar puntaje, combina la medición objetiva de la performance y la medición subjetiva del dominio. Esto último siempre y cuando el estudiante mantenga la presunción de realismo.

Se sustenta en 4 principios:

- Calcular el puntaje clásico
- Verificar el realismo (ver sección F a continuación)
- Calcular la Confianza promedio y la Imprudencia promedio
- Calcular los "plus metacognitivos" según índices metacognitivos de dominio

[150] Derochette, 2012, p. 100.

## E.2. Principio 1: Calcular el puntaje clásico

Supongamos que el *puntaje clásico* se calcule según el siguiente baremo (lista de cotejo):

- por una respuesta correcta: +1 punto
- por una respuesta incorrecta u omisión: –0,25

El *puntaje clásico* es el total obtenido, *expresado sobre 20 puntos*, que es el valor máximo habitual en Bélgica, o *sobre 7,0 puntos*, que es el máximo usado en Chile. Por ejemplo, si una prueba tiene 42 preguntas, y el estudiante ha dado 30 correctas, 10 incorrectas y 2 omisiones, su puntaje clásico es: (30 x 1) – (12 x –0,25) = 30 – 3 = 27.

En una escala de 0 a 20, su puntaje clásico es 27/42*20 = 0,64 * 20 = 12,8.
En una escala de 0 a 7, su puntaje clásico es: 27/42*7 = 0,64 * 7 = 4,5.

## E.3. Los bonus metacognitivos y sus criterios de obtención

Se calculan la Confianza promedio y la Imprudencia promedio. *Si el estudiante ha mantenido la presunción de realismo* (*ver sección F*), se aplican los siguientes *bonus* ("plus metacognitivos")[151], según las escalas (de 0 a 20 o de 0 a 7):

Tabla 8: Lista de cotejo de Leclercq y Poumay (2003)

| Criterio | Tipo de escala | |
|---|---|---|
| Si la Confianza promedio es | de 0 a 20 | de 0 a 7 |
| >50%, el bonus equivale a | + 0,5 | +0,2 |
| >60%, el bonus equivale a | +1 | +0,4 |
| >70%, el bonus equivale a | +1,5 | +0,6 |
| Si la Imprudencia promedio es | | |
| <50%, el bonus equivale a | + 0,5 | +0,2 |
| <45%, el bonus equivale a | +1 | +0,4 |
| <40%, el bonus equivale a | +1,5 | +0,6 |

Lo anterior implica que:

(1) la metacognición, expresada por los grados de certeza, solo puede mejorar el puntaje clásico, nunca disminuirlo (eso evita el camino 5a),
(2) esta mejora *posible* es de hasta 3 puntos sobre 20 (15%) o de hasta 1,2 sobre 7,0 (17%).

[151] Esta proporción (3 sobre 20) fue decidida en un debate con Jean-Loup Castaigne.

(3) estas mejoras deben ser merecidas por el estudiante de dos formas:
(a) debe mantenerse la presunción de realismo (ver sección F), y
(b) debe tener buenos índices de Confianza y de Imprudencia.

### E.4. Reglas adicionales

(1) Si un estudiante no utiliza los grados de certeza, o si no entrega los grados en un porcentaje suficiente de las respuestas (90% por ejemplo), no tiene acceso a los "plus metacognitivos".
(2) Si un estudiante obtiene, en una prueba, un total superior al máximo (20/20 o 7,0/7,0), el puntaje de los "plus metacognitivos" de esa prueba contribuirán a la suma con los puntajes de las otras evaluaciones del semestre. Por supuesto, si este puntaje final del semestre es superior a 20 o 7,0, la nota oficial del curso será 20 o 7,0.
(3) Si un estudiante ha dado todas las respuestas correctas, no se puede calcular su Imprudencia promedio y no se entregan los "plus" vinculados a ella. Por esta razón, se entrega el doble de puntaje en los "plus" asociados a la Confianza.

### E.5. Algunos resultados del sistema

A continuación se muestran datos de tres años, obtenidos en el examen de enero por los estudiantes que habían rendido las tres pruebas formativas anteriores (de octubre, noviembre y diciembre):

Tabla 9: Promedios de notas clásicas (máxima 20) y de "plus" metacognitivos (máximo 3 puntos) en tres cohortes

| Curso ISE-ULG de 1er semestre de pregrado | 2007 – 2008 | 2008 – 2009 | 2009 – 2010 | |
|---|---|---|---|---|
| Número de estudiantes | 308 | 327 | 277 | |
| Examen de memoria (libro cerrado) de 30 PSN | 11,8 | 10,6 | 11,9 | Nota clásica |
| | 1,8 | 1,5 | 1,9 | Plus metacognitivo |
| Examen de comprensión–análisis (libro abierto) de 30 PSM + SGI | 9,2 | 8,9 | 7,7 | Nota clásica |
| | 1,4 | 1,4 | 0,6 | Plus metacognitivo |

## F. El camino 6: El realismo

### F.1. El realismo: condición de fiabilidad y postulado docimológico

Una cuestión frecuentemente debatida es

"Cuando un estudiante añade grados de certeza a sus respuestas, ¿sabe el docente más sobre el conocimiento de ese estudiante que cuando no hay grados de certeza?"

Responder automáticamente *"sí"* a esta pregunta (sin considerar el realismo del estudiante) constituye el *tercer error* que ha sido cometido en el ámbito de los grados de certeza. Mi respuesta es:

> *Depende: según si el estudiante es realista o no.* Si es realista, el docente sabe más, pero si el estudiante no es realista (si se subestima o sobrestima), el docente tiene más datos, pero estos datos son "ruido" (en términos de teoría de la información). Esta situación es similar a la de los miembros de un tribunal que escuchan un testigo: saben más cosas fiables sobre lo que pasó si el testigo ha dicho la *verdad*, pero si ha dicho una *mentira*, lo que conocen los miembros del tribunal relativo a lo que pasó se vuelve más embrollado que antes de la declaración del testigo.

Mi *postulado*[152] es el siguiente:

| Solo las expresiones realistas de certeza deben ser tenidas en cuenta para dar a los estudiantes puntajes adicionales en sus calificaciones. |
|---|

Habitualmente los profesores atribuyen a los estudiantes la *presunción* de realismo, pero no la verifican, siendo incapaces de detectar las Confianzas y Prudencias *verdaderas* que merecen ser reconocidas con bonos de puntajes.

### F.2. El realismo: una presunción que debe ser verificada

El *principio de caridad interpretativa*, de Quine y Davidson, que aplican los filósofos (Bruno Leclercq, 2008, p. 224) consiste en un "postulado de racionalidad", es decir, considerar, a priori, que el autor que leen es competente en el ámbito discutido, que ha leído los autores fundamentales, que es lógico (racional) y de buena fe… a no ser que entregue evidencias de lo contrario. Smedslund (1997) hace este mismo postulado aplicado a la lógica, y atribuye los errores a una falta de comprensión, no de lógica.

En mi caso, hago la *presunción de realismo* de todos los estudiantes. Cada uno es considerado, *a priori*, como realista, lógico, y de buena fe…, a no ser que entregue datos que evidencien lo contrario. Esto es lo que verifico sistemáticamente, para que solo los que efectivamente son realistas se beneficien de bonos, si sus Confianzas y Prudencias lo merecen.

### F.3. Las zonas de aceptabilidad de la presunción de realismo

Cuando el *n*úmero de *u*sos (*nu*) de un grado de certeza, por ejemplo 20% (es decir *ps* = 0,2), es 5 (*nu* = 5), las *p*robabilidades para cada *n*úmero de respuestas *c*orrectas (*nc*),

[152] "Un postulado es un principio teórico o metodológico fundamental (no una simple hipótesis), siendo una presunción una tesis considerada como verdadera hasta que existen evidencias de lo contrario" (B. Leclercq, 2006, pp. 185-187).

sobre las 5, pueden ser calculadas usando la ley binomial (ver sección H). Es decir, la ley binomial nos permite calcular la probabilidad de que, en las 5 preguntas en que un estudiante ha contestado diciendo "20% seguro", sean correctas 0, 1, 2, 3, 4 o 5 de ellas. A estas probabilidades las llamaremos *pB* (*p*robabilidades *B*inomiales). La Tabla 10 muestra los valores de *pB* para las 6 posibilidades de *nc*, cuando *nu* = 5 y *ps* = 0,2.

Tabla 10: Valores de pB calculados según la ley binomial

| Para nc = | 0 | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|
| pB = | 0,328 | 0,41 | 0,205 | 0,051 | 0,006 | 0,00004 | 1 |

Se puede ver que algunos valores de respuestas correctas (*nc*) son más compatibles que otros con la presunción de realismo, dependiendo del riesgo a equivocarse que quiera correr el docente. Denominaremos *p* a la probabilidad de equivocarse si se rechaza la presunción de realismo del estudiante.

Así, si se quiere trabajar con:

- $p<0,1$ (es decir, menos de 10% de probabilidades de equivocarse si se rechaza la presunción de realismo), se debe mantener esta presunción si se obtiene entre 0 y 2 respuestas correctas; 0 y 2 son los límites de la zona de aceptabilidad del realismo
- $p<0,05$ y $p<0,01$ (es decir, menos de 5% y de 1% de probabilidades de equivocarse si se rechaza la presunción de realismo), se mantiene la presunción si hay entre 0 y 3 respuestas correctas; 0 y 3 son los límites de la zona
- $p<0,0001$ (es decir, menos de 0,01% de probabilidades de equivocarse si se rechaza la presunción de realismo), con entre 0 y 4 respuestas correctas se mantiene la presunción

Para un docente-investigador es fácil calcular los límites de confianza basándose en la fórmula del Error Estándar de Medición (EEM –ver sección H1), cuando el *nu* de cada grado de certeza es un número grande, pues se hace según la distribución de Gauss. Sin embargo, es mucho más complicado cuando esos *nu* son pequeños (lo que con frecuencia ocurre en pruebas donde no puede haber decenas de preguntas), porque se debe calcular según la ley binomial. Afortunadamente, estas zonas de aceptación del realismo se pueden comprobar en tablas (ábacos). Basándome en la tabla de Diem (1963, p. 104), propongo que se utilicen los ábacos que se presentan a continuación (tablas 11 y 12), dependiendo de si se quiere trabajar con $p<0,05$ o con $p<0,01$.

Las tablas 11 y 12 presentan las zonas de realismo o los "intervalos de aceptación de la presunción de realismo"; la 11 a $p<0,05$ (5 posibilidades sobre 100 de equivocarse rechazando la presunción de realismo), y la 12 a $p<0,01$ (1 posibilidad sobre 100 de equivocarse al rechazarla). Se presentan los intervalos para *nu* variando de 5 a 30;

con *nu* inferior a 5 no hay datos suficientes para ninguna operación, y simplemente se mantiene la presunción de realismo. En cada celda se ubica la pareja de números que constituyen el límite inferior y el límite superior de esta zona de aceptabilidad de *nc*.

Tabla 11: Intervalos (zonas o límites) fuera de los cuales se rechaza la presunción de realismo a p<0,05

| | p<0,05 | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *nu* | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| Probabilidad subjetiva (ps) | *0,05* | 0-2 | 0-3 | 0-3 | 0-3 | 0-3 | 0-3 | 0-3 | 0-3 | 0-3 | 0-4 | 0-4 | 0-4 | 0-4 | 0-4 | 0-4 | 0-4 | 0-4 | 0-4 |
| | *0,2* | 0-4 | 0-4 | 0-5 | 0-5 | 0-5 | 0-6 | 0-6 | 0-6 | 0-7 | 0-7 | 0-7 | 0-8 | 0-8 | 0-8 | 0-8 | 0-9 | 0-9 | 0-9 |
| | *0,4* | 0-5 | 0-6 | 0-6 | 0-7 | 0-8 | 0-8 | 0-9 | 1-9 | 1-10 | 1-10 | 1-11 | 2-11 | 2-12 | 2-12 | 3-13 | 3-13 | 3-14 | 3-14 |
| | *0,6* | 0-5 | 0-6 | 1-7 | 1-8 | 1-9 | 2-10 | 2-11 | 3-11 | 3-12 | 4-13 | 4-14 | 5-14 | 5-15 | 6-16 | 6-16 | 7-17 | 7-18 | 8-19 |
| | *0,8* | 1-5 | 2-6 | 2-7 | 3-8 | 4-9 | 4-10 | 5-11 | 6-12 | 6-13 | 7-14 | 8-15 | 8-16 | 9-17 | 10-18 | 11-19 | 11-20 | 12-21 | 13-22 |
| | *0,95* | 3-5 | 3-6 | 4-7 | 5-8 | 6-9 | 7-10 | 8-11 | 9-12 | 10-13 | 10-14 | 11-15 | 12-16 | 13-17 | 14-18 | 15-19 | 16-20 | 17-21 | 18-22 |

| | *nu* | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Probabilidad subjetiva (ps) | *0,05* | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-6 | 0-6 | 0-6 | 0-6 | 0-6 | 0-6 | 0-6 |
| | *0,2* | 0-10 | 0-10 | 0-10 | 1-11 | 1-11 | 1-11 | 1-11 | 1-12 | 1-12 | 1-12 | 1-12 | 2-13 | 2-13 | 2-13 | 2-13 | 2-14 | 2-14 | 2-14 |
| | *0,4* | 4-15 | 4-15 | 4-16 | 5-16 | 5-17 | 5-17 | 6-18 | 6-18 | 6-19 | 7-19 | 7-20 | 7-20 | 7-21 | 8-21 | 8-22 | 8-22 | 9-23 | 9-23 |
| | *0,6* | 8-19 | 9-20 | 9-21 | 10-21 | 10-22 | 11-23 | 11-23 | 12-24 | 12-25 | 13-25 | 13-26 | 14-27 | 14-28 | 15-28 | 15-29 | 16-30 | 16-30 | 17-31 |
| | *0,8* | 13-23 | 14-24 | 15-25 | 15-25 | 16-26 | 17-27 | 18-28 | 18-29 | 19-30 | 20-31 | 21-32 | 21-32 | 22-33 | 23-34 | 24-35 | 24-36 | 25-37 | 26-38 |
| | *0,95* | 18-23 | 19-24 | 20-25 | 21-26 | 22-27 | 23-28 | 24-29 | 25-30 | 26-31 | 27-32 | 28-33 | 28-34 | 29-35 | 30-36 | 31-37 | 32-38 | 33-39 | 34-40 |

Tabla 12: Intervalos (zonas o límites) fuera de los cuales se rechaza la presunción de realismo a p<0,01

| | p<0,01 | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *nu* | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| probabilidad subjetiva (ps) | *0,05* | 0-3 | 0-3 | 0-3 | 0-4 | 0-4 | 0-4 | 0-4 | 0-4 | 0-4 | 0-4 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 |
| | *0,2* | 0-5 | 0-5 | 0-5 | 0-6 | 0-6 | 0-7 | 0-7 | 0-7 | 0-8 | 0-8 | 0-8 | 0-9 | 0-9 | 0-9 | 0-10 | 0-10 | 0-10 | 0-11 |
| | *0,4* | 0-5 | 0-6 | 0-7 | 0-8 | 0-8 | 0-9 | 0-10 | 0-10 | 0-11 | 0-11 | 0-12 | 1-12 | 1-13 | 1-14 | 1-14 | 1-15 | 1-15 | 2-16 |
| | *0,6* | 0-5 | 0-6 | 0-7 | 0-8 | 0-9 | 1-10 | 2-11 | 2-12 | 2-13 | 3-14 | 3-15 | 4-15 | 4-16 | 4-17 | 5-18 | 5-19 | 6-20 | 6-20 |
| | *0,8* | 0-5 | 1-6 | 2-7 | 2-8 | 3-9 | 3-10 | 4-11 | 5-12 | 5-13 | 6-14 | 7-15 | 7-16 | 8-17 | 9-18 | 9-19 | 10-20 | 11-21 | 11-22 |
| | *0,95* | 2-5 | 3-6 | 4-7 | 4-8 | 5-9 | 6-10 | 7-11 | 8-12 | 9-13 | 10-14 | 10-15 | 11-16 | 12-17 | 13-18 | 14-19 | 15-20 | 16-21 | 17-22 |

| | *nu* | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| probabilidad subjetiva (ps) | *0,05* | 0-5 | 0-6 | 0-6 | 0-6 | 0-6 | 0-6 | 0-6 | 0-6 | 0-6 | 0-6 | 0-7 | 0-7 | 0-7 | 0-7 | 0-7 | 0-7 | 0-7 | 0-7 |
| | *0,2* | 0-11 | 0-11 | 0-12 | 1-12 | 1-12 | 1-12 | 1-13 | 1-13 | 1-13 | 1-14 | 1-14 | 2-14 | 2-15 | 2-15 | 2-15 | 2-15 | 2-16 | 2-16 |
| | *0,4* | 2-16 | 3-17 | 3-17 | 3-18 | 4-18 | 4-19 | 4-19 | 4-20 | 5-21 | 5-21 | 5-22 | 6-22 | 6-22 | 6-23 | 6-24 | 7-24 | 7-25 | 7-25 |
| | *0,6* | 7-21 | 8-21 | 8-22 | 8-23 | 9-23 | 9-24 | 10-25 | 10-26 | 10-26 | 11-27 | 11-28 | 12-28 | 13-29 | 13-30 | 13-31 | 14-31 | 14-32 | 15-33 |
| | *0,8* | 12-23 | 13-24 | 13-25 | 14-25 | 15-26 | 16-27 | 16-28 | 17-29 | 18-30 | 18-31 | 19-32 | 20-32 | 20-33 | 21-34 | 22-35 | 23-36 | 23-37 | 24-38 |
| | *0,95* | 18-23 | 18-24 | 19-25 | 20-26 | 21-27 | 22-28 | 23-29 | 24-30 | 25-31 | 26-32 | 26-33 | 27-34 | 28-35 | 29-36 | 30-37 | 31-38 | 32-39 | 33-40 |

Así, si un estudiante ha utilizado 12 veces el grado de certeza 40% (*nu*=12 y *ps*=0,4), a p<0,05, la presunción de realismo se mantiene si el número de respuestas correctas

(*nc*) está incluido entre 1 y 9. Si este número de éxito es mayor que 9, se podría hablar de subestimación, y si es menor que 1, de sobrestimación.

En inglés llamo a esto *the BLAC procedure*:
*Binomial Law Applied to Confidenced responses.*

En francés y en español, propongo la expresión *Real-Bin*, o "*Real*ismo verificado por la ley *Bin*omial".

## F.4. El reloj de las 13 campanadas

Si la presunción de realismo es rechazada para un solo grado de certeza, es rechazada para toda la prueba. Tal decisión es arbitraria, pero está basada sobre el principio del reloj de las 13 campanadas:

*Cuando un reloj toca 13 campanadas,*
*todas las campanadas que ha tocado se vuelven sospechosas.*

## F.5. Estadísticas de rechazo de la presunción de realismo

La verificación del realismo se hace de acuerdo con los criterios expuestos en la sección F3, lo que determinará si se mantiene o se rechaza la presunción de realismo de cada estudiante. En la Tabla 13 se muestran las estadísticas de los rechazos de presunción de realismo con una cohorte de 326 estudiantes en el curso ISE[153]:

Tabla 13: Tasa (en %) de estudiantes excluidos de la presunción de realismo con $p<0,05$ y $p<0,01$

| Tasa de rechazo de la presunción de realismo sobre 326 estudiantes | a $p<0,05$ | a $p<0,01$ |
|---|---|---|
| Libro abierto (comprensión – vigilancia) | 25,8% | 13,5% |
| Libro cerrado (memoria) | 5,5% | 1,5% |

Se observa que en el examen de memoria los estudiantes son más realistas que en el de comprensión, lo que es normal: es más fácil darnos cuenta de lo que no sabemos antes de lo que no comprendemos.

En los análisis estadísticos realizados por uno de nuestros estudiantes (Derochette, 2012) aparece que:

(1) el sistema de verificación del realismo y de los "plus" metacognitivos no presenta diferencias en sus resultados por género, es decir, no aventaja a hombres versus mujeres o viceversa.

153 Introducción a las Ciencias de la Educación, por D. Leclercq a estudiantes de primer año de pregrado en psicología.

(2) los estudiantes que participaron en todos los test formativos que hubo a su disposición durante el curso (3 tests) ganaron más "plus metacognitivos" que los que participaron en solo dos de ellos.

El estudiante no está obligado a utilizar los grados de certeza. Sin embargo, y aunque está permitido no usarlos, hasta el día de hoy no he encontrado un solo estudiante que haya renunciado a hacerlo, y por tanto haya renunciado a optar al "plus" de puntaje por su dominio subjetivo del conocimiento.

La aplicación del nuevo sistema de cotejo –donde primero se establece si el estudiante mantiene su presunción de realismo, y luego se otorgan los "plus" si corresponde– resulta en puntajes (sobre 20) más o menos equivalentes a los puntajes que resultan del sistema basado en la calificación de cada respuesta de acuerdo con su grado de certeza (ver Tabla 4), con la severidad que considera como puntaje máximo todas las respuestas correctas y 60% de certeza (Derochette, 2012, p. 80). Esta severidad es la más generosa, de modo que los estudiantes prefieren el nuevo sistema.

## G. Límites del método y conclusiones

### G.1. Pagos (payoff) y mediciones

El método de cotejo que ha sido privilegiado en este capítulo implica la distinción conceptual entre *dos tipos de medición*:

(1) la medición de la *cognición* (el puntaje clásico), siendo estimada su fiabilidad (replicabilidad o *reliability*) por los intervalos de credibilidad del porcentaje (que depende del número de preguntas, como lo muestra la Figura 4 del Capítulo 13, sección E3)
(2) la medición de la *calidad subjetiva del dominio* (la Confianza y la Imprudencia), siendo garantizada su fiabilidad mediante la confrontación de las tasas de éxito –con cada grado de certeza– y las probabilidades (ábacos –tablas 11 y 12), para verificar la presunción *de realismo*, que es un modo de *estudiar objetivamente la subjetividad.*

Esta forma de asignar puntaje pretende valorizar la *calidad de dominio subjetivo*, es decir:

- la tasa de éxitos
- el realismo (la autoestimación realista del conocimiento parcial)
- la Lucidez, que combina altos niveles de calidad *subjetiva* de dominio (que se manifiesta por un índice de Confianza alto) y una gran Prudencia (índice de Imprudencia bajo)

El puntaje o calificación final (la nota) NO ES UNA MEDICIÓN, sino un pago (*payoff*) según *proporciones arbitrarias*. Confundir los dos conceptos (medición y pago) fue el *cuarto error* que cometieron los investigadores en los años 1960-1970: calcularon correlaciones (con criterios externos) para verificar la validez (predictiva) y la fiabilidad de puntajes utilizando los grados de certeza, sin tener en cuenta el realismo (error 3), que, desde mi punto de vista, es una condición esencial para que los nuevos puntajes sean más fiables y más válidos que el puntaje clásico. Esta es una razón suplementaria para explicar el fracaso de las investigaciones de aquella época (Hassmen y Hunt, 1994).

### G.2. Aceptabilidad de las listas de cotejo

Es importante que las nuevas listas de cotejo resulten en notas a lo menos iguales (o mejores) que las habituales. Si no lo son, existe el riesgo de que sean rechazadas por los estudiantes (tendrían poca validez de aceptabilidad).

### G.3. La ayuda de programas informáticos

El cálculo de los índices se ve enormemente facilitado por el uso de programas informáticos, siendo los más sencillos los realizados con Excel (que tienen la ventaja de ser fácilmente modificables por los docentes mismos). Algunos de estos programas serán desarrollados y estarán disponibles gratuitamente en el sitio: http://orbi.ulg.ac.be. Luego de ingresar se debe teclear "Leclercq Dieudonné"; los documentos aparecen en orden, desde los más recientes hasta los más antiguos según sus fechas de publicación. Algunos programas tienen nombres como SPECTRAL o SPLIT... u otras denominaciones según la investigación en desarrollo.

### G.4. La necesitad de referencias y de acciones-investigaciones adicionales

Conocemos muy poco acerca de cuáles son las capacidades de realismo y las posibilidades de los seres humanos en términos de calidad subjetiva del dominio en varios ámbitos, y en relación con varios procesos mentales. Espero que los conceptos y ejemplos desarrollados en este capítulo constituyan instrumentos para que se pueda comenzar sobre otras bases la exploración de este subestimado (¡el colmo!) campo de acción y de investigación. Porque, al fin y al cabo, sí se puede...

...estudiar objetivamente la subjetividad.

## H. Anexo: las bases matemáticas de las tablas binomiales

### H.1. El error de medición de una Tasa de Éxito (TE) o de una proporción observada (po)

La Tasa de Éxito (TE), o el porcentaje de respuestas correctas con una probabilidad subjetiva (*ps*) específica, es la *p*roporción *o*bservada (*po*) de respuestas correctas dentro del grupo de respuestas dadas con una *ps* determinada (es decir, con un determinado grado de certeza: 60% por ejemplo).

La Tasa de Éxito (TE) puede ser usada como una estimación (*a posteriori*) del realismo de la probabilidad *ps* anunciada, y aumenta su fiabilidad a medida que el número de usos (*nu*) de una *ps* (o grado de certeza) particular es más grande. Para que la TE (o *po*) sirva como criterio para mantener o rechazar la presunción de realismo en una *ps*, cada *po* (o TE) se debe enmarcar dentro de su *Error Estándar de Medición de po*, o EEM *po*, calculado con la fórmula clásica:

$$EEM\ po = \sqrt{po_i . qo_i \ / \ nu_i} \qquad \text{donde } qo_i = 1\text{-}po_i$$

Por ejemplo, consideremos un estudiante ficticio que ha utilizado el grado de certeza 80% (*ps* = 0,8) treinta veces en un test (*nu* =30). Supongamos que 21 de estas 30 respuestas son correctas (es decir, el 70%). De este modo, $po_{80} = 0{,}7$.

El Error Estándar de Medición (EEM) de $po_{80}$ (que es *0,7*) se calcula así:

$$\sqrt{(0{,}7 \ . \ 0{,}3)/30} \ = \sqrt{0{,}21 \ / \ 30} = \sqrt{\ 0{,}007 \ } = 0{,}084.$$ Es decir 8,4%.

En consecuencia, en este ejemplo será 8,4% el valor que se usará para establecer los límites del intervalo de confianza que mantenga la presunción de realismo.

### H.2. El intervalo de confianza de una Tasa de Éxito (TE) o de una proporción observada (po)

En el *gráfico de calibración* que se muestra en la Figura 7, cuando el número de usos (*nu*) es grande (igual o mayor que 30), se establece un intervalo de confianza para la tasa de éxito (TE) de cada una de las *ps* posibles (es decir, para cada uno de los seis grados de certeza). Este intervalo de confianza está formado por un valor que está *1,96* veces el EEM para esa *ps* por encima de la TE, y otro que está 1,96 veces el EEM por debajo del valor de TE. Dentro de este intervalo, el verdadero valor de *po* (o TE) tiene 95% de probabilidades de ubicarse, de estar incluido. Dicho de otra forma, hay solo 5% de probabilidades de que el verdadero valor de *po* se ubique fuera del intervalo de confianza.

En el ejemplo del estudiante ficticio, sobre el valor 70% de *po* se debe añadir 1,96 veces 8,4% (que es el EEM para $po_{80}$); es decir, 16,5%. Así, el límite superior del intervalo de confianza es 70% + 16,5% = 86,5%. Debajo de este 70% también se debe sustraer 1,96 veces 8,4%, de modo que el límite inferior del intervalo de confianza

es 70% - 16,5% = 53,5%. En la Figura 7 los dos estudiantes representados tienen la misma TE para la $ps_{80}$ (o para el grado de certeza 80%), por lo que este intervalo de confianza para $TE_{80}$ es el mismo.

Si se quiere tener *solo un 1% de probabilidades de equivocarse* ($p<0,01$) al declarar que el valor verdadero de *po* está dentro de la zona de confianza, se debe sumar y restar *2,58* veces el Error Estándar de Medición a la TE, generando un intervalo de confianza más amplio en torno a la *po*. Si se acepta la posibilidad de *equivocarse 10% de las veces*, se debe sumar y restar *1,62* el EEM al valor de TE, generando un intervalo de confianza más estrecho (lo que es más exigente para el estudiante, porque aumenta las posibilidades de perder la presunción de realismo).

## H.3. ¿Cuándo se rechaza la presunción de realismo?

La presunción de realismo, de un estudiante en una prueba, se rechaza cuando:

- tenemos suficientes datos (por ejemplo 5 respuestas) con el mismo grado de certeza, y
- el porcentaje de respuestas correctas (TE) con ese grado de certeza (*ps*) no está incluido en el intervalo de confianza para esa *ps*.

En el caso de la Figura 7, las *po* (o TE) de cada grado de certeza están enmarcadas (verticalmente) por líneas que delimitan el intervalo de confianza, con el EEM multiplicado por 1,62, 1,96 y 2,58.

Se ve que con estos límites,

- para el estudiante B (a la derecha) todos los intervalos de confianza incluyen la diagonal, lo que significa que NO se puede rechazar la presunción de realismo para ninguno de los grados de certeza.
- para el estudiante A (a la izquierda) los intervalos de confianza de *po* (o TE) para los grados de certeza 5% y 95% NO incluyen la diagonal. En estos dos casos se puede rechazar la presunción de realismo, con probabilidad $p<0,1$ de equivocarse.

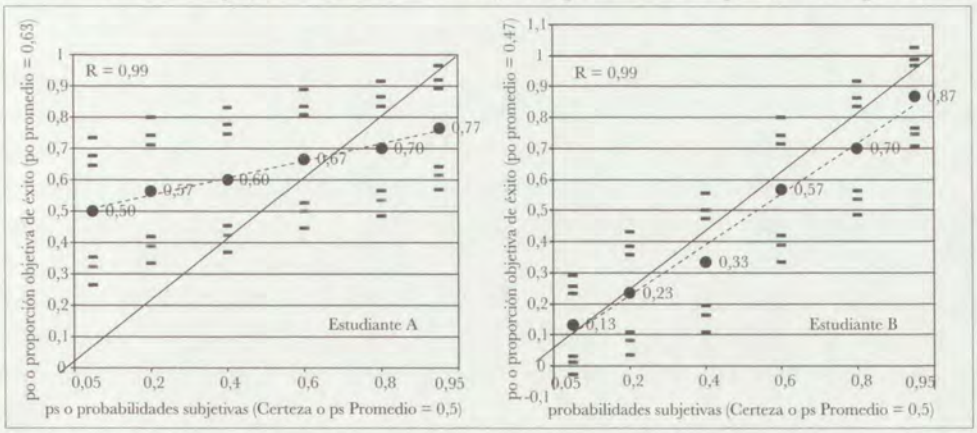


Figura 7: Tasas de Éxito enmarcadas por sus EEM con $p<0,01$, $p<0,05$ y $p<0,1$ para dos estudiantes (A y B)

## H.4. ¿Por qué la ley binomial?

El uso del EEM para determinar los límites de la zona de realismo sería correcto si se trabajara siempre con grandes números de respuestas, para los cuales se puede considerar que el Error Estándar de Medición (EEM) se distribuye en torno a la tasa de Éxito (TE) de un modo normal (según una curva de Gauss). Ahora bien, en numerosos tests el número de preguntas es menor que 30. En consecuencia, el número de uso de UN grado de certeza en particular (por ejemplo 40%) no alcanza los 30 usos. Con números tan pequeños el error de medición (EEM) no es normal (no se distribuye según una curva de Gauss), y no se pueden aplicar los valores de 1,62 o 1,96 o 2,58 para determinar los límites de confianza. Esto hace necesario aplicar la ley binomial.

## H.5. La ley binomial: un ejemplo

Si un estudiante ha elegido el grado de certeza 20% (*ps* = 0,2) 5 veces (*nu* = 5), ¿cuál es la probabilidad de que todas esas 5 respuestas sean correctas, si el estudiante se autoevalúa con realismo?

Esta pregunta se puede responder gracias a la ley binomial, calculando la probabilidad binomial (*pB*) correspondiente. En este ejemplo:

- La *pB* de que de una bolsa en la cual se mezclan 20% de bolas verdes (las correctas) y 80% de bolas rojas (las incorrectas), si 5 bolas son extraídas al azar, todas (las 5) sean verdes.

O, lo que es lo mismo:

- La *pB* de que el número de respuestas correctas (*nc*) sea 5, dado que el número de usos (*nu*) = 5, y *ps* = 0,2.

Esta probabilidad Binomial (*pB*) es 0,00004, o 4 posibilidades sobre 100.000, y se calcula según la fórmula (combinatoria) siguiente:

$$pB_{nc}|_{nu}|_{sp} = C^{nu}_{nc}\,sp^{nc}(1\text{-}sp)^{nu\text{-}nc}$$

donde, *C* = el símbolo de la función de Combinación
*nu* = el número de uso de este grado de certeza
*nc* = el número de respuestas correctas con este grado de certeza
*ps* = la probabilidad subjetiva (entre 0 y 1)

La Tabla 14 presenta los valores de *pB* para los 6 valores de *nc* cuando *nu* = 5 y *ps* = 0,2.

Tabla 14: Valores de pB calculados según la ley binomial

| Para nc = | 0 | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|
| pB = | 0,328 | 0,41 | 0,205 | 0,051 | 0,006 | 0,00004 | 1 |

Recordemos que los valores 1,96 y 2,58, multiplicados por el Error Estándar de Medición (EEM), sirven para obtener un riesgo de equivocarse inferior al 5% (1,96*EEM) o al 1% (2,58*EEM), al rechazar la presunción de realismo. Estos dos valores son válidos solo cuando la distribución de los errores de medición tiene una forma de Gauss. En la Figura 8 se pueden ver las distribuciones de las probabilidades EXACTAS de cada evento *nc* (número de respuestas correctas), para 5 valores de *nu* (número de uso de un grado de certeza), cuando la *ps* = 0,2. Se observan las variaciones de los *números absolutos de nc* que son las Modas o cimas de la curva (es decir, que tienen la mayor probabilidad cuando *ps* = 0,2). Para *nu* = 5, la cima es 1 (porque 1/5 = 0,20). Para *nu* = 10, la cima es 2. Para *nu* = 20, la cima es 4. Para *nu* = 30, la cima es 6. Para *nu* = 40, la cima es 8. Los otros *valores absolutos* de *nc* tienen probabilidades exactas que se distribuyen en forma de campana de Gauss cuando *nu* es mayor a 30. Pero cuando *nu* es inferior a 30, la distribución es asimétrica porque limitando a la izquierda está el cero como mínimo (no existe un evento del tipo "observamos –3 respuestas correctas").

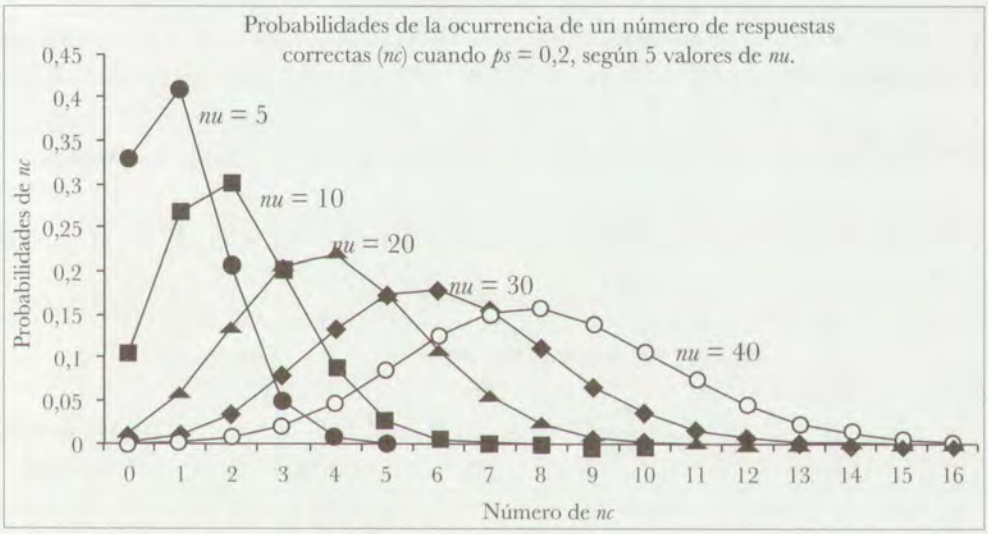


Figura 8: Distribuciones de las probabilidades de observar *nc* cuando *ps* = 0,2 para 5 valores de *nu*

Ahora bien, lo que nos interesa no son los valores exactos de las probabilidades, sino los límites de confianza de una proporción observada (*po*). A causa de la asimetría de las distribuciones, cuando *nu*<30 (lo más frecuente en las pruebas escolares), el modo de calcular estos límites de confianza cambia. Se mantiene el principio de establecer estos límites desde el valor observado $-n$*EEM al valor observado $+n$*EEM, pero *n* ya *no* es 1,96 (con p<0,05) o 2,58 (con p<0,01). Cuando *nu*<30, el proceso de calcular los valores de *n* para cada *nu* es complicado. Felizmente, algunos autores han elaborado tablas con estos valores, como la de Diem (1968). Luego de consultar estas fuentes, y teniendo en cuenta las límites de confianza, he construido los ábacos que se muestran en la sección F3, tablas 11 y 12, para dos valores de p (p<0,05 y p<0,01).

## H.6. Comparación entre coherencia, Centración, calibración y aplicación del ábaco BLAC

El método BLAC presenta dos innovaciones principales, en comparación con métodos ya existentes:

(1) aplica la ley binomial y (2) la aplica a cada grado de certeza (a partir de *nu* = 5).

A diferencia de BLAC, los métodos que se muestran a continuación (coherencia, centración, calibración) calculan *UN índice para la prueba total.*

(1) La *coherencia interna* de la persona se calcula con la fórmula de la *correlación* Bravais-Pearson entre los seis valores de *ps* (0,05; 0,2; 0,4; 0,6; 0,8; 0,95) y la tasa de Éxito (TE) de las respuesta dadas con estas *ps*. Existen casos de coherencia casi perfecta que sin embargo corresponden a un realismo débil.

En la Figura 7 esta correlación es la misma para los dos estudiantes: 0,99, casi perfecta. Pero la calibración 1-MECA[154] de A (0,79) es más baja que la de B (0,92).

> En consecuencia, la coherencia interna no puede ser un índice de realismo.

(2) La *Centración* (Yates, 1990) de cada estudiante en el *total de la prueba* se calcula con la fórmula:

$$Centración = MC - TE_T.$$

donde *MC* = la Media de las Certezas dadas por el estudiante durante la prueba
*TE* = la Tasa de Éxito en el test *total* (o *po* = proporción de Éxito en el test *total*)

En la Figura 7 se ve que el estudiante A tiene una MC (Media de las Certezas) que vale 50%, y una TE que vale 47%, lo que resulta en una centración de 3% (*solo 3% de sobrestimación en el total*). Sin embargo, al mismo tiempo se ve que este valor resulta de la compensación entre sobrestimaciones y subestimaciones.

> En consecuencia, la centración no puede ser un índice de realismo

(3) Existen varios índices de *calibración*. Están basados en el concepto de Error de Calibración (EC) para cada grado de certeza (GC), que se calcula así:

$$\text{Error de Calibración de la certeza}_i = EC_i = GC_i - TE_i.$$

[154] MECA: *Mean Absolute Discrepancy Score*. Ver punto (3) más adelante.

Después, estos errores son sumados (con su ponderación dependiendo de *nu*) para resumir en un solo índice la calibración. Este índice es el MEC o Media de los Errores de Calibración, y se calcula con la fórmula:

$$\text{MEC} = \Sigma\,(nu_i \,.\, EC_i) \,/\, NP$$

Donde $nu_i$ es el número de uso del grado de certeza i y el MEC ideal es 0 (ningún EC)

Con el índice MEC, una sobrestimación en un grado de certeza puede ser compensada por una subestimación en otro grado, resultando en un valor total cercano a 0, lo que puede dar la ilusión de realismo.

En consecuencia, el índice MEC no puede ser un índice de realismo

Por eso Adams y Adams (1961) calculan un "Mean Absolute Discrepancy Score", utilizando los valores *absolutos* de los EC (Errores de Calibración), y el índice *MECA* (Media de los Errores *Absolutos* de Calibración). Otros autores como Murphy (1974) recomiendan utilizar los valores al cuadrado de los ECs (lo que elimina también el signo negativo). Una revisión de estos índices propuestos por autores como Brier (1950), Murphy (1974), Lichtenstein *et al.* (1977) puede verse en Leclercq (1982, p. 229ss) y Schraw *et al.* (2013). Por ejemplo, el realismo puede expresarse como "*1- MECA*". En este caso, 1 es el realismo perfecto (MECA siendo 0).

El valor 1-MECA sí que puede servir como un índice de realismo, y ser utilizado en correlaciones con otras variables.

Pero, si queremos utilizarlo para verificar (o rechazar) la presunción de realismo,

- se debe indicar el valor de este índice 1-MECA a partir del cual la rechazamos (y decir sobre cuáles argumentos)
- el índice 1-MECA no muestra cuál(es) uso(s) de cuál(es) grado(s) de certeza explican el valor del índice.

Además, este índice necesita cálculos simples pero numerosos.

En consecuencia,

- si se trata de tener un índice métrico del realismo, el valor 1-MECA puede convenir.
- si se trata de tener un criterio que se puede consultar rápidamente, para rechazar o mantener la presunción de realismo, recomiendo el uso del ábaco BLAC

## Referencias

Adams, J.K. y Adams, P.A. (1961). Realism of confidence judgments. Psychological Review, 68, 33-45.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability, Montly Weather Review, 75, 1-3.

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.

Derochette, D. (2012). Analyse d'indices cognitifs et métacognitifs pour un cours de première année universitaire: Quelles différences selon des sous-groupes? (Genre, âge et orientation). Master Tesis en educación. Universidad de Liège.

Diem, K. (1963). Tables scientifiques. Bâle: Geigy.

Hassmen, P. y Hunt, D. (1994). Human Self-Assessment in Multiple-Choice Testing. Journal of Educacional Measurement. Vol. 31, 2, 149-160.

Jacobs, S. (1971). Correlates of unwarranted confidence in responses to objective test items. Journal of Educational Measurement, 8, 1, pp. 15-20.

Kahneman, D. y Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), pp. 263-291.

Koehler, R. (1974). Over confidence on probabilistic tests. Journal of Educational Measurement, 11, 101-108.

Leclercq, B. (2006). Introduction à la logique et à l'argumentation. Editions de l'Université de Liège.

Leclercq, B. (2008). Introduction à la philosophie analytique. Bruxelles: De Boeck.

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. 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. (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 Monseur, Ch. (2012). Vérifier le réalisme par degrés de certitude au moyen de la loi binomiale. Artículo sometido a publicación.

Lichtenstein, S., Fischhoff, B. y Phillips, L. (1977). Calibration of probabilities: The state of the art. In Jungermann y De Zeeuw (Eds): Decision making and change in human affairs. Proceedings of the 5th spudm Conference, Darmstadt: Reidel Publishing Company, 275-324.

Michael, J. (1968). The reliability of multiple choice examination under various test-taking instructions. Journal of Educational Measurement, 5, 307-314.

Murphy, A.H. (1974). A sample skill score for probability forecasts. Montly Weather Review, 102, 48-55.

Schraw, G., Kuch, F. y Gutierrez, A. (2013). Measure for measure: Calibrating ten commonly used calibration scores. Learning and Instruction 24, p. 48-57.

Smedslund, J. (1997). The forgotten variable of understanding. Cahiers de Psychologie Cognitive – Current Psychology of Cognition, 16 (1-2), 217-221.

Yates, J. F. (1990). Judgment and decision making. Englewood Cliffs: Prentice-Hall.