# Análisis del proceso de construcción de un cuestionario sobre probabilidad condicional

**CARMEN BATANERO E CARMEN DÍAZ***

**Resumen**

Describimos brevemente el proceso de elaboración de un cuestionario de evaluación, y lo analizamos desde el enfoque onto-semiótico de la educación matemática. La finalidad es reflexionar sobre las instituciones y procesos de muestreo implicados, así como sobre las posibilidades de generalización y criterios de idoneidad de las tareas e instrumentos de evaluación.

***Palabras claves:*** evaluación; probabilidad condicional; enfoque ontosemiótico; validación de cuestionarios.

***Abstract***

*This paper summarises the process of building an assessment questionnaire and analyses it from the standpoint of the onto-semiotic approach to mathematics education. The aim is to reflect on the institutions and sampling processes involved, the possibilities of generalization, and the suitability criteria for assessment instruments and tasks.*

**Key-words:** *assessment; conditional probability; onto-semiotic perspective; questionnaire validity.*

***Resumo***

*Descrevemos brevemente o processo de elaboração de um questionário de avaliação, e o analisamos sob o ponto de vista do enfoque onto-semiótico da Educação Matemática. O objetivo é refletir sobre as instituições e processos de amostragem implicados, assim como sobre as possibilidades de generalização e critério de idoneidade das tarefas e instrumentos da avaliação.*

**Palavras-chave:** *avaliação; probabilidade condiconal; enfoque ontosemiótico; validação de questionários.*

* Universidad de Granada. E-mail: batanero@ugr.es, mcdiaz@ugr.es

## Introducción

La construcción de instrumentos de evaluación es habitual en la investigación en educación, donde se siguen las normas metodológicas habituales en psicometría. Desde el punto de vista del Enfoque ontosemiótico para la educación matemática (Godino, 1999, 2002, 2003; Godino, Batanero y Roa, 2005), un instrumento de evaluación tiene como finalidad proporcionar información sobre los *significados personales* de un grupo de estudiantes sobre un objeto o un grupo de objetos matemáticos dados.

La investigación encaminada a la construcción de estos instrumentos o el análisis de las respuestas a los mismos trata de describir la *estática de significados sistémicos*, esto es, la caracterización de la trama de las funciones semióticas (o al menos una muestra representativa de tal trama) en las cuales un objeto se pone en juego en un contexto y circunstancias fijadas. La "medida" de tales significados (sistemas de prácticas) tendrá un carácter cualitativo y será relativa a una persona, institución, contexto fenomenológico y momento temporal especificado.

Puesto que los significados dependen de los contextos sociales y de los sujetos, su carácter es relativo. Respecto al significado institucional, en este marco teórico se diferencia entre el global (qué significa la media), referencial (qué significado de la media se considera en una enseñanza o investigación), pretendido (qué se pretende enseñar de la media en una experiencia de enseñanza), implementado (qué se logra enseñar) y evaluado (qué parte se evalúa).

Por otro lado, ni el significado global de un objeto de enseñanza ni siquiera los significados pretendidos o implementados en una enseñanza sobre dicho objeto pueden ser abarcados en un solo instrumento de evaluación. Tampoco el significado personal del alumno puede ser explicitado completamente en las respuestas a tareas o la observación de su actividad durante una prueba.

En este trabajo describiremos brevemente el proceso seguido en la construcción de un cuestionario dirigido a evaluar la comprensión de la probabilidad condicional por estudiantes de Psicología. La finalidad principal es reflexionar sobre la complejidad de la función evaluadora, los diferentes niveles y tipos de significados personales e institucionales involucrados y sobre la información que diferentes tipos de análisis psicométricos pueden proporcionar respecto a las posibilidades de generalización en la investigación educativa.

## El contexto de la investigación

La construcción del cuestionario citado forma parte de una investigación realizada con estudiantes de psicología en la Universidad de Granada (Díaz, 2004; Díaz y de la Fuente, 2006). Se eligió el concepto de probabilidad condicional, que es fundamental en las aplicaciones de la Estadística, porque permite incorporar cambios en nuestro grado de creencia sobre los sucesos aleatorios a medida que adquirimos nueva información. Es también un concepto teórico básico requerido en el estudio de la inferencia estadística, tanto clásica como bayesiana, así como en el estudio de la asociación entre variables, la regresión y los modelos lineales. En el terreno profesional e incluso en la vida cotidiana, la toma de decisiones acertadas en situaciones de incertidumbre se basa en gran medida en el razonamiento condicional.

## Constructos y variables

Al tratar de evaluar la comprensión sobre un cierto concepto de un grupo de alumnos, hemos de tener en cuenta que es un *constructo inobservable* (León y Montero, 2002), por lo que sus características deben ser inferidas de las respuestas de los alumnos. Un constructo es un atributo psicológico que caracteriza los comportamientos de los individuos y permite explicar patrones de comportamiento. Sólo pueden ser observados indirectamente y están sujetos al cambio, de aquí la dificultad de su evaluación, que llevamos a cabo mediante alguna *variable* observable, por ejemplo, la puntuación en un cuestionario (Osterlind, 1989). Generalmente hay más de una posible forma de definir el constructo, cuya definición se realiza a dos niveles:

- *Definición semántica*: en términos de comportamientos observables o reglas de correspondencia entre el constructo y la conducta.
- *Definición sintáctica:* en términos de las relaciones lógicas o matemáticas del constructo con otros constructos o variables dentro de un marco teórico.

En este trabajo nos limitaremos a la definición semántica, que considera la especificación detallada de la variable de interés (en este caso la comprensión de la probabilidad condicional). Desde el enfoque ontosemiótico, las especificaciones del contenido podrían describirse como tipos de prácticas operatorias y discursivas (Godino, 2003) asociadas al

objeto "probabilidad condicional". La diferenciación entre constructos y variables también se recoge en este marco teórico, donde se diferencia entre el dominio de las ideas u objetos abstractos (personales e institucionales) y el dominio de los significados o sistemas de prácticas de donde emergen tales objetos inobservables (Godino, 1999), lo que permite plantear con nitidez la dificultad del problema de la evaluación.

## Significado referencial de la probabilidad condicional

El primer paso en la construcción del cuestionario es elaborar un procedimiento para identificar, mediante una definición semántica precisa, el constructo (Martínez Arias, 1995). Para dotar de una mayor objetividad a esta definición, en el caso de la probabilidad condicional, se llevó a cabo un análisis de contenido de una muestra de libros de texto que utilizan los alumnos de Psicología en las asignaturas de análisis de datos. Además se tuvieron en cuenta los errores y dificultades señalados en las investigaciones previas sobre la probabilidad condicional. Con todo ello se delimita el *significado referencial* en el estudio. Hacemos notar que este es un significado parcial, puesto que el objeto probabilidad condicional tiene un significado más completo, si se tienen en cuenta los elementos aportados a dicho significado desde la matemática (por ejemplo, en el estudio que analizamos no se trata el concepto de intercambiabilidad), la historia (sucesivas concepciones históricas de la probabilidad condicional y los campos de problemas que las originaron), psicología y didáctica. Todo ello constituiría el *significado holístico o global* del concepto.

### Investigaciones previas

En primer lugar se analizaron las principales investigaciones relacionadas con la comprensión de las ideas de probabilidad condicional e independencia, tanto en el campo de la Psicología, como en el de la Educación, recopilando, además, los ítems usados en las mismas, que serían la base posterior de la construcción de un banco de ítems. Podemos clasificar las investigaciones encontradas en los apartados siguientes:

- *Comprensión intuitiva de la probabilidad condicional y sus relaciones con la probabilidad simple y dependencia* (Maury, 1985, 1986; Kelly y Zwiers, 1986; Totohasina, 1992; Sánchez, 1996).

- *Condicionamiento y causación:* La existencia de una relación condicional indica que una relación causal es posible, pero no segura. Desde el punto de vista psicológico, la persona que evalúa una probabilidad condicional percibe en forma diferente las relaciones causales y diagnósticas (Tversky y Kahneman, 1982a). La relación de causalidad también se asocia, a menudo, con la secuencia temporal (Falk, 1986; Gras y Totohasina, 1995).
- *Intercambio de sucesos en la probabilidad condicional* (Eddy, 1982; Falk, 1986; Batanero y cols., 1996).
- *Confusión de probabilidad condicional y conjunta* (Tversky y Kahneman, 1982b; Einhorn y Hogarth, 1986; Pollatsek, Well, Konold y Hardiman, 1987; Ojeda, 1995).
- *Situaciones sincrónicas* (que ocurren simultáneamente) *y diacrónicas* (son consecutivas en el tiempo) (Gras y Totohasina, 1995; Ojeda, 1995).
- *Razonamiento bayesiano* (Bar-Hillel, 1983; Totohasina, 1992; Teigen, Brun & Frydenlund, 1999- )
- *Influencia del formato y los datos* (Pollatesk y cols.,1987; Fiedler, 1988; Gigerenzer, 1994)
- *Enseñanza de la probabilidad condicional* (Sedlmeier, 1999; Martignon & Wassner, 2002).

Esta revisión permitió comprobar que las investigaciones se habían centrado en puntos aislados de la comprensión del concepto y sugirió la necesidad de construir un cuestionario comprensivo. Por otro lado, se enriquece el significado puramente matemático del concepto, al tener en cuenta aspectos psicológicos involucrados, tales como la falacia de la conjunción (Tversky y Kahneman, 1982b) o la falacia del eje temporal (Falk, 1986).

## Análisis de contenido de libros de texto de estadística para psicólogos

El estudio de los libros de texto es una forma – limitada – de acercarse al significado institucional de la probabilidad condicional en la institución "análisis de datos en psicología", es decir en los cursos universitarios de análisis de datos para este tipo de estudiantes. El análisis de contenido se basa en la idea de que las unidades del texto pueden

clasificarse en un número reducido de categorías (Weber, 1985). Sirve para efectuar inferencias mediante la identificación sistemática y objetiva de las características específicas de un texto (Ghiglione y Matalón, 1989).

El procedimiento seguido consistió en elaborar un listado con las 31 universidades españolas en las que se imparte la licenciatura de Psicología y solicitar a los directores de los correspondientes departamentos el programa y bibliografía recomendada. Se recibieron repuestas de 23 de estas universidades. De un total de 79 libros diferentes recomendados de análisis de datos 20 eran citados por 4 o más universidades. Trece de ellos incluían el tema de probabilidad condicional y fueron analizados. Además se incluyeron otros 5 libros de orientación bayesiana.

El análisis realizado de los libros y las investigaciones previas sirvió para elaborar la tabla de especificaciones de nuestro cuestionario, que se presenta en la Tabla 1.

Tabla 1 – Especificaciones del contenido del cuestionario

| | Contenido |
|---|---|
| Conocimiento conceptual | 1. Definición de la probabilidad condicional |
| | 2. Reconocer que la probabilidad de P(A) >0 para poder definir P(B/A) |
| | 3. Reconocer que una probabilidad condicional cumple los axiomas. |
| | 4. Reconocer que la probabilidad condicional supone una restricción del espacio muestral |
| | 5. Distinguir probabilidad condicional con inversa |
| | 6. Distinguir probabilidad conjunta, condicional y simple |
| | 7. Probabilidad conjunta menor que probabilidad simple |
| | 8. Distinguir sucesos independientes, dependientes y mutuamente excluyentes |
| Conocimiento procedimental | 9. Calcular una probabilidad condicional dentro de un único experimento |
| | 10. Resolver correctamente problemas de probabilidad condicional en un contexto de muestreo con reposición |
| | 11. Resolver correctamente problemas de probabilidad condicional en un contexto de muestreo sin reposición |
| | 12. Resolver correctamente problemas de probabilidad condicional a partir de probabilidades conjuntas y simples |
| | 13. Resolver correctamente problemas condicionales cuando se invierte el eje de tiempo |
| | 14. Distinguir situación condicional, causal y diagnóstica |
| | 15. Resolver correctamente problemas en situaciones diacrónicas |
| | 16. Resolver correctamente problemas en situaciones sincrónicas |
| | 17. Resolver correctamente problemas de probabilidad compuesta haciendo uso de la regla del producto en caso de sucesos independientes |
| | 18. Resolver correctamente problemas de probabilidad compuesta haciendo uso de la regla del producto en caso de sucesos dependientes |
| | 19. Aplicar correctamente el cálculo de la probabilidad condicional en situaciones de sucesos múltiples (regla de la probabilidad total) |
| | 20. Aplicar correctamente el cálculo de la probabilidad condicional en situaciones de probabilidad inversa (regla de Bayes) |

En esta especificación estamos considerando los conceptos y propiedades (conocimiento conceptual) así como los problemas y algoritmos (conocimiento procedimental). No hemos prestado atención especial al lenguaje o a los argumentos, aunque quedan implícitamente evaluados en los ítems de respuesta abierta, que se podrían reanalizar para estudiar la comprensión de este tipo de elementos.

## Proceso de selección de ítems

Un ítem de un cuestionario es una unidad de medida que consta de un estímulo y una forma prescriptiva de respuesta y su fin es inferir la capacidad del examinado en un cierto constructo (habilidad, rasgo, etc.), proporcionando datos cuantificables sobre la persona que lo completa (Osterlind, 1989). Desde el punto de vista del enfoque onto- semiótico, y aún teniendo en cuenta la naturaleza esencialmente compleja del significado de los objetos matemáticos, al analizar las actuaciones de los alumnos, interesa con frecuencia fijar la atención en procesos interpretativos específicos y en las dificultades inherentes a los mismos, por lo que en la respuesta a cada ítem conlleva *significados parciales* de la probabilidad condicional (Godino y Batanero, 2003).

Al considerar el número total de ítems, se ha de cubrir adecuadamente el contenido y asegurar una fiabilidad satisfactoria (Millman y Greene, 1989), teniendo en cuenta la restricción de la posible longitud total del test. En el cuestionario sobre probabilidad condicional se usaron ítems de opciones múltiples y de respuesta abierta. Se comenzó con un conjunto inicial de unos 50 ítems (2-3 por cada especificación de contenido). Una vez concluida la planificación del cuestionario, se seleccionaron los ítems que constituirían el cuestionario piloto, siguiendo los dos sistemas sugeridos en Osterlind (1989):

- El análisis a partir de un juicio requiere pedir a una serie de expertos que valoren los ítems particulares, de acuerdo con algunos criterios.
- La valoración numérica requiere que los ítems se administren a una muestra de sujetos y se basa en el estudio de una serie de indicadores estadísticos de los mismos.

## Juicio de expertos

Millman y Greene (1989) indican que el "experto" lo define el propósito del instrumento y que el grupo elegido de expertos ha de representar una diversidad relevante de capacidades y puntos de vista. En nuestro caso, fueron seleccionados en base a su conocimiento de probabilidades y más particularmente de probabilidad condicional, así como a la experiencia de investigación sobre el tema. Participaron nueve investigadores en didáctica de la estadística, tanto españoles como iberoamericanos. Nuestro objetivo era doble:

- Establecer un consenso sobre la tabla de especificaciones del instrumento, decidiendo cuales especificaciones del contenido eran relevantes para los propósitos del instrumento. De este modo se reforzarían los resultados obtenidos del análisis de contenido de los libros de texto.
- Establecer un consenso de opiniones de los expertos sobre cómo cada ítem particular se ajusta bien para evaluar el contenido específico para el cuál ha sido diseñado que sirviesen como base para elegir los ítems definitivos.

**Contenido 1: Definición de la probabilidad condicional.**

Ítem 1. Explica con tus propias palabras la diferencia entre una probabilidad simple y una probabilidad condicional.

Ítem 2. ¿Qué quiere decir la expresión "la probabilidad condicional de *A* dado *B* es 1/4"?
a) En la cuarta parte de los experimentos obtenemos *A* y *B* simultáneamente
**b) *A* ocurre la cuarta parte de las veces en que ocurre *B***
c) *B* ocurre la cuarta parte de las veces en que ocurre *A*
d) *A* o *B* ocurren la cuarta parte de las veces

| | 1: Nada | 2 | 3 | 4 | 5: Mucho |
|---|---|---|---|---|---|
| El contenido "Definición de la probabilidad condicional" es relevante | | | | | |
| El ítem 1 es adecuado para este contenido | | | | | |
| El ítem 2 es adecuado para este contenido | | | | | |

Figura 1 – Ejemplo de contenido y estructura del cuestionario a expertos

Se proporcionó a cada uno de estos expertos un cuestionario en que se les pedía, para cada unidad de contenido y para cada uno de los ítems asociados a la misma su grado de acuerdo (en una escala 1 a 5) sobre su adecuación a los fines de la evaluación. Un ejemplo de la estructura y contenido del cuestionario a expertos se muestra en la Figura 1.

## Pruebas de ítems

Por otro lado, todos los ítems del banco inicial fueron probados con muestras de estudiantes similares a aquellos a quienes iba destinado el cuestionario. Los ítems fueron divididos en cuatro cuestionarios, con objeto de que el número total a completar en una sesión por un mismo grupo de alumnos no fuera excesivamente largo, de modo que tuviesen tiempo suficiente para responder. Participaron en las pruebas dos grupos diferentes de alumnos, de primer año en la Licenciatura de Psicología (en total 157 alumnos). Los porcentajes de estudiantes con diferentes tipos de bachillerato fueron los siguientes: ciencias 27,6%, letras 72,4%. Se trató de sujetos voluntarios, como es frecuente en investigaciones en ciencias sociales.

Una vez finalizadas las fases anteriores, seleccionamos los ítems que compondrían el cuestionario piloto, con el siguiente proceso:

- Primeramente se desecharon aquellas especificaciones de contenido en las que el grado de acuerdo sobre su relevancia (en escala 1-5) no fuese suficientemente elevado y consensuado (se rechazaron todos los ítems cuya puntuación media fuese 4 o inferior o tuviesen una desviación típica mayor que el resto).
- Para aquellos contenidos con alto grado de acuerdo respecto a su relevancia se examinaron los resultados de los ítems, desechándose aquellos que no fuesen altamente valorados por los expertos (media 4 o menor).
- De entre los ítems bien valorados se eligió para cada contenido el que hubiese presentado menor dificultad en las pruebas pre-piloto (mayor porcentaje de respuestas correctas). Algunos ítems también se desecharon porque algún distractor fue elegido mayoritariamente (mientras que lo ideal es que las respuestas se distribuyan homogéneamente entre distractores) o porque el contexto muy marcado distraía la atención del estudiante y forzaba la respuesta.

## Pruebas del cuestionario piloto

Finalizado el instrumento piloto, se deben realizar pruebas del mismo, con objeto de obtener información empírica sobre las características de esta primera versión del instrumento y comprobar que sea útil para los

objetivos pretendidos. Al mismo tiempo se analizan sus limitaciones, con objeto de identificar aquellos puntos en que se puede mejorar.

Una primera prueba piloto del cuestionario RPC se llevó a cabo con dos muestras de estudiantes universitarios. La primera de ella estuvo formada por 37 alumnos que cursaban 5º año de la Licenciatura de Matemáticas en la especialidad de Metodología. Se eligió a estos alumnos, debido a su alta preparación y porque el disponer de este grupo de alumnos nos permitiría comparar si los errores más frecuentes en alumnos de psicología se repetían en alumnos con alta preparación matemática.

## Validación del cuestionario

Finalizada esta prueba y revisado nuevamente el cuestionario, se procedió a la validación definitiva con una muestra de 414 estudiantes de psicología de cuatro universidades diferentes, después de haber estudiado el tema. Otra muestra de 170 estudiantes tomó el cuestionario antes de la enseñanza, para servir en el estudio de discriminación. La validación consta de varias fases y puede hacerse desde diversos modelos teóricos. Nosotros utilizamos el modelo lineal clásico de la teoría de tests (Muñiz, 1994; Barbero, 2003).

### Análisis de items

En primer lugar, se analizan las respuestas obtenidas en cada uno de los ítems. En los ítems de opciones múltiples, como el ítem 3, estudiamos las frecuencias y proporciones de respuestas (ver Tabla 2)

Ítem 3. Una caja tiene cuatro focos, de los cuales dos son defectuosos. Se sacan dos al azar, uno tras otro sin reemplazamiento. Si el primer foco fue defectuoso, entonces:

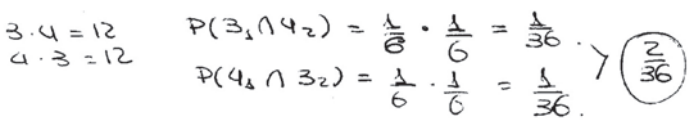
a) Es más probable que el segundo sea defectuoso
b) Es más probable que el segundo no sea defectuoso
c) La probabilidad de que el segundo sea defectuoso es igual a la probabilidad de que no lo sea

Tabla 2 – Análisis de respuestas al ítem 3

| Respuesta | Frecuencia (n=414) | Porcentaje |
|---|---|---|
| a) | 4 | 1,0 |
| **b)** | **367** | **88,6** |
| c) | 42 | 10,1 |
| Blanco | 1 | 0,2 |

En los ítems de respuesta abierta se ha puntuado de acuerdo a la mayor o menor completitud. Por ejemplo, en el ítem 4 se ha seguido el siguiente criterio: 0) No se responde, o se responde incorrectamente; 1) Identifica los casos favorables y posibles, plantea el problema pero comete algún error y 3) Resuelve el problema correctamente, como en el caso siguiente: *"{(2,6), (3,4), (6,2), (4,3)}; 2/4=1/2= 0,5".*

Un ejemplo de respuesta puntuada como 3 es el siguiente en que el alumno ha identificado los casos favorables, percibe la independencia de los sucesos y la importancia del orden pero no identifica el problema como de probabilidad condicional, sino lo confunde con otro de probabilidad simple. Por ello, encuentra la probabilidad de cada caso, aplicando la regla del producto (correctamente) y suma las probabilidades (aplicando el axioma de la unión). El alumno muestra un razonamiento probabilístico bueno, pero aparece un *conflicto* al no interpretar correctamente el enunciado del problema. En la Tabla 3 se muestran los resultados en este ítem




**Ítem 4.** Hemos lanzado dos dados y sabemos que el producto de los dos números obtenidos ha sido 12 ¿Cuál es la probabilidad de que ninguno de los dos números sea un 6?

Tabla 3 – Resultados en el ítem 5 ($n$ = 414)

| | Frecuencia | Porcentaje | Porcentaje acumulado |
|---|---|---|---|
| 0 | 186 | 44,9 | 44,9 |
| 1 | 86 | 20,8 | 65,7 |
| 2 | 142 | 34,3 | 100,0 |

## Índice de dificultad

El índice de dificultad valora la dificultad que entraña la resolución del ítem para la población a la que va dirigido el test. Se define en la forma siguiente (Muñiz, 1994)

$$ID = \text{Aciertos} / \text{Numero de respuestas}$$

El valor del índice es un número entre 0 y 1. 0 es indicativo de alta dificultad y (nadie ha acertado) y 1 de máxima facilidad. Este índice, permite ordenar los ítems en función de su dificultad, y puede ayudar a la selección de ítems en función de los objetivos y facilitan la discriminación de personas. Los índices de dificultad media son los que discriminan mejor. Puesto que el índice de dificultad es una proporción (la proporción de alumnos que superan el ítem) calculamos adicionalmente los intervalos de confianza de dicha proporción con la fórmula habitual (Ver tabla 4).

## Índice de discriminación

Los índices de discriminación tratan de valorar si un ítem discrimina entre los sujetos que tienen un valor alto o bajo en el constructo que se trata de medir (en este caso, que conocen o no la probabilidad condicional). Se pueden calcular por dos sistemas diferentes:

1. Como correlación entre el ítem y la puntuación total del cuestionario. Los sujetos habilidosos tendrán una alta puntuación en el test y responderán correctamente al ítem; los sujetos no habilidosos tendrán una baja puntuación en el test y no responderán bien al ítem. Los valores del índice de discriminación varían entre -1 y 1.
2. Mediante el estudio de la diferencia en proporción de aciertos al ítem en los estudiantes que difieren en el atributo. En nuestro caso entre estudiantes del mismo curso antes y después de haber estudiado el tema (aproximadamente 200 estudiantes en cada grupo). Todos los índices fueron estadísticamente significativos, aunque de valor moderado, pues algunos ítems se resuelven intuitivamente incluso antes de la enseñanza.

Tabla 4 – Índices de dificultad y discriminación de algunos ítems del cuestionario

| | n = 414 | | | | n = 393 |
|---|---|---|---|---|---|
| | Índice dificultad | I. confianza 95% | | Índice discriminación Correlación /total | Índice discriminación Diferencia medias |
| I1 | 0,428 | 0,36 | 0,50 | 0,451** | 0,30** |
| I3 | 0,886 | 0,84 | 0,83 | 0,264** | 0,18** |
| I4 | 0,343 | 0,28 | 0,31 | 0,378** | 0,13** |

** Significativo a nivel 0,01

## Fiabilidad

Para cualquier instrumento, es necesario conocer la precisión de las medidas que nos proporciona, ya que esta precisión se puede usar para extender los resultados de la muestra particular a una población más general La medida siempre produce un cierto error aleatorio, pero dos medidas del mismo fenómeno sobre un mismo individuo suelen ser consistentes. Un instrumento de medida se considerará *fiable* si las medidas que se obtienen a partir de él no contienen errores o los errores son suficientemente pequeños. La fiabilidad es esta tendencia a la consistencia y se define como correlación entre las puntuaciones verdadera y observada (Martínez Arias, 1995).

El *coeficiente de fiabilidad* es un indicador de la fiabilidad teórica de las puntuaciones observadas, en el sentido de proporcionar un valor numérico del *grado de confianza* que podíamos tener en dichas puntuaciones como estimadores de las puntuaciones verdaderas de los sujetos (Thorndike, 1989). Se define como la correlación entre las puntuaciones obtenidas por el sujeto en un test cuando se le pasa dos veces sucesivas. Muñiz (1994) indica que, si no hubiese errores de medida, las puntuaciones coincidirían y la correlación sería perfecta, por lo que este coeficiente sería igual a 1.

Este coeficiente de fiabilidad es un valor teórico que debe ser estimado por algún procedimiento empírico, a través de las respuestas de un grupo de sujetos a un conjunto de ítems (Carmines y Zeller, 1979). Entre los diversos procedimientos para el cálculo del estimador del coeficiente de fiabilidad (Díaz, Batanero y Cobo, 2003) se han usado en este estudio los que se describen a continuación.

**Test-retest**

En este método se administra el mismo test dos veces a las mismas personas y se calcula el coeficiente de correlación entre las puntuaciones obtenidas en las dos ocasiones. Este procedimiento intenta medir el porcentaje de variabilidad debido a las fuentes que contribuyen a que un sujeto tenga diferente puntuación en aplicaciones repetidas de la "misma prueba": temporales, ambientales, estado, sentimientos... y da una medida de la *estabilidad* del rasgo durante el periodo de tiempo dado. El intervalo de tiempo que se deja entre las dos administraciones del test tiene que ser suficiente para que no recuerden la tarea pero no demasiado amplio para que no se den cambios en los sujetos (aprendizaje, maduración...).

En nuestro estudio se estimó la fiabilidad test-retest al pasar dos veces el cuestionario a una muestra de 106 estudiantes, obteniéndose un valor de 0,871 para el coeficiente de estabilidad.

**Covariación entre los ítems del test**

Su cálculo se basa en el análisis relativo de la varianza de la puntuación total del cuestionario y de las varianzas de los ítems particulares. También es una cota inferior de la que se obtendría por el método de la prueba repetida si se comparase el test dado y otro cualquiera paralelo de igual cantidad de ítems (Carmines y Zeller, 1979). Se estima mediante el coeficiente Alfa de Cronbach: refleja el grado en el que covarían los ítems que constituyen el test.

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} s_j^2}{S_n^2}\right)$$

Siendo $n$ el número de ítems, $S^2$ la varianza de la puntuación total, $s_j^2$ las varianzas de cada ítem, que en el caso de ser dicotómicos son iguales al producto $p_j q_j$, siendo $p_j$ el índice de dificultad del ítem. En el cuestionario RPC se obtuvo un coeficiente Alfa = 0,79 sobre la muestra de 414 estudiantes.

**Coeficientes basados en el análisis factorial**

Cuando se sospecha que el cuestionario es multidimensional se pueden calcular los siguientes coeficientes, basados en los resultados del análisis factorial (Barbero, 2003):

1. El coeficiente Theta que tiene en cuenta los pesos de los ítems en el primer factor. Se trata del coeficiente $\alpha$ cuando las varianzas verdaderas (suma de las covarianzas) y total se calculan a partir de las puntuaciones factoriales derivadas del primer factor común, antes de las rotaciones (Morales, 1988). En el caso del cuestionario RPC fue bastante alto, debido a que el primer factor explicó mucho mayor porcentaje de varianza que los siguientes. Viene dado por la siguiente expresión, donde $\lambda_1$ es el primer autovalor en el análisis factorial.

$$\theta = \frac{n}{n-1}\left(1-\frac{1}{\lambda_1}\right) = 0{,}82$$

2. El coeficiente Omega, que en el cuestionario RPC dio todavía un valor superior. Como es de esperar se cumplió la relación: $\alpha < \theta < \Omega$. En la expresión que sigue, $h_j$ son las comunalidades de las variables (correlación múltiple de cada una con las demás) y $r_{jh}$ las correlaciones entre pares de variables.

$$\Omega = 1 - \frac{n - \sum h_j^2}{n + 2\sum r_{jh}} = 0{,}896$$

## Validez

La validez es un criterio de calidad relacionado con la adecuación de las puntuaciones del test para el objetivo que suscitó su aplicación. La validez es un concepto unitario, aunque hay diferentes formas de recoger evidencias de la misma (validez de constructo, contenido, criterio...) (Martínez Arias, 1995). Además, la validez suele entenderse con relación al uso que se dé al cuestionario y la interpretación de sus puntuaciones; no validamos sólo el cuestionario, sino también las inferencias o interpretaciones que hacemos a partir de las puntuaciones obtenidas con el mismo (Messick, 1989, 1995).

En lo que sigue se describen las formas más usuales (validez de contenido, criterio y constructo); indicando la metodología de su análisis para el caso del cuestionario RPC.

### Estudios de validación de contenido o dominio

La *validez de contenido o dominio* es el grado en que los ítems de un cuestionario representan la totalidad de los contenidos de un dominio

previamente delimitado y definido (Carmines y Zeller, 1979). El proceso de validación de contenido es eminentemente lógico, si bien pueden utilizarse algunos procedimientos estadísticos para resumir la información obtenida a partir de jueces expertos en el tema que valoran la congruencia entre los diversos ítems y los diversos objetivos.

En el caso del cuestionario RPC la validez de contenido se trata de alcanzar realizando una planificación cuidadosa de los ítems y de cómo estos ítems pueden contribuir a la medida del constructo subyacente. Una vez elegidos los ítems que formarían el cuestionario (18 ítems con 22 subítems) se analizó su contenido primario y secundario. Por ejemplo, el ítem 5 fue diseñado para evaluar el contenido primario *"distinguir sucesos dependientes, independientes y mutuamente excluyentes".* Pero al analizarlo aparecen otros contenidos secundarios, ya que el alumno debe también "*resolver correctamente problemas de probabilidad compuesta haciendo uso de la regla del producto en experimentos independientes"* para reconocer que el distractor d) es incorrecto.

Ítem 5. Se extrae una carta al azar de una baraja americana: sea *A* el suceso "se extrae un trébol" y *B* el suceso "se extrae una reina" ¿Los sucesos *A* y *B* son independientes?

a) Sí, en todos los casos.

b) No son independientes porque en la baraja hay una reina de tréboles.

c) Sólo si sacamos primero una carta para ver si es reina y se vuelve a colocar en la baraja y luego sacamos una segunda carta y para mirar si es trébol.

d) No, porque P(reina de trébol)= P(reina) x P(trébol)

Los otros distractores tratan de detectar diferentes errores descritos en la literatura de investigación sobre probabilidad condicional:

- En el distractor b) se trata de detectar la posible confusión entre sucesos excluyentes y sucesos independientes.
- En el distractor c) se trata de detectar la posible creencia errónea que sólo pueden ser independientes los sucesos de experimentos diacrónicos.
- En el distractor d) se presenta la definición correcta de la regla del producto, junto con la afirmación incorrecta que esta regla

no se cumple en el caso de sucesos independientes. Precisamente en este caso se cumple porque los sucesos son independientes y tratamos de ver si el alumno detecta el error en la afirmación presentada en el distractor.

Este análisis para cada uno de los ítems permitió comprobar que todas las unidades de contenido especificadas en la tabla 1 quedaban cubiertas al menos dos veces en el cuestionario. En esta tabla representamos, por tanto el significado de la probabilidad condicional *evaluado* por el cuestionario, que es más restringido que el significado holístico y en ciertos aspectos también que el socio-profesional en la institución de enseñanza – evaluado en el análisis de textos. Pero incluye elementos psicológicos, tomados de las investigaciones previas, que no se tienen en cuenta en la enseñanza y por tanto son nuevos respecto al significado institucional socio-profesional.

También se usó una metodología adecuada para revisar la congruencia entre los ítems y las especificaciones (Osterlind, 1989) mediante el juicio de expertos, definiendo previamente el universo de observaciones admisibles, identificando expertos en el campo, pidiendo a los expertos que emparejen ítems con objetivos y resumiendo la información numéricamente.

### Estudios de validación de criterio

Este tipo de validez se justifica si se encuentra una correspondencia entre las mediciones de un cuestionario y otras variables externas referidas al mismo grupo de sujetos. En el contexto de las normas APA, AERA y NCME (1999) se han propuesto dos tipos de validez de criterio empírico externo: a) *Validez concurrente*: cuando se compara el cuestionario con otro método que sepamos que es un buen estándar para medir la misma variable; b) *Validez predictiva* sería la capacidad de un instrumento de predecir comportamientos, actitudes o eventos futuros.

En nuestro ejemplo, el criterio utilizado fue el haber estudiado específicamente probabilidad condicional o no en el curso en que se pasó el cuestionario RPC (criterio dicotómico). Se compararon las puntuaciones totales y en cada ítem en dos grupos de estudiantes (un grupo antes de la enseñanza y otro después). La prueba de diferencia de puntuación media total en el cuestionario (mayor puntuación media en los alumnos que

habían estudiado la probabilidad condicional) y el análisis discriminante (88% de estudiantes correctamente clasificados en su grupo) se usaron como indicadores de validez respecto a este criterio.

## Estudios de validación de constructo

Es la característica más importante de un cuestionario cuando se elabora con fines científicos. Pretende investigar qué propiedades mide, determinando para ello el grado en que ciertos constructos y sus relaciones hipotetizadas y formalizadas a través de un modelo matemático o estadístico se corresponden con los datos empíricos aportados por las respuestas al cuestionario. El procedimiento para valorarla consta de las siguientes fases:

1. Formulación de hipótesis que se pretenden comprobar, lo que se lleva a cabo en la definición semántica y sintáctica de la variable. En el ejemplo del cuestionario RPC, una de las hipótesis formuladas fue que el constructo "comprensión de la probabilidad condicional" era multidimensional y que los sesgos de razonamiento descritos en la literatura previa serían independientes de la capacidad matemática para resolver problemas de probabilidad condicional.
2. Obtención de los datos empíricos y analizar se verifican o no las hipótesis planteadas. En el caso de que así sea, queda confirmado mediante una investigación que el test mide el constructo de interés

En la validación de constructo del cuestionario RPC se utilizó el análisis factorial, que es una técnica estadística multivariante que sirve para estudiar las dimensiones que subyacen a las relaciones entre varias variables. Del conjunto de ítems (22 variables) se obtuvo un total de 7 factores diferenciados; cada factor viene definido por un conjunto de las variables originales, que tienen correlaciones fuertes en el factor y los factores son independientes entre si (Tabla 5).

Tabla 5 – Resultados del análisis Factorial Matriz de componentes rotados

| | Componente | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| I16.Teorema de Bayes | 0,76 | | | | | | |
| I11. Probabilidad total | 0,76 | | | | | | |
| I15. Probabilidad compuesta dependencia | 0,75 | | | | | | |
| I13. Probabilidad compuesta independencia | 0,67 | | | | | | |
| I12. Probabilidad condicional; con reemplazamiento | 0,43 | | 0,42 | | | | |
| I6c. Probabilidad condicional (tabla) | | 0,79 | | | | | |
| I6d. Probabilidad condicional (tabla) | | 0,77 | | | | | |
| I6b. Probabilidad compuesta (tabla) | 0,32 | 0,61 | | | | | |
| I6a. Probabilidad simple (tabla) | | 0,61 | | | | | |
| I8. Probabilidad condicional; experimento simple | | | 0,67 | | | | |
| d I1. Definición | | | 0,59 | | | | |
| I2. Espacio muestral | 0,40 | | 0,45 | | | | |
| I17b. Falacia del eje temporal, diacrócnico | | | | 0,71 | | | |
| I14. Falacia del eje temporal | | | | 0,70 | | | |
| I7. Probabilidad condicional (de simple y compuesta) | | | | | 0,66 | | |
| I9. Falacia conjunción | | | | | 0,62 | | |
| I5. Probabilidad condicional sin reemplazamiento | | | 0,39 | | 0,44 | | |
| I17a. Probabilidad condicional sin reemplazamiento | | | | | | 0,66 | |
| I10. Condicional transpuesta | | | | | | -0,65 | |
| I4. Independencia | | | | | | | 0,68 |
| I3. Teorema Bayes (falacia de tasas base) | 0,34 | | | | | | 0,48 |
| I18. Probabilidad condicional, sin reemplazamiento | | | | 0,353 | | | -0,45 |

Los tres primeros factores (que explicaron la mayor parte de la varianza total en las puntuaciones) correspondieron a componentes matemáticos: resolución de problemas de probabilidad condicional (factor 1), definición y restricción del espacio muestral (factor 2), lectura de tablas dobles y cálculo de probabilidades simples, compuestas y condicionales (factor 3). El resto de los factores, mide cada uno un sesgo de razonamiento (confusión de una probabilidad condicional y su transpuesta, confusión de sucesos independientes y mutuamente excluyentes, etc.). Se confirma así la hipótesis previa de que estos sesgos están incorrelacionados con el conocimiento matemático, representado por los primeros factores y por tanto se proporciona evidencias de validez de constructo al cuestionario.

## Reflexiones desde el enfoque de la TFS

### Instituciones implicadas en el proceso de evaluación

El proceso de construcción y prueba de un cuestionario de evaluación descrito nos permite reflexionar sobre la complejidad de la tarea evaluadora y la diversidad de significados involucrados respecto a un mismo objeto matemático. En la descripción aparece claramente la

dialéctica *institucional-personal* (Godino, 2003), puesto que la adecuación de los significados personales de los alumnos sólo puede llevarse a cabo desde una institución de referencia. Ahora bien, dicha faceta aparece en relación con diversas instituciones que intervienen en diferentes fases, que representamos esquemáticamente en la Tabla 6.

Tabla 6 – Significados y procesos de muestreo en el estudio realizado

| Instituciones/ sujetos involucrados | Significado de interés | Instrumento de evaluación | Proceso de muestreo | Significado evaluado |
|---|---|---|---|---|
| Didáctica / Matemática | Holístico | | | |
| Enseñanza (Estadística en Psicología) | Institucional (socio-profesional) | Análisis del contenido de libros de texto recomendados | 23 entre 31 Universidades; 18 entre 60 libros | Referencia |
| Investigación en Educación Estadística | Objetivamente evaluado | Juicio de expertos Pruebas pre-piloto de ítems | 9 investigadores muestras de estudiantes | Evaluado |
| Un estudiante particular | Personal (Un estudiante) | Cuestionario | Unidades de contenido Ítems | Declarado por una persona |
| Estudiantes de psicología en general | Personal (Un tipo de estudiantes) | Cuestionario | Muestra de estudiantes | Declarado en una muestra |

Nuestro interés es evaluar el significado personal que los estudiantes de psicología alcanzan sobre la probabilidad condicional tras un proceso "estándar" de instrucción, pero esta evaluación ha de tener como pauta el significado institucional en la correspondiente institución de enseñanza (estadística en psicología). No olvidamos que este significado es sólo una parte del sistema de prácticas más complejo y global – el significado global u holístico del concepto, del que aquí se aborda tan sólo una parte y que en este trabajo se considera como dado, no es objeto de investigación.

Por otro lado, el instrumento es elaborado desde una institución diferente, que podemos denominar como *investigación en educación estadística*; en esta institución se comparte un cierto significado del objeto probabilidad condicional, sobre qué sería una evaluación objetiva y los criterios para construir un instrumento aceptable (por ejemplo referidos a fiabilidad, validez, etc.). El investigador es un sujeto de dicha institución; por tanto puede introducir elementos subjetivos tanto en la elaboración del instrumento, como en la interpretación de resultados, elección de muestra de estudiantes, etc. Es aquí donde se usa el juicio de expertos – a su vez una práctica dentro de dicha institución- como control de la objetividad de proceso. Tanto esta práctica como otras – la serie de análisis

psicométricos efectuados- tienen como finalidad evaluar o investigar el significado personal de una forma objetiva.

## Procesos de muestreo implicados

Para poder comprender la diferencia entre los significados que son objetos del estudio y lo que realmente es posible determinar en la investigación, es importante también reflexionar sobre todo el proceso de inferencia llevado a cabo, desde la definición operacional de la variable hasta la interpretación dada a las respuestas de los estudiantes. Ello nos permite diferenciar entre lo que queremos evaluar y lo que podemos evaluar, así como hasta qué punto podemos generalizar los resultados de un estudio de evaluación y definir en consecuencia algunos criterios que permitan mejorar la generalizabilidad.

Usualmente en la investigación didáctica somos conscientes de que los alumnos participantes en el estudio constituyen una muestra de una población real o potencial, a la que queremos extender nuestras conclusiones. Sin embargo, muestreamos también las tareas, contextos, etc. En concreto en el estudio analizado podemos identificar los siguientes procesos de muestreo:

- El investigador está interesado en determinar el significado institucional para poder precisar su variable: qué se entiende como conocimiento de la probabilidad condicional en una cierta institución de enseñanza. Pero no es posible acceder a las clases que se dan en esta institución (que además podrían potencialmente variar de un curso a otro, incluso para un mismo profesor). Una forma de acercarse al significado de interés es el análisis de libros de texto recomendados, que ni siquiera es completo, pues hay un muestreo de Universidades y de libros. Es innegable que el resultado del estudio es un significado diferente, que sería el significado de referencia del proceso de evaluación, en cuanto en base a él se organiza la construcción del cuestionario e interpretación de las respuestas de los alumnos.
- Una vez fijado el *significado de referencia* hay infinitos posibles instrumentos de evaluación; incluso infinitos posibles cuestionarios. Podemos variar el tipo y número de ítems, el nivel de formalización de los mismos, su dificultad, su contexto. El investigador trata de construir un instrumento lo más objetivo

posible, y para ello recopila diferentes ítems, tomados de investigaciones previas en las que han sido evaluados y han dado resultados probados, para cada una de las unidades de contenido. Pero un investigador puede involuntariamente introducir elementos subjetivos en la elección de los ítems o tareas. El recurso a juicio de expertos trata de crear un *significado compartido* de lo que sería un instrumento de evaluación. Las pruebas de los ítems con estudiantes tratan de asegurar la legibilidad y la idoneidad cognitiva de los mismos. El cuestionario resultante define un significado nuevo, sería el *significado evaluado*, diferente de los anteriores.

- Finalmente el instrumento se prueba con una muestra de estudiantes. La respuesta que cada uno de ellos proporciona a cada ítem no es la única que puede dar. Dependiendo del interés, cansancio, concentración y otros factores, sus respuestas reflejan una parte de lo que el estudiante realmente conoce. Sería el *significado declarado* que es el finalmente accesible al investigador.

**Idoneidad de un cuestionario de evaluación**

Pero el interés del estudio no se limita a este significado declarado. El investigador estaría interesado en el significado personal de los alumnos respecto a las tareas propuestas (las respuestas dadas se suponen una muestra representativa de las que darían los mismos estudiantes en la misma prueba en otras ocasiones).

El estudio significativo de los objetos matemáticos debe poner en juego una muestra representativa de las prácticas que constituyen el significado sistémico de los mismos en el seno de un contexto institucional dado (Godino, 1999). Más aún, si las tareas son suficientemente representativas (para evaluar las unidades de contenido definidas), podríamos hacer una inferencia sobre lo que cada alumno de la muestra sería capaz de hacer y decir en otras tareas relacionadas con el concepto.

Si las unidades del contenido están bien definidas y representan el concepto de probabilidad condicional, entonces podríamos acercarnos al significado personal de los alumnos de la muestra sobre la probabilidad condicional. Finalmente, los alumnos particulares son una muestra (que suponemos representativa) de otros estudiantes de psicología. Mientras

que un profesor se interesa sólo por los alumnos a su cargo, el investigador aspira a obtener conocimiento generalizable sobre las dificultades y capacidades de los estudiantes.

Es claro que la posibilidad de generalizar en cada uno de los pasos descritos depende de la representatividad y la variabilidad de la muestra elegida en cada uno de los procesos de muestreo. Aunque la tarea de conseguir una generalizabilidad completa parece imposible, la investigación didáctica debe aspirar a dar criterios que permitan la construcción adecuada de instrumentos de evaluación o de reinterpretar los criterios clásicos en psicometría. En este sentido, pensamos que podríamos aplicar o extender el concepto de *idoneidad* y sus tipos (Godino, 2003; Godino, Contreras y Fonts, 2006) al caso de la evaluación, en el siguiente sentido:

- La *dificultad* de un ítem o tarea daría una medida de su *idoneidad cognitiva*; es decir del grado de representatividad de los significados evaluados respecto a los significados personales.
- La *discriminación* de un ítem valoraría su *idoneidad evaluadora,* un item puede ser adecuado cognitivamente, pero no diferenciar (por ser demasiado fácil) los alumnos que tienen un mayor o menor conocimiento del concepto. Esta idoneidad podría ser un componente de la idoneidad *instruccional*, en cuanto uno de los objetivos de la instrucción es la función evaluadora.
- La *validez de contenido* de un cuestionario indicaría una idoneidad *epistémica*, o grado de representatividad del instrumento en cuanto al significado objeto de evaluación.
- La *fiabilidad* daría una medida de la estabilidad de la respuesta, es decir sería otro componente de la *idoneidad evaluadora* También incluiríamos aquí la *validez de criterio y constructo* que indicaría la utilidad del instrumento para el fin que fue diseñado.
- La *validez externa* y *generalizabilidad a otros estudiantes,* sugeriría una *idoneidad generalizadora o externa* en cuanto los resultados se generalizarían a otros estudiantes.

Como conclusión señalamos que el marco teórico nos ha permitido analizar y reflexionar sobre un proceso de investigación – incluso realizado desde una perspectiva muy diferente como es la psicométrica- y reinterpretar desde nuestra perspectiva algunas de sus prácticas y conceptos. El enfoque de la TSF a priori no presupone una metodología única de

investigación, sino que puede beneficiarse de múltiples perspectivas a las que a su vez puede enriquecer introduciendo algunos de sus conceptos.

## Reconocimientos

Este trabajo es parte del Proyecto SEJ2004-00789 y Beca FPU: AP2003-5130.

## Referencias

AMERICAN PSYCHOLOGICAL ASSOCIATION, AMERICAN EDUCATIONAL RESEARCH ASSOCIATION Y NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (1999). *Standards for educational and psychological testing*. Washington, DC, American Psychological Association.

BAR-HILLEL, M. (1987). "The base rate fallacy controversy". In: SCHOLZ, R. W. (ed.). *Decision making under uncertainty*. Amsterdam, North Holland.

BARBERO, M. (2003). *Psicometría II. Métodos de elaboración de escalas*. Madrid, UNED.

BATANERO, C.; ESTEPA, A.; GODINO, J. y GREEN, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education*, v. 27, n. 2, pp. 151-169.

CARMINES, E. G. y ZELLER, R. A. (1979). *Reliability and validity assesment*. Sage University Paper.

DÍAZ, C. (2004). *Elaboración de un instrumento de evaluación del razonamiento condicional. Un estudio preliminar.* Trabajo de Investigación Tutelada. Universidad de Granada.

DÍAZ, C.; BATANERO, C. y COBO, B. (2003). Fiabilidad y generalizabilidad. Aplicaciones en evaluación educativa. *Números,* n. 54, pp. 3-21.

DÍAZ, C. y de la Fuente, I. (2006). "Assessing psychology students' difficulties with conditional probability and bayesian reasoning". In: Rossman, A. y Chance, B. (eds.). *Proceedings of ICOTS – 7*. Salvador, International Association for Statistical Education (CD-ROM).

DÍAZ, C. y de la Fuente, I. (2005). Razonamiento sobre probabilidad condicional e implicaciones para la enseñanza de la estadística. *Epsilon,* n. 59, pp. 245-260.

EDDY, D. M. (1982). "Probabilistic reasoning in clinical medicine: Problems and opportunities". In: KAHNEMAN, D.; SLOVIC, P. y TVERSKY (eds.). *Judgement under uncertainty: Heuristics and biases*. New York, Cambridge University Press.

EINHORN, H. J. y HOGART, R. M. (1986). Judging probable cause. *Psychological Bulletin*, n. 99, pp.3-19.

FALK, R. (1986). "Conditional Probabilities: insights and difficulties". In: DAVIDSON, R. y SWIFT, J. (eds.). *Proceedings of the Second International Conference on Teaching Statistics*. Victoria, Canada, International Statistical Institute.

FIEDLER, K. (1988). The dependence of the conjunction fallacy on subtle lingüistic factors. *Psychological Research,* n. 50, pp.123-129.

GHIGLIONE, R. y MATALÓN, B. (1991). *Les enquêtes sociologiques. Théories et practique.* Paris, Armand Colin.

GIGERENZER, G. (1994). "Why the distinction between single-event probabilities and frequencies is important for psychology (and vice-versa)". In: Wright, G. y Ayton, P. (eds.). *Subjective probability*. Chichester, Wiley.

GODINO, J. D. (1999). "Implicaciones metodológicas de un enfoque semiótico-antropológico para la investigación en didáctica de la matemática". In: ORTEGA, T. (ed.). *Actas del III Simposio de la Sociedad Española de Investigación en Educación Matemática*. Universidad de Valladolid.

____ (2002). Un enfoque ontológico y semiótico de la cognición matemática. *Recherches en Didactiques des Mathematiques*, v. 22, n. 2-3, pp. 237-284.

____ (2003). *Teoría de las funciones semióticas. Un enfoque ontológico-semiótico de la cognición e instrucción matemática*. Granada, El autor.

GODINO, J. D. y BATANERO, C. (2003). "Semiotic functions in teaching and learning mathematics". In: ANDERSON, M.; SÁENZ-LUDLOW, A.; ZELLWEGER, S. y CIFARELLI, V. V. (eds.). *Educational perspectives on mathematics as semiosis: From thinking to interpreting to knowing*. New York, LEGAS.

GODINO, J. D.; BATANERO, C. y ROA, R. (2005). An onto-semiotic analysis of combinatorial problems and the solving processes by university students. *Educational Studies in Mathematics,* v. 60, n.1, pp. 3-36.

GODINO, J. D.; CONTRERAS, A. y FONTS, V. (2006). Análisis de procesos de instrucción basado en el enfoque ontológico-semiótico de la cognición matemática. *Recherches en Didáctique des Mathematiques*, v. 26, n.1, pp. 39-88.

GRAS, R. y TOTOHASINA, A. (1995). Chronologie et causalité, conceptions sources d'obstacles épistémologiques à la notion de probabilité conditionnelle. *Recherches en Didactique des Mathématiques,* v. 15, n. 1, pp. 49-95.

KELLY, I. W. y ZWIERS, F. W. (1986). Mutually exclusive and independence: Unravelling basic misconceptions in probability theory. *Teaching Statistics*, n. 8, pp. 96-100.

LEÓN, O. G. y MONTERO, I. (2002). *Métodos de investigación en psicología y educación.* Madrid, McGraw-Hill.

MARTIGNON, L. y WASSNER, C. (2002). "Teaching decision making and statistical thinking with natural frequencies". In: PHILLIPS, B. (ed.). *Proceedings of the Sixth International Conference on Teaching of Statistics.* Ciudad del Cabo, IASE (CD-ROM).

MARTÍNEZ ARIAS, R. (1995). *Psicometría: teoría de los tests psicológicos y educativos.* Madrid, Síntesis.

MAURY, S. (1985). Influence de la question dans una épreuve relative a la notion d'independance. *Educational Studies in Mathematics,* n. 16, pp. 283-301.

____ (1986). *Contribution à l'étude didacique de quelques notions de probabilité et de combinatoire à travers la résolution de problémes.* Tesis doctoral. Universidad de Montpéllier II.

MESSICK, S. (1989). "Validity". In: LINN, R. L. (ed.). *Educational Measurement*. 3 ed. New York, Collier Macmillan.

____ (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into scoring meaning. *American Psychologist,* n. 9, pp. 741-749.

MILLMAN, J. y GREENE, J. (1989). "The specification and development of test of achievement and ability". In: LINN, R. L. (ed.). *Educational Measurement*. London, Macmillan.

MORALES, P. (1988). *Medición de actitudes en psicología y educación*. San Sebastián, Universidad de Comillas.

MUÑIZ, J. (1994). *Teoría clásica de los tests.* Madrid, Pirámide.

OJEDA, A. M. (1995). Dificultades del alumnado respecto a la probabilidad condicional. *UNO,* n. 5, pp. 37-55.

OSTERLIND, S. J. (1989). *Constructing test items.* Boston, Kluwer.

POLLATSEK, A.; WELL, A. D.; KONOLD, C. y HARDIMAN, P. (1987). Understanding Conditional Probabilities. *Organization, Behavior and Human Decision Processes*, n. 40, pp. 255-269.

SÁNCHEZ, E. (1996). "Dificultades en la comprensión del concepto de eventos independientes". In: HITT, F. (ed.). *Investigaciones en Educación Matemática*. México, Grupo Editorial Iberoamérica.

SEDLMEIER, P. (1999). *Improving statistical reasoning. Theoretical models and practical implications.* Mahwah, NJ, Erlbaum.

TEIGEN, K. H.; BRUN, W. y FRYDENLUND, R. (1999). Judgments of risk and probability: the role of frequentistic information. *Journal of Behavioral Decision Making*, v. 12, n. 2, p. 123.

THORNDIKE, R. L. (1989). *Psicometría aplicada*. Mexico, Limusa.

TOTOHASINA, A. (1992). *Méthode implicative en analyse de données et application á l'analyse de conceptions d'étudiants sur la notion de probabilité conditionnelle.* Tesis Doctoral. Universidad Rennes I.

TVERSKY, A. y KAHNEMAN, D. (1982a). "Causal schemas in judgment under uncertainty". In: KAHNEMAN, D.; SLOVIC, P. y TVERSKY, A. (eds.). *Judgement under uncertainty: Heuristics and biases*. Cambridge, MA, Cambridge University Press.

____ (1982b). "On the psychology of prediction". In: KAHNEMAN, D.; SLOVIC, P. y TVERSKY, A. (eds.). *Judgement under uncertainty: Heuristics and biases*. Cambridge, MA, Cambridge University Press.

WEBER, R. P. (1985). *Basic content analysis.* Londres, Sage.

*Recebido em nov./2006; aprovado em nov./2006.*