II.8

VALIDEZ Y FIABILIDAD DE LAS OBSERVACIONES SOCIOLOGICAS

Margarita Latiesa

1. Introducción

Una forma de abordar la problemática sobre la medida en sociología consiste en delimitar diferentes formas de medir y operacionalizar los conceptos por medio de las dimensiones, los indicadores y las escalas. Pedro Gonzalez Blasco [1], en el capítulo anterior así lo hace y examina estos aspectos esenciales de toda observación científica.

Pero, una vez especificadas las mediciones surge en el presente capítulo la pregunta de hasta qué punto éstas ofrecen una información significativa y correcta. Así, la observación de cualquier realidad social implica una aproximación por medio de la medición y utilización de algún instrumento de medida o técnica de recogida de datos, que debe reunir las siguientes características (Selltiz, 1976; Repetto, 1977):

1. Relevancia. La medida en términos de atributo presupone que la persona o el objeto pueden ser descritos adecuadamente.
2. Fiabilidad. Constancia de las observaciones que produce el instrumento de medida. Se deben ofrecer medidas fiables, de manera que se obtengan los mismos resultados al volver a medir el rasgo o aspecto, bajo condiciones similares del individuo u objeto en cuestión.
3. Validez. El instrumento de medida que se utiliza en una situación concreta y con propósito determinado debe realmente medir el rasgo que pretende medir. En otras palabras, la medida reproduce el atributo «verdadero».
4. Sensibilidad. Posibilidad de hacer suficientes distinciones con el instrumento de medida y permitir la especificidad y la exactitud de los atributos que se miden.

[1] La observación y la medida a las que nos vamos a referir en este capítulo se entienden, al igual que en el capítulo anterior por Pedro Gonzalez Blasco, en el sentido de correspondencia entre números y propiedades, desde la perspectiva de la sociología empírica, cuyo enfoque es similar al de las ciencias naturales y físicas.

Tipificación. El instrumento de medición debe poseer unas normas o patrones estandarizados con los que comparar las puntuaciones que el individuo ha obtenido al realizar la prueba.

s técnicas de recogida de datos para ser útiles a la observación científica han nplir todos estos requisitos, pero muy especialmente los de validez (grado en la observación mide el concepto «verdadero» y posición de la persona) y fia- l (grado en el que la observación es consistente y constante).
medición en sociología está sometida a error, lo que generalmente se deno- :los problemas de la medida». Los objetivos de este capítulo son dos de estos mas, la validez y la fiabilidad [2], que están implícitos en todo tipo de actividad ca que implique la operacionalización y la medición de conceptos y han sido mente estudiados y medidos por diversos investigadores. El estudio y esclare- to de estos conceptos permite depurar la metodología necesaria para llevar a ı medición en ciencias sociales, para contrastar las hipótesis y para acumular ocimientos científicos.
:ctuar las pruebas de validez y fiabilidad parece no sólo recomendable sino :indible en las ciencias sociales. Sin embargo, hay que indicar, que tradicio- ıte en la investigación sociológica no se suelen efectuar ni en el análisis de ido, ni en las encuestas, y mucho menos en otros tipos de observaciones más :ivas. Más recientemente, esta escasez de análisis está cambiando, ya que el » que observamos desde los años treinta de encuestas sociales y de «medidas :alidad social», ha permitido actualmente la existencia de numerosos bancos)s y la acumulación de observaciones que han de ser analizadas. Esta sobre- ıncia hace que «volvamos la mirada» sobre la validez y la forma en que los on obtenidos, la fiabilidad de nuestras observaciones y de nuestras encuestas :finitiva, a que seamos más meticulosos sobre el proceso de medición y los nas de la medida (Blalock, 1968, 96).

medida y el error

lir es asignar números para representar propiedades (Campbel, 1921). La ın conecta los conceptos «abstractos», las propiedades, con los «indicadores os», los números (Blalock, 1968, 6; Riley, 1963, 23). Pero esta conexión que :iona la medida de cualquier fenómeno resulta imposible de realizar sin erro-

.ervadores diferentes no tendrán la misma percepción de un fenómeno. Esto que ningún método de observación está libre de sesgos. El tipo de observa- ıue interesan a la sociología (los conocimientos humanos, las opiniones, los s, los comportamientos, las capacidades, etc.) impide que los instrumentos tan estas observaciones sean completamente fiables y válidos. Pretender que ltados sean idénticos en todas las observaciones y que la medida refleje una acterística es imposible. Por otra parte, esto no debe limitar la obligación

capítulos de este libro se ocupan también de estos temas, pero, desde una perspectiva muy a la que se adopta aquí. El capítulo de Francisco Alvira expone la validez de los diseños de ión y el capítulo de Eduardo López-Aranguren expone la validez y la fiabilidad en el análisis ido. En ambos se trata de evaluaciones específicas aplicadas al diseño en el primer caso y al ; contenido en el segundo. Sin embargo, en el presente capítulo se lleva a cabo un enfoque esde el punto de vista de la teoría de la medición.

que tiene el investigador de esmerarse en conseguir que sus observaciones sean potentes desde el punto de vista de la validez y de la fiabilidad.

Los resultados obtenidos por una medida no solamente dependen de las diferencias que existen entre los individuos respecto a una característica, sino que también dependen de factores transitorios y estables. Los factores transitorios o errores aleatorios determinan la fiabilidad y los factores estables, o errores sistemáticos, determinan la validez. La variación entre las medidas que produce un instrumento procede de la variación real entre los sujetos y de los errores de medición.

La observación y la medida tienen lugar en situaciones en las que numerosos factores contribuyen a ocasionar los errores (grado de atención o cansancio, grado de cooperación...) y el objetivo es incrementar la validez y la fiabilidad reduciendo tales errores.

En todas la ciencias, físicas naturales o sociales, el producto de una observación o de una medición surge de la interacción entre un OBSERVADOR el INSTRUMENTO DE MEDICION y el sujeto u objeto OBSERVADO. Y las fuentes de error proceden de: *a)* cada uno de estos tres aspectos considerados aisladamente, *b)* la interacción entre los mismos, y *c)* el medio y situación física y/o social en la que se inscriben y desarrollan.

Parece indiscutible que las posibles fuentes de error son superiores en las ciencias sociales, donde los instrumentos de medición son más imprecisos, las situaciones donde se produce el proceso de aplicación más difícilmente reproducibles y los cambios reales de las características y hechos sociales muy probables.

2.1 *Fuentes de error*

Como consecuencia de lo expuesto, no es posible hacer una enumeración exhaustiva de todas las fuentes de error que existen en la investigación sociológica, pero sí al menos podemos mencionar aquellas fuentes más conocidas y citadas. Así, los tipos de errores que propician la no validez y la no fiabilidad son los siguientes (Webb y otros, 1966; Selltiz, 1976):

A) *Errores del investigador (características y conducta de la persona que recoge los datos)*

Una primera fuente de error emana de los propios supuestos y presunciones del investigador. El investigador, posee unas estructuras de orientación, prejuicios, etc. que pueden interferir en la interpretación de las conductas ajenas.

Todos estos valores y supuestos del investigador deben ser conocidos por el mismo, con el fin de reconocer los errores que como consecuencia se pueden presentar. En este punto no interesan cuáles sean los supuestos, tan sólo que el investigador sea consciente de ellos.

El clínico, el experimentador o el observador participante tienen serias dificultades en detectar *la desviación de su propia orientación,* al mismo tiempo que recogen los datos. Sin embargo, el director de una encuesta encuentra mayores facilidades en aislar la problemática propia de su orientación y la problemática correspondiente a las desviaciones de los codificadores, entrevistadores, etc, pudiendo, así, descontar estos últimos efectos en la interpretación de los datos.

Es indudable que las características y la conducta de la persona que recoge los

influye en la calidad de los mismos e incluso las variaciones que se producen administración del instrumento. El entrevistador en la encuesta, el codificador análisis de contenido, o el observador de un grupo, pueden estar más o menos idos, cansados, etc.; o por el contrario volverse más diestros con las aplicacio-

Errores del instrumento y del análisis

Jna segunda fuente de error se puede deber a factores de claridad del instruo de medida. Las personas pueden interpretar de modo distinto las preguntas nidas en el instrumento de medida y las respuestas pueden reflejar diferencias interpretación más que diferencias en el contenido de la pregunta.
'ambién hay que tener en cuenta los factores debidos al formato del instrumen-l cuestionario recoge respuestas verbales, la observación recoge los aspectos uctuales, ello hace que aunque se mida la misma característica los resultados n. Igualmente, no es lo mismo recoger la respuesta en una escala de 10 items, en una dicotómica de si/no.
(por último hay que considerar los factores debidos al análisis. Errores en la icación. tabulación y análisis estadísticos y no estadísticos.

Errores de los investigados

.. Conductas de los estudiados como sentirse amenazados, cooperar o por el contrario no cooperar con la investigación, y factores personales como humor, salud, distracción, etc., influyen en las respuestas y especialmente el grado de interés que tenga la persona en el tema que se le propone. Por ejemplo, alguien que no está interesado por las cuestiones políticas, ni por las acciones o palabras que pronuncian los miembros del gobierno, contestará con desinterés y desgana a las preguntas de este tipo. La misma medición hace que la observación en un instante y sobre unos temas afecte a los siguientes.
2. Elección de un rol que el estudiado considere adecuado para la situación, pero, que no coincida con su opinión (deseabilidad social).
   La deseabilidad social consiste en la tendencia de las personas a dar una buena imagen de sí mismas, a dar respuestas de cierta respetabilidad y aceptación social.
   Las personas se diferencian en el grado en el que manifiesten este aspecto y ello afecta a los resultados de la medición. Así, se ha confirmado la propensión que existe de dar respuestas deseables, de tal forma que las diferencias en las respuestas a determinadas preguntas pueden manifestar el grado en que una persona está dispuesta a admitir que mantiene conductas indeseables, más que diferencias respecto al contenido de la pregunta (Edwards 1957; Crowne y Marlowe, 1964).
3. La tendencia al asentimiento y a dar respuestas positivas o negativas, respuestas estereotipadas, etc (aquiescencia).

Algunos autores han identificado este aspecto como una propensión de las perıs a expresar conformidad o disconformidad (Cronbach, 1950; Phillips, 1971;

Couch y Keniston, 1960) . Sin embargo, Scott (1968) observa que no se trata de una tendencia de los individuos, sino que es una característica del instrumento de medida (Scott, 1968)

Los problemas de aquiescencia y deseabilidad social afectan a la validez de las mediciones y han sido muy estudiados en el campo de la psicología [3].

D) *Factores de la situación*

1. Factores que influyen debido a la situación en que la medida tiene lugar: ambiente relajado, tenso, presencia o ausencia de determinadas personas, etc.
2. Factores mecánicos. Aspectos tan triviales como la falta de espacio para anotar, la rotura de la vestimenta o de un lapicero, pueden contribuir a distorsionar la medida.

E) *Errores de muestreo*

1. Limitaciones de la población a la que se puede generalizar los resultados.
2. La inestabilidad de las características de la población que cambian a través del tiempo hace que los datos de estudios longitudinales no sean comparables.
3. La hetereogeneidad de la población a traves de las zonas hace que un determinado procedimiento de recogida de datos sea más útil o efectivo en unas zonas que en otras.
4. Las limitaciones que impone el método de recogida de datos o de medición sobre los contenidos para los que resulta apropiado.
   También hay que tener en cuenta las limitaciones del método de muestreo de items, ya que no es posible utilizar un universo entero con todos ellos.
5. La inestabilidad de los contenidos de la investigación a través del tiempo impide la comparabilidad.
6. La inestabilidad de los contenidos a través de las zonas geográficas.

2.2. *Cómo evitar las fuentes de error*

El sesgo que introduce el *investigador* debe de reducirse en lo posible, por medio de dos procedimientos: la clarificación teórica y la aplicación de los métodos de investigación social.

La clarificación teórica es un requisito previo e indispensable para determinar qué se va a medir. En efecto, para asegurarnos la calidad de la investigación debemos saber qué cuestiones vamos a investigar y especificar tales cuestiones definiendo los conceptos.

La aplicación de los métodos en el establecimiento de procedimientos de medida implica la utilización de técnicas de recogida de datos y unas normas para la uti-

[3] Ante los problemas de la aquiescencia y la deseabilidad social, la American Psychological Association recomendó que los investigadores eliminaran estas influencias en la construcción de los tests.

de los mismos. En palabras de Selltiz (1976), los *procedimientos de medida* /en «las definiciones de trabajo» de los conceptos y las *normas* «facilitan el)s datos».

:sar de que es imposible evitar los errores que propician la falta de *validez* y d podemos al menos tratar de minimizarlos. Para ello, actualmente se está o gran interés a un enfoque que consiste en llevar a cabo el proyecto de in-ón de manera que maximice la validez y fiabilidad, en lugar de evaluarlas a :i. En otras palabras, cuando se planea la investigación se hacen los *diseños*)s para disminuir los errores.

)tra parte, como ningún instrumento es totalmente válido y fiable, ya que por n, no existe ninguno que refleje exclusivamente diferencias en la característi-iada, debido a que la conducta humana está influida por múltiples factores, se aconseja utilizar *varios instrumentos* en la misma investigación (Campbell 1959) , y emplear *varias medidas* de un mismo concepto.

ı nivel más general, podemos hacer disminuir las fuentes de error durante el investigador mediante el *control* riguroso de todos y cada una de las fases de igación. La mayoría de los tipos más obvios de error pueden ser superados de ıera. El control se refiere a los requisitos metodológicos y técnicos necesarios cumplimiento planificado de las observaciones. Es decir, utilizar correctamen-étodos adecuados en la aplicación de las técnicas de recogida de datos, en la ı y preparación de entrevistadores o de observadores en general, en la depura-visión de observaciones y cuestionarios, así como la codificación.

control estricto es especialmente necesario en sociología, por la peculiar rela-: se establece entre observador y observado. Es más, aunque el investigador ue no puede dedicar el tiempo, esfuerzo y dinero necesarios para la compro-e la fiabilidad y validez de su observación, análisis, cuestionario o entrevista, debe controlar cuidadosamente el trabajo de campo, las preguntas, las ins-:s dadas a los codificadores, entrevistadores, etc., para asegurarse de que se ıdo fuentes de error.

ı literatura existente sobre los problemas de las medidas en ciencias sociales, concede la importancia que tiene para minimizar los errores, al control y al nto exhaustivo en todos los momentos de la investigación. Esto quizás se ue es prácticamente imposible cuantificar la influencia positiva que tiene un estigador, diligente y siempre pendiente de las fuentes de errores que se origi-ontraste con otro menos preocupado y sensible a estos temas.

rores aleatorios y sistemáticos

s las fuentes de error que hemos enumerado propician que en la investiga-al se incorporen los *errores aleatorios* y los *errores sistemáticos* .

ror aleatorio obedece al azar, es debido a las situaciones *coyunturales* que he-:rito de la persona, de la situación, del procedimiento de medida, etc y afecta lidad.

:rrores aleatorios son aquellos factores azarosos que distorsionan la medi-ı fenómeno; son un conjunto de variables cuyos efectos nos son desconoci-previsibles y sus efectos múltiples y actúan de forma tal que no siguen nin-que no sea, precisamente, la del azar [4]. La cantidad de error aleatorio está ente relacionada con el grado de fiabilidad de la medida del instrumento.

[4] ecisamente «del desorden de la aleatoriedad el científico extrae el orden de la predicción y ıtíficos» (Kerlinger, 1975, 96).

Estos errores estan siempre presentes en cualquier medición que llevemos a cabo (Stanley, 1971, 356) y, por tanto, en todos los tipos de investigación (encuesta, observación participante, análisis de contenido, simulaciones y experimentos). Es endémico a la investigación social y, también, en otras áreas como la física o la biología.

El error sistemático, por el contrario, influye *estructuralmente* en la característica objeto de medida y aunque muchas veces pasa inadvertido al investigador afecta a la validez.

Los errores sistemáticos siguen reglas fijas. Un ejemplo de error sistemático sería un termómetro que siempre registra dos grados por encima al tomar la temperatura.

Otro ejemplo donde podemos comparar los errores sistemáticos con los errores aleatorios sería el siguiente: si un rifle está mal calibrado y siempre que se dispara con él tiene un desvío hacia arriba y hacia la izquierda, todas las personas que lo utilicen cometerán un error al disparar sobre una diana. Este error es sistemático y afecta a todas las personas por igual. Ahora bien, si en el momento en el que dispara la persona «X» existe una ráfaga de aire y desvía ligeramente la bala y sin embargo esto no ocurre con la persona «Y», nos enfrentamos a un error aleatorio. También sería un error aleatorio si una misma persona en un disparo tiene una gran concentración y en otro disparo tiene menor concentración. En todos estos casos el error no afecta por igual a los disparos que se producen, ni a los individuos que los efectúan.

En resumen, el estudio de la fiabilidad se centra en la determinación de en qué grado las diferencias de las puntuaciones se deben a influencias aleatorias o casuales. Cuanto menos influyen los errores de azar mayor será la fiabilidad y la consistencia de nuestras mediciones. Pero, un instrumento puede ser fiable y tener errores sistemáticos y por tanto, ser no válido. El estudio de la validez se centra en la determinación de las influencias sistemáticas, que impiden que un instrumento mida realmente los comportamientos sociales.

Para algunos autores, como Selltiz, la validez incluye los errores aleatorios y sistemáticos, de tal forma que si se demuestra que un instrumento de medida es válido, no es necesario demostrar que es fiable. Otros autores, sin embargo, asocian los errores sistemáticos exclusivamente con la validez, y los errores aleatorios con la fiabilidad. A efectos prácticos, es lo mismo optar por un enfoque que por otro, ya que demostrar completamente la validez de un instrumento es imposible y en consecuencia, debemos estudiar siempre la fiabilidad y llevar a cabo un estudio detallado de los errores de tipo aleatorio que pueden influir.

3. Validez y fiabilidad

Este apartado constituye el núcleo central del capítulo y será, por tanto, en el que más nos vamos a extender. Se expone en primer lugar, la definición de los conceptos de validez y fiabilidad, en segundo, la clasificación de los diferentes tipos y estrategias que existen (criterios, definiciones, fórmulas, etc.), en tercero, las relaciones existentes entre la validez y la fiabilidad y por último, la evolución histórica que han tenido estos términos.

3.1. *Conceptos*

La validez hace referencia a que el procedimiento utilizado mide lo que realmente pretende medir y la fiabilidad hace referencia a la propiedad del instrumento que produce los mismos resultados en diferentes pruebas.

lidez trata de determinar el modo en que la «realidad» queda reflejada en . que hacemos de ella. Es decir, si tratamos de captar la «inteligencia», que tengamos de nuestra medida sea la «inteligencia» y no otra cosa. Podemos ›ariencias», «realidades engañosas» y de esta forma nuestras mediciones idas.

fine como el grado en que una medida mide «realmente» lo que pretende en otras palabras, el grado en que las diferencias de puntuación reflejan :e» las diferencias existentes entre los individuos en la característica estu-

lefinición tan sencilla de la validez implica serias dificultades, ya que no sa- al es la «realidad», ni la «verdadera» puntuación del sujeto y por tanto no compararla con la puntuación obtenida por el instrumento de medida. Así lo, si queremos determinar cuál es la estructura de clases sociales en Espa- ›os medir la «estructura de clases real». Pero es evidente que esta adecua- : la medida y la realidad no es fácil, porque cuando los sociólogos decimos mos y medimos «las clases sociales», ¿qué es lo que medimos? Desde luego clases en su sentido absoluto, sino en sentido relativo, de acuerdo con de- ıs definiciones apriorísticas.

bilidad hace referencia a la exactitud de las medidas, para ello el instru- medida debe ser adecuado y estar bien «calibrado» y medir lo más exacta- ;ible lo que pretendemos captar.

›blema de la fiabilidad es el de la determinación del grado en que las medi- ıidas con los instrumentos de medición, estan libres de errores de tipo alea- ı puntuación observada (X) es una estimación de la realidad verdadera nbién incluye un cierto error (E):

$$X = XV + E$$

ma es algebraica, es decir, X puede sobreestimar o subestimar el valor de ıbilidad será mayor cuanto menor sea el error de medición y depende, por que la puntuación del sujeto sea independiente o no del error.

ocedimiento de medida es fiable en el grado en que las medidas de un fe- lan resultados similares. El concepto de fiabilidad se ha aplicado para muy ›peraciones y conceptos y se ha utilizado de forma muy diferente para su ción, no obstante, siempre supone algún tipo de medida de la constancia ncia [5]. Para algunos autores como Köning (1973) esta constancia de la me- el principio más general e importante del desarrollo de la observación

›eficientes de fiabilidad aportan una medida de la coincidencia o falta de :ia de las respuestas, pero no indican cuáles son las razones por las que no ıcidencia. En otras palabras, los coeficientes de correlación que operacio- fiabilidad, cuantifican la diferencia pero no delimitan qué parte corres- ada una de las posibles fuentes de influencia. En el caso del test-retest las atribuyen a cambios en las características transitorias de las personas o a n las condiciones de medición. En el caso de diferentes observadores con

—

›osos autores han propuesto el cambio del término «fiabilidad». Así, por ejemplo, Cron- '0, indicó que sería preferible utilizar el término «generalización». Loevinger (1957) acon- separadamente el término «homogeneidad» del término «fiabilidad». Sin embargo, ningu- propuestas ha tenido éxito y los diversos autores se siguen refiriendo a la fiabilidad para os, conceptos y aplicaciones.

un mismo instrumento las fuentes se atribuyen a diferencias entre los observadores. En el caso de formularios diferentes de un test las fuentes se atribuyen a diferencias de muestreo de los items.

3.2. *Tipos o estrategias*

Para medir la validez de un instrumento de medida se comparan los resultados que proporciona con otras «evidencias significativas» y se examina la exactitud o inexactitud de los mismos. Existen, no obstante, diversas formas de llevar a cabo estas comparaciones y, por tanto, diversos tipos de validez. Para hacernos una idea de la complejidad conceptual del término algunos autores como Broown (1970) indican que existen en la literatura aproximadamente 40 tipos de validez, posteriormente Brinberg y Mcgrath (1982) enumeran 10 términos distinos de uso frecuente y Messick (1980) encuentra 17 enfoques de validez.

Esta gran variabilidad se justifica en parte por la diversidad de los métodos de observación, ya que los instrumentos de medida difieren entre sí en los siguientes aspectos (Manheim):

- Unos son más estructurados que otros. El experimento de laboratorio sería el extremo de mayor rigidez, formalidad y se fija en ciertos comportamientos muy específicos; la observación participante sería el otro extremo de menor estructuración y mayor flexibilidad en la recogida de datos y en los aspectos que estudia.
- Con unos se pueden hacer más inferencias y generalizaciones que con otros, es decir, permiten en mayor o menor medida la abstracción y el alejamiento de los datos inmediatos tal y como los capta el investigador.
- También difieren en la medida en que los sujetos observados son o no conscientes de este hecho, ya que si son conscientes existe la posibilidad de modificación de la conducta.

Estas diferencias en los métodos de investigación implican una cierta diversidad en las características estudiadas e influyen notablemente en los tipos de validez y fiabilidad que podemos emplear.

A pesar de que numerosos autores se siguen refiriendo a los *tipos de validez* conviene hacer notar que actuamente, a raiz del artículo de Lawshe (1985), se habla más de *tipos de estrategias* para verificar el significado de la variable y usos pontenciales del instrumento. En estas páginas utilizaremos, no obstante, las dos denominaciones, tanto en lo que se refiere a la validez como a la fiabilidad.

Los medios que se emplean para determinar la validez de un instrumento son diferentes según el tipo de validez a la que se refiera. Así por ejemplo, la validez aparente o la validez de contenido se refieren al instrumento de medida y son independientes de cómo se interpreten las respuestas. Sin embargo, en la validez de constructo, no se trata de una característica del instrumento, sino de una propiedad de las inferencias que pueden hacerse a partir de las puntuaciones con él obtenidas.

Y las mismas observaciones que estamos efectuando con la validez, son aplicables a la fiabilidad. Dadas las grandes diferencias que existen entre los numerosos métodos de observación existentes, cada uno de ellos tiene su propia forma de calcularla. Para llevar a cabo la comprobación de la fiabilidad desde el punto de vista de las observaciones se pueden efectuar las siguientes acciones:

Repetición de los actos de observación por la misma persona.
Repetición de los actos de observación por personas diferentes, bien en el momento o en distintos momentos.
Constrastación con otros materiales, como son resultados de observaciones es que han sido registradas por escrito. Siempre que sea posible en una in-ión, deben de contrastarse las informaciones con informes estadísticos, aun-ıralmente estos también pueden tener errores [6].

· la clasificación de los tipos de fiabilidad es más conocida desde el punto de la aplicabilidad del instrumento. De forma resumida podemos identificar :ctos:

tabilidad. El procedimiento consiste en efectuar varias aplicaciones del mis-ɔ instrumento de medida con el fin de comparar los comportamientos de los lividuos en diferentes momentos del tiempo.
uivalencia. El procedimiento consiste en efectuar varias aplicaciones con :tintos instrumentos con el fin de comparar los resultados.
nsistencia interna. El procedimiento consiste en una sola aplicación del smo instrumento, con el fin de medir el grado en que las respuestas de los lividuos a los distintos ítems, u otros componentes de una medida, son con-:entes.

aspectos implican diferentes formas de contrastar la fiabilidad y, por tanto, ipos de fiabilidad.
ıbilidad en el método de observación se suele calcular por medio del acuer-jueces, ya que de la observación de los jueces proviene la mayor parte de ɔrales, 1981) . Se puede calcular el acuerdo para toda la sesión, o bien por : o por categorías. También se puede calcular por medio de la estabilidad :rvación en momentos diferentes, pero es menos frecuente.
almente se emplea el método test-retest en las encuestas sociológicas, es abilidad entendida como estabilidad, mientras que en los test psicológicos n los métodos de las dos mitades o fiabilidad entendida como equivalencia ncia interna.

'pos de validez

ımos a referir a la clasificación más aceptada:

z de contenido.
z de criterio.
z de constructo.

clasificación hay que efectuar dos observaciones: en primer lugar que incluye la validez aparente en el apartado de validez de contenido, consi-ue requiere una mención especial porque es fundamental en el campo so-ɔor el tipo de variables y observaciones en las que se interesa; en segundo

ɔ es aplicable cuando en vez de buscar el conocimiento exacto de hechos objetivos, el in-sca saber qué es lo que la gente cree que son los hechos.

lugar, que no nos vamos a referir a la validez interna y externa porque esta tipología se centra en los diseños de investigación y ha sido expuesta previamente en este libro. Podemos, no obstante, recordar de forma muy simplificada que la validez interna se basa en comparaciones entre las características diferentes de ciertos grupos y la validez externa se ocupa de las generalizaciones (Alvira, 1976, 1991). O en otros términos, la validez interna responde a la pregunta ¿se obtendrían resultados diferentes si se hubieran utilizado procedimientos diferentes?, mientras que la validez externa se pregunta ¿cuán generalizable es el procedimiento utilizado? (García Ferrando, 1985, 34).

Validez de contenido

En este tipo de validez se utilizan los juicios con frecuencia consensuados, para determinar si el contenido de los ítems es apropiado o no lo es.

La validez de contenido se pregunta si el instrumento de medida sirve para medir el comportamiento que pretende. Para ello, debemos especificar el universo de los comportamientos y la variedad de formas en que podemos medirlos; y en el caso de los test identificar todos los posibles ítems útiles para la medición. Se asume que todas las mediciones sirven para medir el concepto objeto de estudio, pero no podemos asegurar que se estan midiendo todas las dimensiones del concepto.

La validez de contenido se refiere al grado en que los ítems de una escala o test representan un determinado universo temático. Se fija en una propiedad del instrumento de medición (formulación de los ítems) y no en las inferencias que pueden hacerse con las puntuaciones obtenidas, por ello, algunos autores cuestionan que pueda hablarse con propiedad de validez.

La validez de contenido trata de determinar la relevancia o representatividad de los ítems en cuanto muestra adecuada de un dominio previamente especificado. La adecuación del contenido de los ítems con un plan previo, responde a la necesidad de construir un instrumento y nada tiene que ver con el significado de las puntuaciones. Por esta razón, algunos autores han llegado a cuestionar que la validez de contenido sea un tipo de validez y a afirmar que no es apropiada para determinar si se «mide lo que se pretende medir» (Messick, 1980, 1981; Tenopyr, 1977; Cronbach y Meehl, 1955). Otros autores, sin embargo, como Yalow, Popham (1983) y Ebel (1983) aclaran que si el contenido es relevante y representativo esto significa que existe una inferencia implícita, porque, aunque es una propiedad del instrumento, una vez que se obtienen las respuestas, la validez de contenido hace posible su interpretación. Así, el contenido permite y justifica interpretar resultados.

Algunos autores también denominan validez aparente a la validez de contenido, pero conviene que delimitemos sus características propias por la transcendencia que en el campo sociológico tiene este tipo de validez.

Un instrumento tiene validez aparente si mide lo que parece. Este tipo de validez se considera en muchos casos trivial y no se tiene en cuenta, pero sin lugar a dudas tiene su importancia.

La validez aparente suele emplearse en dos sentidos o matices:

1) Un instrumento tiene validez aparente si parece a apropiado a la situación. Así por ejemplo, un test de aptitud numérica parece apropiado para predecir el rendimiento en matemáticas.
2) Un instrumento tiene validez aparente si la denominación que se le d-

rece a simple vista coherente con la formulación de los items y, por lo tanto, mide lo que se dice que mide.

validez aparente se ocupa de las medidas que se efectúan directamente. Así, idad de un atleta se cronometra observando su comportamiento. Las pregun- cuestionario como el sexo, los metros cuadrados que ocupa la vivienda, los rendimiento, etc., son ejemplos de observaciones directas, donde la validez rumento de medida se presenta en su propia apariencia. El problema que se lucir es que la apariencia puede ser momentánea y no representar la realidad ia.

nbién se denomina validez facial o análisis lógico, porque la validez de los parece a simple vista utilizando el sentido común. Se aplica cuando los datos asados en la observación directa y no se necesitan inferencias. La mayor parte reguntas que se efectúan en los cuestionarios en Sociología pertenecen a este xo, edad, etc.

de criterio (concurrente y predictora)

/alidez de criterio se mide directamente por medio de los coeficientes de co- n que se establecen con otras medidas o criterios. Suele dividirse en concu- ambas medidas se obtienen simultánemanete) y predictora (la medida en el es posterior).

/alidez concurrente hace referencia a alguna característica por la que los in- s difieren en el presente. La validez predictora hace referencia a alguna ca- ica por la que los individuos se diferenciarán en el futuro. Lo importante es ambos casos se trata de una validez *pragmática* porque la validación se efec- rastando con otro procedimiento de medida empírico. No interesa aquí si se liendo algún concepto teórico determinado, lo que interesa es saber si el ins- o de medida es *útil* para medir una conducta en el presente o en el futuro. nplo, si queremos predecir el rendimiento académico de los alumnos y en- os determinadas características como potenciadoras del mismo, no nos inte- er por qué esas características son buenas predictoras, sino que el hecho de emuestre que efectivamente lo son, ya es suficiente.

enomina validez pragmática porque comparamos los resultados con algún independiente, como por ejemplo, contrastar las fuentes de datos secunda- las fuentes de determinadas preguntas del cuestionario elaborado por el in- or. El criterio también puede ser alguna predicción basada en los resultados stan validando, o puede basarse en grupos que se sabe ocupan posiciones en la variable que se está midiendo.

eórica o de constructo

os los tipos de validez están relacionados: la validación pragmática y la vali- eórica no son excluyentes, ya que un instrumento de medida puede tener ojetivos y por otra parte, no debemos quedar satisfechos si tan sólo hemos o la validez pragmática y no hemos respondido al por qué de su efectividad. parte no se trata de demostrar un «tipo» de validez y quedar satisfechos,

hay que validar desde diferentes aspectos y difícilmente podemos concluir de manera certera que nuestras observaciones son válidas.

La validez no queda demostrada por unos coeficientes de correlación aislados (validez concurrente o predictiva), ni por juicios sobre la relevancia del contenido, sino que queda integrada en todo un proceso de verificación de hipótesis teóricas, donde entra tanto lo empírico como los juicios racionales. Como consecuencia, podemos afirmar que la validez de constructo subsume las nociones tradicionales de validez de contenido y de validez de criterio (Morales, 1988, 371). Así, la cuestión del contenido tiene que ver con la relevancia y la validez del criterio con la utilidad del instrumento, pero, ni el contenido, ni las relaciones con otras variables, pueden separarse de la interpretación del constructo. En este sentido la validez de constructo integra todos los tipos de validez.

La validez de constructo (o teórica) es un proceso mediante el cual acumulamos evidencias, pruebas empíricas, sobre relaciones teóricamente importantes que apoyan —o demuestran— en cierto grado una determinada inferencia o interpretación de las puntuaciones. De las relaciones observadas entre lo que medimos y otras variables, deducimos la naturaleza de las respuestas.

Por una parte hay que considerar el papel que tiene el instrumento de medida, ya que Fiske (1973) ha demostrado que las intercorrelaciones entre un conjunto de variables que miden la personalidad varían según sea el instrumento de medida, incluso utilizando una misma muestra de personas. Pero, la validez de constructo implica la validez no solamente del instrumento sino, también, de la teoría. No se interesa por la predicción de comportamientos, sino por la medición de alguna característica construida a nivel teórico (que no se puede observar ni medir directamente) y que se supone que nuestro instrumento de medida cuantifica. La caracerística que se mide trata de reflejar el grado en el que realmente la posee un individuo, relacionando el propio instrumento de medición con la estructura teórica general [7]. Así, la inteligencia, la actitud hacia el aborto, etc., son constructos teóricos y de su validación como tales se ocupa la validez de constructo, aunque normalmente donde más se emplea es para los rasgos de personalidad.

Para determinar la adecuación de la medida del concepto Campbell y Fiske (1959) proponen dos tipos de validación:

* Validación convergente. Diferentes medidas de un concepto proporcionan resultados semejantes. Para comprobarlo es necesario medir el concepto con varios procedimientos.
* Validación divergente. La medición del concepto se diferencia de otros conceptos. Para comprobarlo es necesario medir con el mismo procedimiento otros conceptos de los que se supone se debe diferenciar el de la investigación.

El método de Campbell y Fiske se basa en la matriz multi-rasgo multi-método y trata de verificar los dos tipos de hipótesis (convergente y divergente). En la validación convergente se comprueba la correlación entre métodos distintos que presumiblemente miden el mismo rasgo. Así, cada instrumento de medición se concibe como un método-rasgo. En la validación divergente rasgos distintos medidos con el mismo método o métodos semejantes no estan relacionados. Las correlaciones entre métodos distintos que miden el mismo rasgo, deben ser mayores que las correlacio-

[7] La validez de constructo ha generado muchos desacuerdos entre los sociólogos (Sjoberg, Nett, 1968).

re métodos semejantes que miden rasgos supuestamente distintos. Luego, el
› de estos autores «trata de clarificar el significado utilizando diversos enfo-
métodos para medir lo mismo (convergencia) y se distingue de conceptos afi-
/ergencia)» (Morales, 1988, 435).
›ra bien, estas pruebas de validación de constructo no son universales y
rales. Así, una actitud perfectamente medida y validada con un instrumento
ser muy útil para una época determinada, pero no para otra. Las generaliza-
›robadas no son estables en las ciencias físicas y mucho menos en las ciencias
:. En consecuencia, tenemos que afirmar con Nunnally (1978), que las prue-
validación de constructo más que probar la «verdad» de la teoría o construc-
ícito en un instrumento de medición, muestran su grado de pertinencia y uti-
ara investigar una realidad.

Tipos de fiabilidad

este apartado vamos a presentar los tipos de fiabilidad y las formas de deter-
›s.

'lidad como estabilidad

estabilidad de los resultados de un instrumento de medida se establece com-
› los mismos en aplicaciones repetidas. Las diferencias en los resultados y por
a inestabilidad de las mediciones, puede ser debida a cambios reales ocurri-
cambios debidos al azar y a factores extraños. Es difícil determinar qué parte
›nde a los cambios reales y qué parte corresponde a los errores, aunque al-
utores han expresado algoritmos para hallarlos, demostrando que es posible
ir los cambios verdaderos de los aleatorios mediante tres aplicaciones o bien
:e los indicadores múltiples de los modelos paht (Heise, 1969; Wiley, Wiley,
)74).
ndo el instrumento de medida se basa en la observación, es necesario reali-
:levado número de aplicaciones; en la encuesta se utilizan normalmente dos
›nes; y el análisis de contenido es un caso intermedio entre los dos expues-

›étodo más conocido de comprobación de la fiabilidad como estabilidad se
con el nombre test-retest. En una investigación por encuesta, se aplica el
:uestionario a las mismas personas en distintos momentos del tiempo, bajo
›nes equivalentes y posteriormente se comparan los resultados. Las concor-
obtenidas entre las puntuaciones de las dos pruebas se establecen mediante
lación. En consecuencia, el coeficiente de fiabilidad se define matemática-
or medio del coeficiente de correlación.
:l test-retest también se pueden distinguir las dos medidas que hemos indica-
una parte el cambio real o su inversa, el coeficiente de estabilidad; y por
cambio aleatorio o su inversa, el coeficiente de fiabilidad. A mayor cambio
›or será el coeficiente de estabilidad y a mayor cambio aleatorio, menor será
iente de fiabilidad.
método plantea algunos inconvenientes que se derivan de su aplicación. En
ugar, el mismo proceso de medida al repetirlo puede aumentar las diferen-
›eras, como por ejemplo el interés y la motivación pueden ser menores en la

segunda aplicación debido a que el encuestado ya participó activamente en la primera. También puede recordar las respuestas que emitió en la primera aplicación y no contestar espontáneamente en la segunda. Así, el cuestionario es el mismo pero la situación de la entrevista puede ser diferente.

También puede darse la posibilidad de cambios reales entre las dos aplicaciones como consecuencia de la primera aplicación. Así por ejemplo, hay encuestados a los que el descubrimiento de un tema por medio del cuestionario, hace que se interesen por el mismo y cambien su opinión o actitud realmente como consecuencia de la información adicional.

Todos estos problemas y limitaciones se concretan según Carmines y Zeller (1979) en cuatro:

1) A menudo los investigadores sólo pueden hacer una aplicación, no sólo por el elevado coste que supone aplicar dos veces el mismo cuestionario, sino porque además hay veces que es imposible hacerlo, aunque se dispusiera de dinero suficiente.
2) El cambio verdadero [8] que ha podido suceder en determinadas características o actitudes, se interpreta en este método como inestabilidad y subestima la fiabilidad, haciendo que los coeficientes sean menores. Esto sólo se puede solucionar separando el efecto del cambio verdadero del cambio aleatorio o error. Heise (1969) ha demostrado que esto es posible con tres aplicaciones del mismo cuestionario.
3) La reflexividad en las ciencias sociales. Es decir, la medición de un fenómeno puede inducir a modificaciones del fenómeno mismo. Al obtener información de una persona en un momento de tiempo puede hacer que la persona se sensibilice hacia el tema y cambie su respuesta en el momento dos. Consecuentemente, la fiabilidad puede ser menor debido a esta reflexividad.
4) También se puede sobreestimar la medida de fiabilidad debido a la memoria. Si el tiempo que media entre las dos aplicaciones es corto, la memoria en la segunda puede influir y los temas que se recuerdan pueden dar coeficientes de correlación más altos que aquellos otros temas que no se recuerdan (Nunnally, 1964).

Para obviar estos problemas se aconseja que transcurra un tiempo suficiente entre las dos aplicaciones para borrar los efectos de la primera aplicación y no sobreestimar la fiabilidad. Ahora bien, tampoco puede ser demasiado tiempo, para que no se produzcan cambios verdaderos y no subestimemos la fiabilidad. Si el tiempo no es suficiente puede influir la memoria de lo que se contestó en la primera aplicación y dará lugar a coeficientes de fiabilidad más altos; pero si se distancian mucho las aplicaciones, pueden darse cambios de hecho en el sujeto y dar coeficientes más bajos de los reales. En general, el intervalo de tiempo puede ser más reducido cuanto menor sea la edad de los sujetos y cuanto más susceptible sea el rasgo medido de experimentar cambios.

Especialmente en psicometría se recomienda abiertamente que no se utilice el test-retest, o que si se hace, «que sea con aquellos instrumentos en los que menos

[8] Sobre este tema de la diferencia entre cambio verdadero y cambio aleatorio, o lo que se denomina comúnmente como el problema de la estabilidad de la medida, se recomienda la lectura de los siguientes autores: Carmines y Zeller (1979); Heise, (1969); Wiley and Wiley (1970, 1974); Erikson (1978)

: la memoria, la práctica o el aprendizaje, como son todos aquellos que impli- ıreas más o menos rutinarias; sin embargo, para aquellos elementos que exi- azonamiento, desarrollo de estrategias, ingeniosidad..., no es adecuado» (R. , 1986, 196). Así por ejemplo, en los test utilizados en psicometría sobre reso- ı de problemas, puede ocurrir que si en la primera ocasión se ha resuelto, en unda se puede ir directamente a la solución, sin pasar por las etapas interme- . Por el contrario, la mayor parte de las variables que se miden en sociología nen esta problemática.

omparativamente existe un mayor número de publicaciones dedicadas a la dad entendida como equivalencia o consistencia interna que a la entendida estabilidad. Esto se debe a que tradicionalmente ha sido el campo de la Psi- ía la que en mayor medida se ha ocupado de estos problemas, y para calcular ɔilidad de sus test son apropiados los conceptos de «equivalencia» o «consis- ı interna». Sin embargo, el concepto de «estabilidad» que se recoge por el dimiento test-retest es más adecuado para la disciplina sociológica, que se ha ɔorado con algún retraso al estudio de esta problemática.

bilidad como equivalencia

: trata de disponer de dos o más formas paralelas de un mismo instrumento :dida, que permita obtener dos o más puntuaciones del mismo sujeto, aplica- multánemanente o dejando un lapso de tiempo entre ellas. Se supone que los ıs deberían responder de forma equivalente a estas muestras de ítems. La di- :ia en la respuesta será la medida del error y se llama coeficiente de equiva- o fiabilidad. Matemáticamente se expresa por la correlación entre las dos aciones.

ı cierta medida este procedimiento es similar al método test-retest, porque :re también aplicar dos test a las mismas personas, la diferencia es que no se .a segunda vez el mismo test, sino otro test alternativo, diferente, aunque se ıe que mide lo mismo que el primero.

ıs características que deben cumplir los instrumentos de medida paralelas

Deben tener el mismo número de elementos intercambiables uno a uno.
La redacción y la estructura de cada elemento debe ser idéntica en los elementos paralelos de ambas formas.
El contenido y el objetivo apreciado, elemento a elemento, deber ser el mismo.
Los índices de dificultad de los elementos deben ser iguales.
Las instrucciones dadas para la realización de la prueba, el tiempo asignado, las condiciones en que se aplica, ..., han de ser las mismas.
También deben ser idénticos los aspectos externos: presentación, formato,...

problema más grave que comporta este procedimiento es, sin duda, la cons- ón de las formas paralelas de manera que sean equivalentes. Así, las medias, sviaciones típicas, etc., verdaderas, de ambas formas deben ser idénticas y las ncias encontradas empíricamente, casuales o aleatorias.

atemáticamente el modelo de las pruebas paralelas es muy estricto en sus su- ɔs. Dos pruebas son paralelas si cumplen las siguientes condiciones:

1. Tienen idénticas medias, varianzas y covarianzas.
2. Ambas correlacionan en idéntico grado con las puntuaciones verdaderas [9].
   La correlación entre las dos pruebas paralelas indica la proporción de verdadera varianza, o varianza que corresponde a un único factor común (Linn, Werts, 1979). La unidimensionalidad, por tanto, está implícita en este modelo [10].
3. La varianza de cada prueba paralela no explicable por las puntuaciones verdaderas se debe a errores de medición.

La variable no experimenta cambio real, o en otras palabras, las puntuaciones verdaderas son siempre las mismas para cada sujeto; las diferencias entre las puntuaciones observadas se deben a que los errores de medición varían; los componentes verdadero y de error son independientes, en consecuencia la correlación entre los mismos es cero.

Este método no tiene en cuenta la inestabilidad que proporciona el tiempo como fuente de no fiabilidad, ya que se administran a los mismos individuos en una misma sesión dos formularios supuestamente equivalentes del mismo test y aunque los dos formularios contienen ítems distintos, todos ellos tratan de medir la misma característica. La correlación entre las puntuaciones de los dos formularios representa la medida de hasta qué punto los ítems miden la misma característica y son consistentes.

Tampoco se tienen en cuenta las fluctuaciones azarosas de las personas o de las situaciones en las que transcurre la administración, debido a que las dos formas se administran a un tiempo en una sola sesión.

Para Carmines y Zeller (1979), este método presenta las siguientes ventajas y desventajas respecto al método test-retest:

1) Tiene la ventaja de que reduce los problemas de memoria, que afectan a la fiabilidad en el método test-retest.
2) Al igual que el test-retest tampoco distingue los cambios verdaderos de los errores aleatorios, por lo que la fiabilidad es menor.
3) La desventaja respecto al test-retest es, como ya hemos visto, la dificultad de construir formas alternativas que sean paralelas: «si ya es difícil construir un test mucho más difícil lo es construir dos que midan exactamente lo mismo» (Carmines, Zeller, 1979, 41). Así, los problemas que plantea su utilización son los siguientes:

   3.*a)* La falta de equivalencia de los ítems muestreados, ya que aunque se construyen con la intención de que sean equivalentes, la formulación de los mismos puede dar origen a diferencias en las puntuaciones.

   3.b) La diferente experiencia y conocimiento que un sujeto puede tener en uno o en otro de los universos de ítems aunque sean equivalentes, ya que el sujeto puede entender mejor, o conocer mejor, unos aspectos que se reflejan en una aplicación que los que se reflejan en otra.

El procedimiento de los formularios alternos se desarrolló sobre los test de inteligencia y aptitud, donde es más facil medir una característica repetidamente por medio

[9] Esta condición es imposible de comprobarla.
[10] La unidimensionalidad se define en el sentido de que todos los ítems miden una sola y única dimensión.

formas paralelas. Pero la dificultad de conseguir construir dos formas equiva-
de una medida ha impedido una mayor utilización.
ı psicometría se considera un buen procedimiento para los instrumentos que
den apreciar la velocidad de ejecución (número de elementos resueltos por
l de tiempo). Se recomienda que entre las dos formas alternativas medie una
ción temporal de dos semanas.
pesar de la defensa que determinados psicometras hacen de este método, es el
ás inconvenientes presenta porque se parte de unos supuestos que es impro-
que se den en la práctica, razón por la cual no se utiliza mucho esta técnica
iología.

ɔilidad como consistencia

isten dos posibilidades: *a)* Se construye un instrumento de doble longitud, que
de partir en dos mitades, y todos los items miden el mismo rasgo o caracterís-
ɔr lo que debe darse una coherencia o consistencia en las respuestas de los su-
n los dos subconjuntos de la prueba; *b)* O bien se establece hipotéticamente el
so total de items que miden un rasgo o característica y se compara con la
ıción observada en una muestra de items; o bien se supone un número indefi-
e pruebas paralelas y se compara con la muestra obtenida. En ambos casos,
ıtarse de puntuaciones que se obtienen en la misma prueba, obtenemos una
ı de consistencia interna.

ɔ *de dos mitades*

aplica una sola vez, pero en la aplicación de la prueba se obtienen dos pun-
ıes para cada sujeto y nos permite calcular un coeficiente de fiabilidad, que
ıtarse de puntuaciones que se obtienen en la misma prueba, se llama coefi-
de consistencia interna.
-a obtener dos puntuaciones se divide el conjunto de elementos que integran
umento de medida en dos mitades equivantes respecto a la característica me-
n realidad, son dos subconjuntos, lo que puede hacer pensar que son dos for-
ralelas.
método de las dos mitades es similar al anterior de formas paralelas equiva-
de hecho, ambos implican los mismos supuestos y restricciones (unidimen-
dad, etc.). La diferencia con el procedimiento anterior es que allí cada for-
una prueba que se aplica independiente de la otra, mientras que aquí se
una sola prueba que después se parte en dos mitades. Por lo demás, consti-
na modalidad del procedimiento de equivalencia, ya que en este caso la
lencia de los resultados de diferentes muestras de items se estiman mediante
isis interno de las respuestas a los items en un solo test. Es lo que se deno-
ı «corrección de corte por la mitad» y que supone un caso especial de medi-
alelas o alternas.
forma más usual de obtener dos mitades es eligiendo los elementos pares y
ares, pero existen otras muchas, como incluir en la primera mitad los prime-
ns y en la segunda mitad, los segundos. Sin embargo, esta partición no per-
comparar los resultados, ya que la fatiga y tensión en la primera prueba
a a la segunda.

También se procede para igualar ambas mitades a considerar la dificultad de los ítems. La dificultad se determina dividiendo para cada elemento el número total de sujetos que los resolvieron correctamente por el de aquellos que lo intentaron. El valor oscilará entre 0 (no resuelto por nadie, difícil) y 1 (resuelto por todos, fácil). De esta forma se distribuyen los elementos a una u otra mitad. También es conveniente, a la vez, procurar un equilibrio entre ambas mitades en el contenido y los objetivos apreciados por los elementos que las integran.

En este procedimiento se puede demostrar que existe relación entre la longitud (número de elementos de la prueba) y la fiabilidad. A mayor número de elementos, mayor fiabilidad o consistencia interna. Ahora bien, esta relación no se da con cualesquiera elementos o items que se añadan, sino cuando son equivalentes a la que integran la prueba, de forma que la estructura y composición de la misma queda inalterada.

Existen tres procedimientos de cálculo para la estimación de la fiabilidad de las pruebas de longitud doble.

a) Ecuación de Spearman-Brown (Spearman, 1910; Brown, 1910), que fue descubierta al mismo tiempo de forma independiente por estos dos autores.

$$R_{xx} = \frac{2r_{xx}}{1 + r_{xx}}$$

R_{xx} = Coeficiente de fiabilidad total.
r_{xx} = Correlación entre las dos mitades (varía de 0 a 1)

b) Ecuación de Rulon

$$r_{xx} = 1 - \frac{S_d^2}{S_t^2}$$

c) Guttman

$$r_{xx} = 2\left(1 - \frac{S_{1a}^2 + S_{2a}^2}{S_t^2}\right)$$

d) Posteriormente se ha desarrollado la fórmula de Kuder-Richardson (1937). Permite como las anteriores obtener un coeficiente de consistencia interna, pero además un coeficiente de homogeneidad entre los elementos, ya que puede haber instrumentos con mitades consistentes, pero heterogéneos en sus elementos.

Se utiliza para instrumentos calificados en dos categorías (acierto, error; sí, no).

$$r_{xx} = \frac{n}{n-1}\left(\frac{S_t - \Sigma\, p\, q}{S_t^2}\right)$$

nero de elementos.
iente entre el número de sujetos que aciertan y el total.
- p.

inconveniente del método de dos mitades es que existen numerosas estima-
de la fiabilidad. Así, en una escala de 10 items, existen 125 formas de dividir
anto, obtener diferentes estimaciones de fiabilidad. La ventaja es que se apli-
sola vez a los mismos individuos (Carmines, Zeller, 1979).
dos mitades se consideran formas alternas del mismo test y el coeficiente
mide indica la consistencia interna del test. Un elevado coeficiente supone
situación del individuo no se halla determinada por el muestreo concreto de
los que ha contestado en cualquiera de las dos mitades.
un principio, las aplicaciones que se llevaron a cabo de este procedimiento
ue las dos mitades debían ser equivalentes, y cada una de ellas representara
dad del test (Guilford, 1954). Se comparan los items pares con los impares
te el coeficiente de correlación que es una estimación del coeficiente de equi-
1 de la mitad del test y a partir de la cual se puede calcular la fiabilidad de
test, por medio de la formula de Spearman-Brown.
embargo, estudios más recientes estiman que si todos los items del test tratan
ir la misma característica, deberían ser comparables cualesquiera dos mitades
calculen al azar y no dos que se consideren equivalentes. Los coeficientes que
calculado basándose en esta nueva concepción son el coeficiente alpha y la
20 de Kuder-Richardson. Estos dos índices expresan la correlación media
por la mitad para todos los casos posibles de división del test en dos partes.
rencia del procedimiento anterior (pares, impares) estos coeficientes de equi-
lo son para el test completo. La tendencia continúa en esta dirección y po-
ıfasis en el concepto de homogeneidad entre todos los items de un test. La
a es la siguiente: ¿hasta qué punto miden la misma característica todos los
un test? [11].

del Universo de items [12]

ıos indicado que una limitación importante del método de las dos mitades
odemos hallar numerosos coeficientes de fiabilidad. Así, el coeficiente halla-

nbién existen otros métodos de medición de la homogeneidad basados en la comparación de
a de las puntuaciones obtenidas de una medición con la varianza que resultaría si todos los
vieran no correlacionados y con la varianza que resultaría si todos los items estuvieran corre-
: (Scott, 1960).
; recientemente Crombach y colaboradores (1972) han ampliado el modelo original del uni-
ems. En el modelo original se controla como fuente de error el contenido de los items. En el
posterior de este modelo se amplía la interpretación del universo, que ya no es solamente de
., sino de situaciones, y en él se incluyen otras fuentes de error (tiempos, contextos, etc.). En
os modelos se controlan diversos tipos de errores y no hay un único coeficiente de fiabilidad,
er tantos como fuentes de error se analicen. Un resumen de este modelo lo lleva a cabo
1976).

do para la primera y segunda mitad es diferente del hallado para los pares e impares y así sucesivamente. Sin embargo, en el modelo de universo de items se obtiene un solo coeficiente.

Por otra parte, en el método de las dos mitades existen unos supuestos muy restrictivos. Por el contrario, en el modelo de la muestra de items procedentes de un determinado dominio, no supone necesariamente unidimensionalidad [13], y aunque de hecho se interprete que la tiene, no se exige explicitamente.

La dicotomía de puntuación verdadera y error se sustituye por el universo de items cuya suma definiría la puntuación verdadera. El coeficiente de fiabilidad indica si una puntuación observada (en una muestra de items) estima bien la puntuación verdadera (universo de items). En realidad es como si se supusiera un número indefinido de pruebas paralelas, que se forman aleatoriamente por items. La diferencia con las pruebas paralelas es que en ellas todos los items tienen idéntica varianza e idéntica correlación con el total del dominio de items, mientras que en las pruebas formadas aleatoriamente no se exigen estos requisitos. De esta forma y como afirma Morales (1988) el modelo de las pruebas paralelas viene a ser un caso particular del modelo del universo de items.

Los presupuestos del modelo de universo de items son los siguientes (Nunnally, 1978):

1/ Se supone que existe una población de items que pertenecen al dominio o constructo que deseamos medir; esta población o universo de items podemos considerarlo como hipotéticamente infinito.
2/ Cualquier instrumento, escala, test, etc., está compuesto por una muestra de k items, tomados aleatoriamente de ese conjunto hipotético de items; esta muestra puede ser tan pequeña o tan grande como se quiera.
3/ La puntuación verdadera de un sujeto es la que obtendría si respondiera a todos los items de la población; la puntuación observada es la que proviene de sus respuestas a una muestra de items, y es sólo una estimación de su puntuación verdadera.
4/ Las puntuaciones observadas serán fiables en la medida que exista una correlación alta con la hipotética puntuación verdadera; es decir, si el instrumento compuesto por k items tiene una correlación alta con el instrumento compuesto por todos los items. La fuente de error que se considera aquí es la que proviene de un deficiente muestreo de items, la que tiene que ver con el contenido de los items; existen otras fuentes de error, pero esta es la verdaderamente importante que se considera.
5/ Un elemento básico de este modelo es suponer que existe una matriz de correlaciones con las intercorrelaciones de todos los items del hipotético universo de items. La correlación media de esa matriz indicaría en qué grado todos los hipotéticos items tienen algo en común, que es precisamente lo que se pretende medir con la muestra de items. Lo que no se supone necesariamente es la existencia de un único factor. El modelo supone que las puntuaciones de los items se han tipificado, con lo que las varianzas de los mismos son idénticas (e iguales a la unidad).
6/ Las correlaciones de cada item con la suma de todos los demás deben ser iguales.

[13] No se identifica varianza verdadera con un único factor común (Morales, 1988).

partir de estos presupuestos Nunnally (1978) ha comprobado que:

) La correlación entre un test de k items (puntuaciones observadas) y el test compuesto por todos los items de la población (puntuación verdadera), es igual a la raiz cuadrada de la correlación media de una serie de test paralelos todos con idéntico número k de items procedentes de la misma población o dominio de items:

$$r_{kv} = \sqrt{r_{KF}}$$

Las fórmulas para hallar las puntuaciones típicas son:

$$r_{11} = (K/\ K\text{-}1)\ \ ((R\text{-}K)\ /\ R)$$

$$r_{11} = (K/\ K\text{-}1)\ \ (1 - (k/R))$$

es la suma de todas las correlaciones en la matriz, incluyendo los valores de la ıal que son las varianzas de los items.
K es la suma de las correlaciones menos las varianzas de los items.

Las fórmulas para hallar las puntuaciones directas que en la práctica es lo normal son:

$$r_{11} = (K/\ K\text{-}1)\ \ ((C - \Sigma\ \sigma_i^2)\ /\ C)$$

$$r_{11} = (K/\ K\text{-}1)\ \ (1 - (\Sigma\ \sigma_i^2/\sigma_t^2))$$

mero de items

matorio de la varianza de los items

rianza total

a es la fórmula más conocida como el coeficiente alpha de Crombach. Este a demostrado que su coeficiente «alfa» equivale a la fiabilidad media que se ría dividiendo el test en todas sus posibles mitades y aplicando la fórmula de an-Brown en cada posible división el test en dos mitades,.
:oeficiente alpha de Crombach es una generalización del coeficiente de Kuichardson (1937) para estimar la fiabilidad de las escalas dicotómicas. Así, si de items dicotómicos utilizaremos la fórmula de Kuder-Richardson 20, que ıso particular de la fórmula más general de Crombach (1951):

$$KR20 = N\ (N\text{-}1)\ \ [1 - \Sigma\ p_i\ q_i\ /\ \sigma_x^2]$$

nero de items dicotómicos.

porción de respuesta positiva a los items.

p_i

ianza total compuesta.

Tiene la misma interpretación que el coeficiente alpha: es una estimación de la correlación esperada entre un test y una forma alternativa hipotética que contiene el mismo número de items.

Actualmente, ya no se usa el método de las dos mitades (de Spearman-Brown a partir de la correlación entre dos formas paralelas del mismo test). Se ha impuesto el coeficiente alpha de Crombach (o Kuder-Richarson 20). Las ventajas son evidentes, como hemos expuesto con anterioridad.

El coeficiente alpha responde a un modelo conceptual muy claro y simple: la proporción de varianza verdadera (fiabilidad) es igual a la varianza compartida dividida por la varianza total. La varianza verdadera queda definida operativamente por la suma de las covarianzas: por lo que discriminan los items precisamente por estar relacionados unos con otros. Este «estar relacionados» es lo que con propiedad se llama también consistencia interna u homogeneidad. El problema es la confusión que se da entre estos términos y el de unidimensionalidad.

El coeficiente alpha indica en qué proporción discriminan los items precisamente por estar relacionados entre sí. La interpretación más clara del coeficiente es la que se limita a expresar lo que señala la misma fórmula: «es la proporción de co-variación, que si es grande implica relaciones claras entre los items» (Morales, 1988, 248). En este caso el término consistencia interna parece adecuado. Sin embargo, no esta tan claro que se pueda identificar consistencia interna y unidimensionalidad, a pesar de que es precisamente Crombach el que introduce la idea de un único factor.

El coeficiente de fiabilidad alpha se utiliza como criterio para evaluar hasta qué punto un test o escala está compuesto por items lo suficientemente homogéneos, como para justificar que su suma constituya una medida del constructo subyacente. Lo que se verifica no es la correlación con unas supuestas puntuaciones verdaderas (contenidas en un universo de items), sino cuánto hay de común en los items, ya que no todo es común.

Este coeficiente nos dice cuánto hay de interrelación, pero no cómo es esa interrelación. El cómo se relacionan las variables nos lo dirá la matriz de correlaciones y el análisis factorial, pero según Crombach, y en general según la teoría clásica, un test es poco fiable y sus items poco homogéneos, si una proporción apreciable de varianza se debe a que los items son distintos; lo que se busca es que las diferencias en las puntuaciones observadas provengan fundamentalmente de que los sujetos son distintos en aquello que se intenta medir, no de que los items sean distintos.

El coeficiente alpha tiene limitaciones si se interpreta como índice de unidimensionalidad. Por otra parte, no guarda una relación clara ni con la magnitud u homogeneidad de las correlaciones, ni con su estructura factorial, por eso se presta a interpretaciones equívocas. No puede interpretarse este coeficiente de forma automática como un índice de homogeneidad o de unidimensionalidad sin tener en cuenta algunas matizaciones. Así por ejemplo, una alta consistencia interna (fiabilidad) no implica unidimensionalidad, pero la unidimensionalidad sí implica una alta consistencia interna. Green y colaboradores (1977) han demostrado que pueden darse valores altos de alpha y una estructura pluridimensional. En efecto, un coeficiente de fiabilidad alto puede darse cuando una proporción grande de la varianza está determinada por varios factores comunes, pero en situaciones en las que la unidimensionalidad puede ser muy discutible: la varianza explicada por el primer factor puede ser mínima y los factores comunes pueden no ser comunes a todos los items; unos factores pueden tener relación grande con unos items y ninguna con otros, sin que haya realmente factores comunes en todos los items.

En conclusión, a pesar de que el coeficiente alpha tiene el atractivo de su clari-ıd conceptual y es el método más utilizado para hallar la fiabilidad como consis-ncia interna, también tiene algunas limitaciones [14].

1/ Dependencia del número de items. La fiabilidad aumenta al incrementar el número de items y las intercorrelaciones entre los mismos. Sin embargo, existen limitaciones porque no se puede incrementar el número de items indefinidamente (Magnusson, 1968). Esta dificultad la reconoció el mismo Crombach (1951) y la expresó con la siguiente analogía: un galón de leche no es más homogéneo que un cuarto de galón.
2/ Los presupuestos en los que se basa no se dan en la práctica. Se presupone en la teoría clásica de la fiabilidad que todos los items miden lo mismo y con la misma intensidad. Esto expresado matemáticamente significa que la correlación inter-items es idéntica a la media de todas las correlaciones y las varianzas idénticas. Pero lo cierto es que aunque los items midieran lo mismo, no lo hacen con la misma intensidad. Frecuentemente miden más de un constructo o dimensión y, además, unos y otros items, de manera desigual.

El hecho de que los items no midan exactamente lo mismo, ni con la misma in-ısidad, lo pone de manifiesto el análisis factorial, al identificar lo que podemos nominar subconstructos o factores (la pluridimensionalidad de la escala o test) y peso o correlación de cada item en cada factor. Por estas razones los últimos ances, tanto conceptuales como metodológicos, en el estudio de la fiabilidad están acionados con el análisis factorial.

Por último, nos vamos a referir al valor mínimo que debe de tomar el coeficien-alpha y sobre lo que hay que indicar que no hay normas, ni práctica común. Nun-lly (1978) propone 0,70. Guilford (1954) propone un coeficiente de 0,50 para in-stigaciones básicas. Pfeiffer y colaboradores (1976) dan orientaciones más nítidas ıiendo en cuenta el uso del instrumento, y ponen el límite en 0,85 si se van a ıar decisiones sobre individuos y de 0,60 para otros usos.

!.3. *Elección del tipo de coeficiente de fiabilidad*

La elección del tipo de coeficiente de fiabilidad a utilizar está relacionado en mer lugar con la naturaleza del rasgo o característica a medir. Así por ejemplo, las pruebas de conocimiento funciona bastante bien hallar la fiabilidad por el ɔcedimiento de dos mitades.

Los instrumentos destinados a resolver problemas, destrezas, encontrar una ley una serie, establecer analogías, completar series, etc, hacen poco conveniente el ɔcedimiento de la repetición e incluso, de las formas equivalentes. Sin embargo, instrumentos que miden tareas rutinarias o de atención, admiten bien los proce-ıientos de repetición. En este mismo caso se encuentran las características esta-s, como edad, sexo o estado civil.

[14] Armor tras reconocer las limitaciones del coeficiente alpha propone un nuevo coeficiente «The basado en el análisis factorial. Además de este coeficiente han surgido otros como el coeficiente ega de Heise y Bohrnstead para suplir algunos inconvenientes.

También influye en la elección del tipo de coeficiente de fiabilidad la finalidad a la que se destina. Si nuestro objetivo es predecir puntuaciones en un tiempo posterior, procede emplear la técnica de repetición.

Otra fuente que influye en la elección es la modalidad de la prueba. Así por ejemplo, si las pruebas son de velocidad hay procedimientos que dan lugar a coeficientes exageradamente altos, como es el caso de las dos mitades o de Kuder-Richardson, por lo que deben ser evitados.

A estas recomendaciones realizadas por R. Juste (1986) para la elección del procedimiento hay que añadir al menos una más: la disciplina en la que trabajamos y el tipo de variables que se utilizan. Si en psicología las herramientas de trabajo son test y escalas (Gomez-Bueno, 1991) y, por tanto, variables ordinales, en Sociología, son las encuestas que recogen variables nominales y, ocasionalmente, ordinales y de intervalo. Esto revierte ciertas peculiaridades que conviene tener en cuenta, ya que no son siempre los mismos problemas los que nos ocupan a los psicólogos y a los sociólogos.

Los psicólogos se han centrado fundamentalmente en el instrumento de medición porque construyen escalas, sobre las que han de garantizar la precisión y la fiabilidad. Las construcciones que se llevan a cabo en psicología implican que todos los items deben de medir lo mismo (inteligencia, autoritarismo,...). En contraposición, los sociólogos no se pueden centrar tan sólo en el instrumento de medición, ya que éste es un aspecto que influye al mismo nivel que otros como son los rasgos medidos, las características de los sujetos, las épocas del año, los estados transitorios, las condiciones ambientales, las personas que intervienen en la medición, etc. Las construcciones de los items que se llevan a cabo en sociología tienen un valor independiente. Así, las frases que empleamos para medir una actitud no tienen la exigencia de medir todas lo mismo, sino que cada frase mide literalmente lo que se dice.

3.3. *Relaciones entre la validez y la fiabilidad*

La fiabilidad es básicamente un asunto empírico, porque se ocupa del grado de consistencia de los resultados al repetir las medidas. La validez esta orientada más teóricamente, porque ha de contestar a la pregunta ¿válido para qué propósito?. Ahora bien, estas afirmaciones conviene matizarlas porque en ocasiones es muy difícil distinguir entre validez y fiabilidad, especialmente entre la validez teórica o de constructo y la fiabilidad entendida como homogeneidad de las medidas paralelas, y también en los contrastes de «consistencia externa», donde las mediciones y los conceptos se comparan con variables externas.

Una prueba puede ser fiable y no ser válida, pero no cabe pensar en la situación opuesta. Se puede efectuar una medida con gran precisión y estabilidad, pero si mide algo diferente de lo que se pretendía medir para nada nos servirá. Ahora bien, si la medición no es fiable, no podemos medir de forma satisfactoria.

Validez y fiabilidad estan íntimamente relacionadas, de forma que las ganancias o las pérdidas en una de ellas repercute en la otra. De ello no se desprende que la relación entre ambas sea homogénea ni se pueda representar en un continuo. Niveles altos de fiabilidad no llevan necesariamente a niveles altos de validez, ya que los resultados pueden tener errores constantes. Tampoco se puede asegurar que necesariamente niveles bajos de fiabilidad impliquen niveles bajos de validez.

En efecto, por una parte los niveles bajos de fiabilidad procedente de las inconsistencias entre los jueces u observadores reduce la validez de la medida; pero, por

parte, los niveles bajos de fiabilidad procedentes de la inestabilidad en los resul-
os del método test-retest puede o no reducir la validez. En este último supuesto,
característica que se mide es bastante estable y persistente, entonces, sí se cum-
que a menor fiabilidad menor validez, pero si la característica puede cambiar
el tiempo, lógicamente la inestabilidad de los resultados de una administración
a no implica necesariamente error y por tanto, puede no afectar a la validez.
También Loevinger (1954), en el caso de la relación entre la fiabilidad entendi-
omo consistencia interna y la validez, ha establecido que el aumento de la ho-
eneidad a partir de cierto límite puede hacer disminuir la validez.

Evolución de los términos

El término fiabilidad comenzó a utilizarse en los test de aptitud y se suponía que
nsistencia o inconsistencia en los resultados de repetidas mediciones, con el
o instrumento o con instrumentos similares, reflejaba la fiabilidad y, por tanto,
rrores de medición. La validez se concebía en relación a un sólo criterio y «se
aba poca atención tanto a la influencia de las características relativamente du-
as, pero extrañas, como a la posibilidad de que el propio proceso de medición
ra modificar la característica» (Selltiz,1976, 257).
medida que se fueron midiendo otras características menos estables como las
des o las opiniones, se observó que la variable que el intrumento medía podía
iar y, por tanto, las inconsistencias no se debían a errores de medición. Tam-
se observó que el instrumento podía reflejar otras características diferentes de
e se pretendía medir y en consecuencia, la consistencia de los resultados no te-
rqué reflejar necesariamente la ausencia de error.
lo referente a la validez, desde principios del siglo xx se esta escribiendo so-
la, especialmente de la validez de contenido y predictiva. En los años cincuen-
American Psychological Assotiation estableció tres tipos de validez: de conteni-
criterio (concurrente y predictiva) y de constructo (Matesanz, 1975). En los
iguientes a la publicación de estas recomendaciones técnicas de la A.P.A., sur-
merosos trabajos centrados sobre todo en el concepto de validez de construc-
nbach, Meehl, 1955).
concepto de validez de constructo nace de la limitación de la validez de con-
cuando se aplica a la medición de rasgos, actitudes, etc, es decir, a conceptos
tos (Nunnally, 1978). La validez de constructo resultó ser la más adecuada y
yor aplicación para las ciencias sociales (Zeller y Carmines, 1980) y su intro-
n operó cambios en los conceptos no solamente de validez sino también de
lad.
dicionalmente se había considerado la validez como el grado en que un ins-
to medía «realmente» lo que se proponía medir. Actualmente el concepto es
plio y se considera validez el grado en que una observación o medición con-
con su referente, lo que implica dar una mayor importancia a los resultados
edición en detrimento del instrumento con el que se mide. Lo más caracterís-
la validez tal y como se entiende actualmente es que no se refiere a una ca-
tica del instrumento, sino de las inferencias que se hacen de las puntuaciones
nstrumento. En otras palabras, no es el instrumento lo que es válido, sino sus
taciones.
recomendaciones técnicas que recoge la A.P.A. en 1974, hacen referencia a
odificaciones y agrupa los tipos de validez en dos grandes bloques: los que

tienen que ver con la naturaleza o significado del atributo (validez de constructo) y los que tienen que ver con el uso de los instrumentos como indicadores de otras variables (validez de criterio).

Posteriormente, se va asumiendo paulatinamente que es la validez de constructo la que engloba a los otros tipos de validez, que se ven como pruebas de la validez de constructo o como coeficientes de utilidad, o de validez de hipótesis específicas. Esto constituye un avance conceptual importante, ya que en la actualidad la validez de constructo es casi sinónima de validez en general [15].

La situación en la que nos encontramos hoy en dia es muy diferente entre la validez y la fiabilidad. Mientras en la validez no existe un método claro y unificado y no se ha impuesto ninguna estrategia en particular, en la fiabilidad sí existen los métodos y las fórmulas específicas que hemos descrito, lo que facilita la comparación de los resultados y de las intepretaciones.

4. Valoración final

Los análisis de validez y fiabilidad, tal y como los hemos descrito, estan relacionados en su surgimiento, evolución y aplicación con los análisis cuantitativos y concretamente con el análisis de la investigación por encuesta. Ha sido en la aplicación de esta técnica donde sistemáticamente se han estudiado y desarrollado los aspectos ligados a «los problemas de la medida», y donde ha existido una preocupación constante por el control y evaluación de los errores. Sin embargo, los análisis cualitativos (entrevistas en profundidad, grupos de discusión, estudios de campo,...) no han mostrado un gran interés por estos temas, ya que se asume el subjetivismo inherente a la investigación social y, por tanto, no existen procedimientos de análisis definidos y comparables [16].

En un extremo, se han sucedido las críticas internas de los cuantitativistas en sus exigencias de control, supervisión y medición de los sesgos en un intento de cientificidad y objetividad, que si ya es imposible para las ciencias físicas, mucho más lo es para las ciencias sociales.

En el otro extremo, también se han sucedido las críticas externas de las perspectivas cualitativistas. Estas, tienden a exagerar la significación de los «errores y sesgos», considerando superficialmente el hecho de que los científicos sociales dependen universalmente de los datos que han sido recogidos por medio de informes orales o escritos, y que estas observaciones y mediciones, con independencia de la forma en que hayan sido recogidas, se hallan sujetas invariable y esencialmente a las mismas fuentes de error y desviación que las recogidas por los entrevistadores de una encuesta.

El clínico, el experimentador o el observador participante, se hallan en el mismo peligro de desviar las respuestas de sus sujetos que los encuestadores. La diferencia fundamental es que cuando los científicos sociales tienen que depender de los informes de entrevistadores a los que han preparado y seleccionado, se dan cuenta mejor de los peligros y dificultades que van implícitos.

Podemos concluir que si por una parte, las posibilidades de controlar los errores,

[15] Son numerosos los autores que defienden este puesto central para la validez de constructo: Messick (1980), Tenopyr (1977), Anastasi (1986).

[16] Ambas corrientes obedecen a dos tradiciones que en sociología se conocen como «comprensión» y «explicación» y que actualmente parece que convergen mediante el pluralismo metodológico (Latiesa, 1991).

dir con precisión las características y de demostrar leyes son muy modestas, no lo, y digámoslo una vez más, hemos de abandonar la idea de intentarlo. No os efectivamente ante una ciencia exacta: hay que tomar decisiones valorati- ay que tener en cuenta las limitaciones de los métodos para interpretar los existen diversas alternativas metodológicas y los conceptos son imprecisos. precisamente por estas mismas razones, podemos y debemos comprobar que ervaciones obtenidas con los métodos de medición son razonablemente vali- fiables y que se encuentran suficientemente respaldadas por construcciones s y por constrastaciones empíricas.

ncias bibliográficas

. y otros
: *Los dos métodos en las ciencias sociales.* CIS.
: «Diseños de investigación», en Latiesa, M. (ed): *El pluralismo metodológico en la investigación social.* SP Universidad de Granada.

: «Theta reliability and factor scaling», en COSTNER, H. L. (ed): *Sociological methodology 1973-1974*, Jossey-Bass, San Francisco.
H. M.
: «The measurement problem», en BLALOCK, H. M.; BLALOCK, A. (ed.): *Methodology in social research*, Mcgraw Hill, New York, pp. 5-27.

: *What is Science*, Dover, New York.
«Psychometric theory», en DUNETTE, M. E. (ed.): *Handbook of industrial and organizational psychology*, Ran Mc Nally, Chicago.
D. T.; FISKE, D. W.
«Convergent and discriminant validation by the multitrait-multimethod matrix», *Psychological Bulletin*, 56, 81-105.
E. G.; ZELLER, R. A.
Reliability and validity assessment series: quantitative applications in the social sciences, Sage, papers 17, London.
; KENISTON, K.
«Yeasayers and naysayers: Agreeing response set as a personality variable», *Journal of Abnormal and Socialk Psychology*, 60, 151-174).
L. J.
«Further evidence on response sets and test design», *Educational and Psychological Measurement.* 10, 3-31.
«Response sets and test validity», *Educational and psychological measurement*, 6, pp. 475-494.
Essentials of psychological testing, Harper, New York.
«Coefficient alpha and de internal structure os tests», Psychometrika, 16, pp. 297-334.
«Beyond the two disciplines of scientific psychology», *American Psychologist*, 30, pp. 116-127.
«What price simplicity?», *Educational measurement issues and practice*, 2,2, pp. 11-12.
«Balancing the qualitative and the qualitative in psychological research», *Psychological assessment*, 2, pp. 3-12.
P.; MARLOWE, D.
The approval motive: Studies in evaluative dependence, N. Y., Wiley.
. L.
Techniques of attitude scale construction, N. Y., Appleton-Century-Crofts.
S.

1978: «Analyzing one variable-three wave panel data: a comparison of two models», *Political methodology*, 5, pp. 221-231.
FISKE, D. W.
1973: «Can a personality construct be validated empirically?», *Psychological Bulletin*, 80, 89-92.
GARCÍA FERRANDO, M.
1985: *Socioestadística*, Alianza.
GÓMEZ BUENO, C.
1991: «Escalas: Problemática subyacente», en LATIESA (ed.): *El pluralismo metodológico en la investigación social. Ensayos típicos*, Centro de publicaciones de la Universidad de Granada, Universidad de Granada.
GREEN, LISSITZ, MULAIK
1977: «Limitations of coefficiente alpha as a index of lext unidimensionality», *Educational and psychological measurement*, 37, pp. 827-838.
GUILFORD, J. P.
1954: *Psychometric methods*, Nueva York, McGraw-Hill.
HEISE, D. R.
1969: «Separating reliability and stability in test-retest correlations», *American Sociological Review*, 34, pp. 93-101.
HYMAN, H.
1971: *Diseños y análisis de las encuestas sociales.* Amorrotu.
HYMAN, H.
1972: *Secondary analysis of sample surveys: Principles, procedures and potentialities*, New York, Wiley).
KERLINGER, F. N.
1975: Investigación del comportamiento, Interamericana.
KUDER, G. F.; RICHARDSON, M. W.
1973: «The theory of the estimation of test reliability», *Psychometrika*, 2, pp. 151-160.
LINN, R. L.; WERTS
1979: «Covariance structures and their analysis», en TRAUB, R.: *New directions for testin and measurement: methodological developments*, 4, pp. 53-74, Jossey-Bass, San Francisco.
LOEVINGER, J.
1954: «Effect of distortions of measurement on item selection», *Educational and Psychological Measurement*, 1954, 14, 441-448)
MAGNUSSON
1968: *Teorías de los tests*, Trillas, México.
MANHEIM, H. L.
1986: *Investigación sociológica*, Ed. Ceac, Barcelona.
MORALES, F.
1981: *Metodología y teoría de la Psicología*, col. Psicología, núm. 2, F. Filosofía y Letras, Madrid.
MORALES, P.
1988: *Medición de actitudes en psicología y educación*, Ed. Itarttalo, San Sebastián.
NUNNALLY, J. C.
1964: *Educational measurement and evaluation*, McGraw Hill, New York.
NUNNALLY, J. C.
1978: *Psychometric theory*, McGraw Hill, New York.
PFEIFFER; HESLIN; JONES
1976: *Instrumentation in human relations training*, La jolla, University of California.
KONING, R.
1973: Tratado de sociología empírica, Madrid, Tecnos.
REPETTO, E., y otros
1977: *Pedagogía experimental*, Uned, Madrid.
RILEY, M. W.
1963: *Sociological research: A case approach*, Harper Brace Jovanovich, New York.
SCOTT, W. A.
1960: «Measures of test homogeneity», en *Educational and Psychological Measurement*, 1960, núm 20, 751-757).