INTIYAN
EDICIONES CIESPAL
521

Métodos de investigación social

Paulina Salinas Meruane
Manuel Cárdenas Castro

Quito - Ecuador
2009

Métodos de investigación social

Primera Edición

AV. Angamos 0610,Antofagasta, Chile
Telefax:(56)(55)355824 / 355826
E-mail: www.periodismo.ucn.cl
ISBN: 978-956-287-266-9

Segunda Edición

1.000 ejemplares - Marzo 2008

ISBN: 978-9978-55-070-0
Código de barras 978-9978-55-070-0
Registro derecho autoral Nº 030584

Portada y Diagramación
Diego Acevedo

Impresión
Editorial "Quipus", CIESPAL
Quito-Ecuador

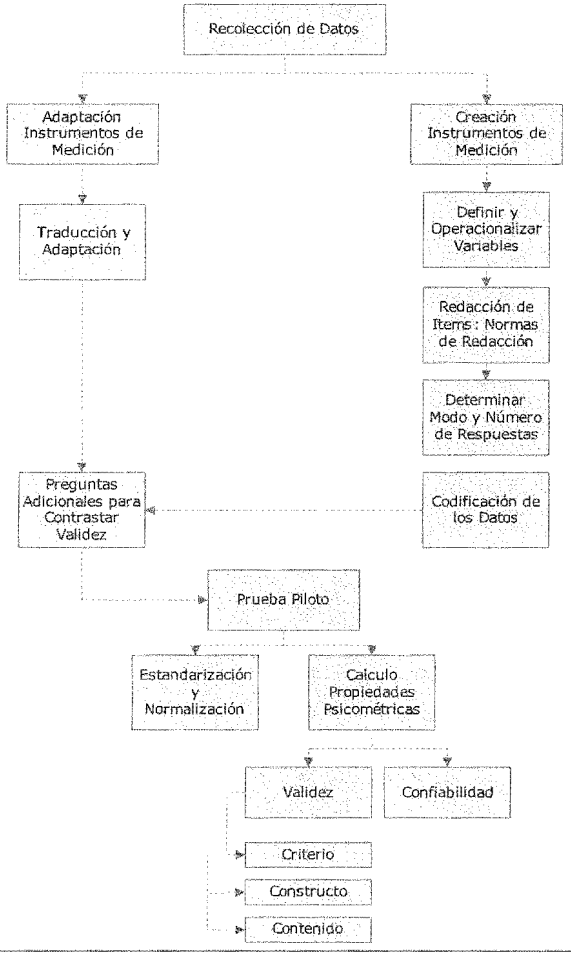
ÍNDICE

Primera Parte
Diseños de Investigación Cuantitativa

Segunda Parte
Diseños de Investigación Cualitativa

CAPÍTULO 7

Recolección de Datos



Capítulo 7

Construcción y Validación de Instrumentos de Medida para la Recolección de Datos

Manuel Cárdenas Castro

En este capítulo presentamos los pasos esenciales para adaptar y construir instrumentos (escalas) de medida que nos permitan recolectar datos sobre el problema de investigación que hemos definido. Se trabaja sobre los requisitos psicométricos que un instrumento debe reunir (confiabilidad y validez) y sobre los métodos para determinarlos. Nos centramos principalmente en los pasos que permitan construcción de escalas de medida y en el modo en que se deben de codificar e introducir los datos en la planilla de modo de dejarlos preparados para los posteriores análisis estadísticos.

Palabras clave: *Escalas de medida, confiabilidad, validez.*

7.1. Introducción

En términos generales, cabe afirmar que construimos instrumentos de recolección de datos para medir algún constructo o evento de la realidad. Medir, consiste en un proceso mediante el cual asignamos números a determinadas características o rasgos de los objetos de la realidad siguiendo unas determinadas reglas (Stevens, 1951). De forma más precisa, medir consiste en generar

correspondencia entre dos sistemas de relaciones, uno empírico (el de las propiedades que deseamos medir) y otro formal o abstracto (el número que asignamos al evento que medimos). Las relaciones presentes en el sistema formal deben quedar adecuadamente representadas en el sistema empírico (Pardo y San Martín, 1994).

Desde este punto de vista, un instrumento de medición adecuado será aquel que registre determinados rasgos de un objeto (físico o social) que representen los conceptos o variables que el investigador tiene en mente (Hernández, Fernández y Baptista, 1998).

La aplicación de instrumentos de medida para indagar en las variables que nos interesa estudiar es generalmente un requisito en la investigación cuantitativa, ya que solo un adecuado material de recolección es lo único que nos permite acercarnos rigurosa y fielmente a las variables que estamos observando.

En este sentido, en este apartado pretendemos presentar una guía práctica para la construcción y validación de escalas o instrumentos de medida que oriente a quienes deseen adaptar o crear nuevos instrumentos que se ajusten de mejor manera a los objetivos de su investigación. Eso sí, antes de comenzar con esta guía debemos revisar algunos conceptos que serán de uso frecuente en ella.

7.2. Test y Escalas

En ciencias sociales es habitual distinguir entre los test y las escalas. Los primeros han sido diseñados para medir habilidades, rasgos de personalidad, valores e intereses. Son generalmente de respuesta cerrada y permiten comparar a un individuo con un grupo, situándolo en referencia a una norma. Las escalas suelen ser diseñadas para medir actitudes y suelen ser en formato de respuesta graduado de modo de situar a los sujetos o grupos dentro

de un continuo. Además de los test y las escalas, nos encontramos con los cuestionarios, que son instrumentos en que cada pregunta tiene un valor independiente, de modo que los deferentes reactivos o ítems del mismo no se suman ni generan una puntuación total. Un ejemplo de estos son los clásicos sondeos de opinión y los cuestionarios de caracterización sociológica (a los cuales no nos referiremos en este apartado).

En términos generales, los test y las pruebas psicológicas (que aquí utilizamos como sinónimos) han sido definidos como medidas objetivas y estandarizadas de una muestra de conducta (Anastasi y Urbina, 1998). Esta medida sería apropiada si logra mostrar una correspondencia entre la conducta que es medida y el desempeño del sujeto en otra área. Por ejemplo, es esperable que una persona que obtiene un alto coeficiente de inteligencia obtenga también buenas calificaciones por su desempeño académico, o que una persona que muestra un adecuado nivel de desarrollo para su edad en un test, muestre un desempeño cognitivo acorde a su nivel de desarrollo.

Que un test sea una medida estandarizada supone que los procedimientos para su aplicación y corrección deben de ser uniformes. Es decir, que existan reglas claras para proceder a su aplicación y corrección que permitan que siempre, independientemente de la persona que funcione como examinador, lleguemos a los mismos resultados. Gracias a esta estandarización es que los resultados de dos personas diferentes se hacen comparables, ya que las condiciones en que se han recopilado los datos son similares.

Ahora bien, el objetivo de los test y las pruebas de inteligencia es comparar a los sujetos con determinados indicadores que funcionan como referencia. Se llama normalización al proceso mediante el cual se consigue comparar los resultados de un determinado sujeto con los de otras personas equiparables pertenecientes a un grupo de referencia o, lo que es lo mismo,

con el desempeño normal de estos (desempeño promedio). En este sentido, podemos señalar que existen al menos dos tipos de normas que nos sirven de criterio de comparación: las normas intragrupo y las normas de desarrollo.

Con las denominadas normas intragrupo se compara el desempeño del individuo al que se le ha aplicado el test o la prueba con el desempeño de su grupo de referencia. Por ejemplo, se compara el desempeño de un niño con el de sus compañeros de curso y se constata su ubicación relativa dentro de dicho grupo. De esta forma, una persona será alta o baja en la variable que se está midiendo teniendo como criterio de comparación el grado en que esa variable se encuentra presente en su propio grupo de referencia.

Para el caso de las denominadas normas de desarrollo, lo que hacemos es comparar al sujeto con el patrón de desarrollo normal y ver cuánto ha progresado con referencia a este. Por ejemplo, si se dice de un sujeto que tiene una edad mental de 15 años si su ejecución en una prueba de inteligencia es similar a la de las personas de 15 años.

En necesario dejar consignado que cualquier norma debe suponerse siempre relativa, ya que se restringe únicamente a la población normativa sobre la que se tomó la muestra. Es decir, las comparaciones que se establecen solo son válidas dentro de una población determinada de la que la muestra utilizada es representativa. El proceso de construcción de normas será revisado más adelante.

Respecto de las escalas de medida podemos afirmar que estas se han utilizado principalmente para medir actitudes, aunque se debe señalar que es posible utilizarlas para medir una serie de rasgos (significados, representaciones, etc.). Se trata de uno de los métodos más utilizados en investigación social en virtud de su simplicidad (la cual no está reñida con la rigurosidad). El

supuesto que está a la base es que una serie de respuestas a un conjunto homogéneo de ítems (todos apuntan a medir el mismo rasgo), referidas a un determinado objeto, sitúan al sujeto dentro de un continuo en la variable medida.

Una escala refiere a un procedimiento de recolección de datos que tiene como característica central el uso de respuestas graduadas en que se ofrece a los sujetos una serie de reactivos con los que pueden, de forma previsible, estar de acuerdo o en desacuerdo. Se trata de un procedimiento por el cual el sujeto aporta directa y explícitamente los datos que se le solicitan. Tienen como características centrales el situar los estímulos dentro de un continuo (se determina cuán valorado es un determinado objeto dependiendo de la ubicación de su opción en la escala graduada), situar a los sujetos dentro de un continuo (cuánto de la variable que se mide posee el sujeto y en qué lugar del continuo del rasgo se encuentra), ofrecer alternativas dentro de cada ítem para que los sujetos manifiesten su grado de acuerdo con los mismos.

- Existen diferentes tipos de escala de medida y si bien la descripción del proceso de construcción que ofrecemos más abajo es de carácter genérico, nos referiremos brevemente a las más conocidas de ellas: las escalas diferenciales (Thurstone, 1928), escalas sumativas (Likert, 1932), escalas acumulativas (Guttman, 1944) y diferencial semántico (Osgood, Suci y Tanembaum, 1976).

- *Escalas diferenciales*: este tipo de procedimientos supone que las opiniones o expresiones de las personas serían sus propias actitudes verbalizadas. Estas opiniones variarían en el grado de intensidad y por lo mismo expresarían mayor favorabilidad o desfavorabilidad. Estas escalas, también denominadas de "intervalos aparentemente iguales" (Thurstone, 1928; Guilford, 1954; Nunnally, 1978), están formadas por una serie de afirmaciones que expresan mayor

o menor grado de favorabilidad respecto del objeto de actitud. El grado de intensidad es determinado por jueces expertos que evalúan uno a uno los ítems de la escala (generalmente en un rango que va de 1 hasta 9) y es representado por el valor de la mediana. Cuando los sujetos responden a la escala se limitan simplemente a escoger aquellas afirmaciones con las que están de acuerdo y su puntuación corresponde a la suma simple de los valores signados a cada ítem. En este sentido, este tipo de escalas no ofrece respuestas graduadas (que era una de las características que definían las escalas más al uso). Entre las limitaciones de este tipo de procedimientos se han señalado el arduo proceso de construcción y la interferencia de las propias actitudes de los jueces a la hora de valorar y asignar puntuaciones a los ítems (Morales, Urosa y Blanco, 2003). Finalmente, este tipo de escala no verifica la unidimensionalidad, por lo que desconocemos si todos los ítems son indicadores del mismo rasgo.

- *Escalas sumativas*: se trata de las más utilizadas en investigación social debido a la simpleza en su construcción y a las buenas propiedades psicométricas que se han reportado. Por otra parte, el valor de cada ítem no depende de las valoraciones de jueces, sino de las propias respuestas del sujeto, lo que las hace un mejor índice de la actitud que se mide. En esta escala la unidimensionalidad es corroborada (por medio de los análisis de correlación ítem-total). La escala en sí consiste en un conjunto de afirmaciones con las que se puede estar más o menos de acuerdo, entregando por ello una escala graduada (que generalmente tiene 5 ó 6 opciones de respuesta que van desde "totalmente de acuerdo" hasta "totalmente en desacuerdo"), lo que permite conocer la intensidad del acuerdo y la dirección de la actitud. Además, la actitud no es medida solo con una pregunta, sino con una serie de ellas que refieren al mismo constructo o a dimensiones

diferentes del mismo constructo. Algunos autores han señalado que su principal limitación es que sujetos distintos pueden llegar a una puntuación total igual aun contestando de forma distinta a las diferentes preguntas (Morales, Urosa y Blanco, 2003).

- *Escalas acumulativas*: pretenden, ante todo, encontrar la unidimensionalidad de la actitud. Es decir, un sujeto que manifiesta acuerdo con una afirmación deberá estar igualmente de acuerdo con afirmaciones más moderadas referidas a la misma actitud. Es decir, los reactivos de estas escalas pueden ordenarse en un continuo de intensidad o dificultad de aceptación (Anastasi y Urbina, 1998). De este modo, la gran dificultad consiste en encontrar un conjunto de ítems que conformen una secuencia ordenada que supongo que escoger un ítem implica la aceptación de todos aquellos de menor intensidad. Entre las principales limitaciones señaladas para este tipo de escala se encuentra la dificultad de la construcción de una gradación de los niveles de intensidad (Morales, Urosa y Blanco, 2003) y de los niveles de dificultad (Nunally, 1978; Kline, 1993), así como su dependencia a las variaciones muestrales (la intensidad lograda en una muestra puede variar con el tiempo o de una muestra a otra) y los estrechos límites a los que somete una actitud (ya que no logra medir diferentes manifestaciones de la actitud).

- *Diferencial semántico*: en este tipo de escala los ítems están compuestos por pares de adjetivos polares u opuestos, los que se le ofrecen al sujeto con una escala que permite que este evalúe un determinado objeto. Se trata de un instrumento de fácil construcción y que tiene como un requisito la existencia de una verdadera polaridad entre los pares de adjetivos. Se trata de una combinación de asociaciones controladas y procedimientos de escala en la que se presenta al sujeto un concepto para que sea evaluado

por medio de un conjunto de escalas de adjetivos bipolares. La tarea consiste en parificar el adjetivo al objeto e indicar la intensidad y dirección de la asociación (Osgood, Suci y Tannembaum, 1976; Cárdenas, 2006). En el diferencial semántico suelen aparecer a lo menos tres dimensiones dentro de las cuales saturan los ítems: evaluación (bueno-malo), potencia (fuerte-débil) y actividad (activo-pasivo). Su principal limitación refiere a que los significados asociados a un objeto varían de forma contextual, es decir serían sensibles a las variaciones del contexto y a la familiaridad del objeto evaluado.

7.3. Fiabilidad y Validez

Dos son los requisitos fundamentales que debe cumplir toda escala de medición o test: debe ser confiable y válido.

La confiabilidad (o fiabilidad) se refiere al grado de consistencia que nos otorga un instrumento para medir de modo preciso y sin error aquello que se desea medir. De esta forma, una escala confiable permitirá obtener mediciones similares cuando aplicamos la escala nuevamente a los mismos sujetos (estabilidad de la medición), de modo que sus elementos (ítems) serán consistentes para medir la misma propiedad (los elementos serán homogéneos) y sin error. De esta forma, podemos afirmar que un instrumento será fiable si cada vez que mide a los mismos sujetos obtiene los mismos resultados. Como vemos, la fiabilidad tiene dos aspectos complementarios: la consistencia interna y la estabilidad temporal (Pardo y Ruiz, 2002). Existen diferentes formas de obtener la confiabilidad: medida de estabilidad (test-retest), formas paralelas, división por mitades y coeficiente alfa de Cronbach.

- *Medida de estabilidad (test-retest)*: en este procedimiento el mismo instrumento se aplica dos o más veces al mismo grupo mediando un lapso de tiempo. Si correlación entre las aplicaciones es alta y positiva el instrumento es confiable.

- *Formas paralelas*: en este caso se administran dos versiones equivalentes del instrumento. Estas versiones deben ser similares en contenido, duración e instrucciones. Nuevamente la medida confiabilidad se obtiene producto de correlacionar las dos formas de la prueba.

- *Método de mitades partidas (split-halves)*: este método requiere una sola aplicación, ya que se divide la muestra en dos mitades y se las compara. Las dos mitades deben ser similares y el coeficiente de correlación obtenido debe ser alto. Este método es sensible a la cantidad de ítems del instrumento y por regla general a mayor cantidad de ítems mayores deberían ser los niveles de confiabilidad.

- *Coeficiente alfa de Cronbach*: este procedimiento requiere duna sola aplicación y no se hace necesario dividir los ítems en mitades. Valora la consistencia de la escala a partir de la correlación inter-elementos promedio (correlación existente entre todos los ítems de la escala).

La validez, por su parte, refiere al grado en que un instrumento mide adecuadamente la variable que dice medir y no otros aspectos diferentes de los pretendidos. Es este concepto el que nos indica lo que puede inferirse a partir de los resultados, pero debemos adelantar desde ya que se trata de un concepto complejo que depende de una serie de indicadores de diverso tipo, por lo que es imposible referirse a la validez como un índice único que pueda apreciarse como alto o bajo (como ocurría con la confiabilidad). Este concepto nos informaría sobre la capacidad de la escala para discriminar entre sujetos que tienen distinto nivel de la característica medida, para evaluar la concordancia entre las mediciones obtenidas por la escala y otros instrumentos (así como por la concordancia con valoraciones hechas por expertos) y la agrupación de los ítems en dimensiones similares a las propuestas por los autores. La

validez es un concepto del que puede obtenerse diferentes formas de evidencia: relacionada con el contenido, con el constructo y con el criterio.

- *Validez de contenido*: se refiere al grado en que instrumento incluye todos los contenidos que debe medir, es decir el grado en que refleja un dominio específico de contenidos. Para aportar evidencia sobre este tipo de validez se debe revisar exhaustivamente cómo ha sido utilizada la variable por otros investigadores. La elaboración de reactivos debe realizarse sobre la base de dicha revisión y debe tener en cuenta una adecuada extracción de dichos ítems por parte de expertos en el tema.

- *Validez de constructo*: se refiere al grado en que la estructura factorial arrojada por los análisis es coincidente con la prevista en el constructo teórico y en que la medición se correlaciona consistentemente con otras mediciones de acuerdo con hipótesis derivadas teóricamente y que conciernen a los constructos que están siendo medidos.

- *Validez de criterio*: en este caso el instrumento se compara con un criterio externo y se comprueba el grado de correlación existente entre nuestro instrumento y otros que miden lo mismo, o entre este y variables que miden algo equivalente. Si el criterio se fija en el presente, se habla de validez concurrente; si se fija en el futuro, se habla de validez predictiva.

Decíamos anteriormente que la confiabilidad y la validez son requisitos fundamentales de cualquier instrumento. Ahora bien, existen ciertos factores que pueden afectar dichos índices de un instrumento de medición y que por lo mismo no nos permitirán lograr el objetivo propuesto al momento de construirlo. Estos factores dicen relación con la improvisación (un adecuado

instrumento requiere que se le dedique tiempo y rigor a la hora de su construcción), la utilización de instrumentos que no han sido validados en nuestro contexto (salvo el proceso de construcción de ítems, los pasos para la validación de instrumentos son fundamentalmente los mismos que los realizados en la construcción de uno), falta de empatía del instrumento (el instrumento resultaría inadecuado para la muestra debido a los usos de lenguaje que utiliza o por no tener inconsideración variables relevantes tales como sexo, edad, grado de conocimientos, etc.) y factores debidos a las condiciones de aplicación del instrumento (ya sea por las condiciones físicas, por lo extenso del instrumento o por problemas mecánicos como la mala lectura de instrucciones).

7.4. Procedimiento para la construcción de escalas

Antes de comenzar con la descripción de los pasos necesarios para la construcción de una escala, debemos referirnos a las razones para construir una escala de varias preguntas y no limitarnos a una sola de ellas (cuestión que también es factible y que de hecho se hace). La primera razón para ello es que nos permite lograr una mejor descripción del constructo que se analiza, sobre todo cuando se trata de objetos complejos, y de la opinión del sujeto acerca de este sobre la base de múltiples indicadores. De este modo, no caemos en la excesiva simplificación del concepto que intentamos medir y podemos dar adecuada cuenta de su complejidad.

Por otra parte, la elaboración de una escala con varios ítems nos permite obtener un mayor grado de validez. Lo anterior debido a que en una pregunta (o pocas de ellas) nuestra medida queda muy expuesta a problemas de redacción o comprensión. Si la pregunta no es claramente entendida y si esta no logra dar cuenta de las múltiples dimensiones implicadas en el concepto, entonces los resultados podrían estar midiendo algo diferente a lo que esperábamos o lo harían de forma poco clara. De este modo una

amplia muestra de preguntas nos permite formarnos una mejor opinión de las actitudes de un sujeto.

Disponer de una escala formada por varios ítems que sean indicadores del mismo rasgo permite aumentar la confiabilidad (Morales, Urosa y Blanco, 2003) debido a que se minimizan las limitaciones de cada ítem y debido al hecho de que varias preguntas merecen más confianza que una sola, otorgándole mayor precisión a la medición.

Finalmente, el mayor número de preguntas permite que la escala diferencie mejor entre las posiciones de varios sujetos, lo que aumenta la varianza, y nos entrega información más detallada y nítida que permita ordenar a los sujetos dentro del continuo medido por la variable. Es por ello que sumar varios indicadores puede ser una buena idea si lo que se desea es comparar personas o grupos.

Ahora ya podemos comenzar a listar la serie de pasos que nos permitirán construir una escala de medición y analizarlos con mayor detalle. La secuencia de pasos que debe realizarse consiste en: 1) Definir y operacionalizar la variable que se desea medir (construir la tabla de especificaciones); 2) Redactar los ítems; 3) Determinar el modo de respuesta y el número de estas; 4) Indicar cómo se codificarán los datos (clave de corrección); 5) Preparar preguntas o instrumentos adicionales para aportar evidencias de validez; 6) Obtener datos de una muestra (Pilotaje); 7) Realizar cálculos de confiabilidad; 8) Seleccionar los ítems definitivos en función de los cálculos anteriores; y, 9) Realizar las comprobaciones de validez. A continuación analizaremos por separado esta secuencia, aunque debemos consignar que según los autores que se revisen esta serie de pasos puede variar, condensarse o expandirse, aunque de todos modos, y salvo algunas leves modificaciones, todos los autores indican básicamente lo mismo.

7.4.1. Definir la variable que se desea medir

El primer paso siempre consistirá en seleccionar el rasgo que se desea medir e intentar definirlo con la mayor precisión que nos sea posible. Es un paso que puede parecer simple a primera vista, pero es fundamental debido a que la coherencia de todo el resto del proceso depende la claridad conceptual que logremos alcanzar aquí. Es por ello que el conocimiento claro de las dimensiones teóricas que componen la variable analizada es fundamental.

Concretamente, lo que proponemos en este paso es generar una tabla en la que se especifique la variable (la que debe ser definida conceptualmente) y las dimensiones que la componen. Posteriormente a la redacción de ítems, estos se incorporarán a la tabla de forma de agrupar adecuadamente nuestro material (el que nos servirá como modelo teórico de comparación con los datos arrojados en los análisis de comprobación de la validez de constructo).

Para cumplir adecuadamente con los requisitos anteriores se debe de realizar una acuciosa búsqueda de información sobre la variable que analizamos, sobre los modos en que ha sido medida, sobre las evidencias disponibles que nos informan sobre sus dimensiones y los indicadores de estas, así como de los instrumentos al uso para medir dicha variable.

Cuando trabajamos sobre la adaptación y validación de una escala construida por otro autor este paso se simplifica bastante ya que en el reporte de dicha escala se explicitará la teoría de base y contextualizará dentro de un marco teórico que le dé sentido.

Además, incluirá las dimensiones de la variable y los ítems que corresponden a estas.

Tabla 1. *Ejemplo de Tabla de Especificaciones*

Variable	Definición	Dimensiones	Ítems
Prejuicio Sutil	Son nuevas expresiones del prejuicio que serían más indirectas y racionalizadas, en las cuales los sentimientos de hostilidad han sido reemplazados por otros de incomodidad, inseguridad, disgusto y miedo (Dovidio y Gaertner, 1986), así como por una dificultad para expresar emociones positivas hacia los exogrupos (Pettigrew y Meertens, 1995).	1. Defensa de los valores tradicionales e idea de que los exogrupos no los estarían respetando.	1. En Chile existen grupos que salen adelante por su propio esfuerzo. Los inmigrantes deberían hacer lo mismo sin que se les de un trato especial. 2. El inconveniente de que los inmigrantes se introduzcan en ciertos lugares (departamentos, hospitales, etc.) es que no saben respetar las normas de convivencia establecidas. 3. El hecho de que los inmigrantes no salgan adelante, es porque enseñan a sus hijos valores y destrezas que no son las adecuadas en esta sociedad. 4. Lo cierto es que si los inmigrantes se esforzaran un poco podrían estar, al menos, tan cómodamente como los ciudadanos chilenos.
		2. Exageración de las diferencias culturales y su uso para justificar la posición del exogrupo.	5. Pienso que los inmigrantes son muy diferentes a los chilenos en sus ideas y valores sexuales o en su práctica sexual. 6. Por lo que he podido ver, los inmigrantes son muy diferentes de los ciudadanos chilenos en su forma de hablar y comunicarse con la gente. 7. Los inmigrantes se diferencias mucho de los chilenos en los hábitos de higiene y la necesidad de limpieza. 8. Por lo que conozco, los inmigrantes son muy diferentes de los ciudadanos chilenos en los valores que enseñan a sus hijos.
		3. Negación de emociones positivas hacia el exogrupo.	9. A menudo he sentido admiración hacia los inmigrantes. 10. Con frecuencia he sentido compasión por la situación en que se encuentran los inmigrantes en nuestro país.

En el ejemplo anterior podemos apreciar cómo puede elaborarse una tabla de especificaciones que nos permita orientar el proceso de construcción de una escala y comparar posteriormente nuestras predicciones teóricas con los análisis factoriales sobre los resultados obtenidos (de modo de aportar indicaciones de validez de constructo). Como se ve, lo que hemos ganado en este paso es la definición conceptual y una operacionalización de la variable que nos interesa (en este caso prejuicio sutil), así como listar las dimensiones que la conforman. Posteriormente, cumplimentado el paso siguiente (redacción de ítems) los incorporaremos a nuestra tabla de especificaciones.

7.4.2 Redacción de ítems

Los criterios o normas para redactar ítems pueden resumirse como sigue: los reactivos de una escala de medida deben ser afirmaciones relevantes, claras, discriminantes y contrastadas sobre una variable sobre la cual los sujetos manifiestan su acuerdo o desacuerdo. Analicemos con más detalle la afirmación anterior.

Los reactivos de una escala pueden redactarse de múltiples formas, pero habitualmente lo hacen como afirmaciones con las cuales los sujetos deben expresar su grado de acuerdo o desacuerdo. Lo relevante respecto de este punto es que los ítems deben ser opiniones (sobre el tema que sea o sobre rasgos del propio sujeto) y no hechos verificables. Este punto es lo que distingue una escala de una prueba de conocimiento. Lo que nos interesa es la valoración del sujeto (la que siempre está teñida por sus grupos de pertenencia) respecto de la serie de temas sobre los que versa la escala de medida.

Por otra parte, los reactivos del instrumento deben estar claramente relacionados con la variable o dimensión de la variable que apuntan a medir. Es decir, cada ítem debe contribuir

claramente a medir el constructo. Se trata como vemos de un criterio de relevancia.

La claridad de la redacción es un elemento insoslayable, ya que debemos precaver que todo aquel que lea el ítem entienda básicamente lo mismo. Los ítems ambiguos no nos sirven, como tampoco aquellos que expresan varias ideas a un mismo tiempo. En este sentido debemos ser rigurosos en la lectura de los ítems que vamos creando, de modo que sean simples y directos, y que expresen una y no más que una idea. En caso de que detectemos la presencia de más de un concepto, debemos dividir el reactivo de modo que no ocurra que el sujeto no sepa cómo contestar, pues puede estar de acuerdo con una parte del ítem y en desacuerdo con otra. Además, buscaremos que su redacción sea positiva para evitar confusiones (ej.: "No me gustan las discusiones sobre política" puede expresarse de mejor forma como "las discusiones sobre política me aburren").

Los ítems deben, además, contribuir a discriminar las posiciones de los sujetos o grupos, por lo que su redacción debe permitir que unos sujetos estén de acuerdo y otros en desacuerdo con la afirmación. Lo que se busca con las escalas es hacer distinciones y comparar sujetos. En otros términos, una buena escala debe contribuir a discriminar adecuadamente a los sujetos que poseen o no una variable (o que son pro o contra-actitudinales) o diversos grados de esta. Por regla general un ítem que contribuye a discriminar entre quienes poseen o no un rasgo, son favorables o no a una posición, o están de acuerdo o no con una determinada afirmación, son aquellos que nos interesan y sirven. Si bien, el grado de discriminación lo obtenemos posteriormente a los análisis, debemos desde la redacción procurar que apunten a detectar las diferencias.

Finalmente, la idea de contraste hace alusión a la posibilidad de generar ítems repetitivos donde la misma idea sea expresada de diferentes formas, lo que nos sirve no solo para analizar

cual redacción funciona mejor, sino además podemos utilizar varios ítems similares para analizar la consistencia de los sujetos a la hora de responder a nuestra escala. Algunos autores han señalado que con ítems repetitivos logramos mayores índices fiabilidad, pero a costa de simplificar demasiado el constructo (Morales, Urosa y Blanco, 2003). En este sentido, una buena escala tendrá ítems variados, toda vez que se hace recomendable que más de una persona revise su redacción para evitar incluir contenidos ambiguos e irrelevantes.

Un último elemento que es importante señalar respecto de la redacción de ítems dice relación con la posibilidad de redactar reactivos tanto positivos como negativos, es decir, ítems que expresen una opinión favorable y otros que lo hagan en sentido desfavorable. Esto permite matizar de mejor manera el constructo, mantener la atención necesaria por parte de la muestra y comprobar la consistencia y coherencia de las respuestas dadas por los sujetos, evitando así la "aquiescencia" o tendencia a mostrar acuerdo con cualquier afirmación (incluso con aquellas que se contradicen entre sí).

7.4.3 Determinar el modo de respuesta y el número de estas

Se trata en este paso de determinar el formato que tomarán las respuestas y el número de alternativas que se pondrá a disposición de los sujetos. Es obvio que la forma que tomen las respuestas de la escala dependerá del tipo de escala escogido. Para el caso analizaremos los dos modos más comunes y que refieren a la escala Likert y al diferencial semántico. En la Tabla 1 podemos apreciar un ejemplo para las preguntas con opción de respuesta para una escala Likert (con cinco opciones de respuesta que van desde "totalmente de acuerdo" a "totalmente en desacuerdo" y que suelen puntuarse desde 1 a 5) y otro para las de diferencial semántico (siete opciones de respuesta –desde "muy democrático" hasta "muy autoritario"- y que suelen puntuarse desde -3 a +3).

Para el caso de las escalas Likert el formato más habitual para las respuestas suele ser ajustado al grado de acuerdo o desacuerdo con una determinada afirmación (aunque también puede apuntar a medir el tipo de interés, el grado de importancia, la frecuencia, etc.). Además, habrá que señalar e número de opciones o de gradaciones que se ofrecerá. Finalmente, se deberá decidir respecto de la inclusión o no de un punto medio neutral o respuesta central, lo que equivale a resolver si se ofrecerá un número par o impar de alternativas.

Resulta frecuente también que solo se ofrezcan las palabras de los extremos, dejando en blanco las opciones de respuesta central.

Para el caso de las escalas de diferencial semántico, lo habitual suele ser ofrecer opciones intermedias entre los dos pares polares de adjetivos, de modo que la elección de una opción nos muestre qué tan cerca de una de las dos palabras se encuentra la evaluación del objeto. También aquí el número de opciones de respuestas puede variar desde dos a siete (pueden ser más pero es poco común).

En términos generales, las opciones de más de dos respuestas obtienen una mayor fiabilidad, la que crece hasta llegar a las siete opciones, luego de lo cual el incremento resulta poco aconsejable. Además, cabe consignar nuestra preferencia por los números pares (cuatro o seis opciones de respuesta), ya que de este modo los participantes de la muestra se ven forzados a tomar una opción más o menos favorable, de modo que las respuestas que manifiestan indiferencia respecto de la afirmación son controladas permitiendo al instrumento un mayor grado de discriminación (además, se trabaja con el supuesto que respecto de objetos conocidos siempre se tiene una actitud más o menos favorable cuando los ítems son relevantes).

7.4.4 Codificación de los datos

A las respuestas siempre se les asignarán números enteros y sucesivos. Además, es preferible evitar utilizar el cero como puntuación. De este modo, para el caso de una escala de cinco opciones de respuesta se deberá puntuar desde 1 hasta 5, y coordinarlo con el sentido del ítem, de modo que una actitud favorable hacia el objeto evaluado deberá tener puntuaciones más altas. Lo anterior vale solo como recomendación, ya que muchas veces una puntuación mayor indicaría mayores niveles de la variable medida.

Ahora bien, lo importante es precisar qué ítems deben puntuarse y en qué sentido. Así, para el ítem "las discusiones sobre política me aburren" la opción "totalmente de acuerdo" tendrá una puntuación de 1 y la opción "totalmente en desacuerdo" se puntuará con un 6 (en caso de que tenga seis opciones de respuesta). De este modo, una puntuación baja nos indicará una actitud poco favorable hacia la política. Si un segundo ítem realiza una afirmación favorable respecto del mismo objeto ("la política es una actividad relevante), entonces tendremos que invertir las puntuaciones de modo que quien esté "totalmente de acuerdo" recibirá una puntuación de 6 y quien manifiesta estar "totalmente en desacuerdo", recibirá una puntuación de 1 (ver Tabla 2).

Tabla 2. *Ejemplo de clave de corrección*

	Totalmente de acuerdo	De acuerdo	Más bien de acuerdo	Más bien en desacuerdo	En desacuerdo	Totalmente en desacuerdo
1. Las discusiones sobre política me aburren	1	2	3	4	5	6
2. La política es una actividad relevante	6	5	4	3	2	1

Estos números son los que asignaremos como puntaje de cada ítem y que más adelante enseñaremos a introducir el SPSS para realizar los análisis de fiabilidad y validez de nuestra escala.

7.4.5. Preguntas e instrumentos adicionales que aporten indicaciones de validez

Para validar un instrumento se necesita aportar datos que nos permitan comprobar cómo funciona nuestra escala relacionándolos con otros elementos de modo de verificar si realmente nuestro instrumento mide aquello que dice medir. En definitiva, se trata de buscar datos que la teoría prediga y que podrían estar relacionados con nuestro constructo. Así, por ejemplo si se trata de una escala que mide actitudes hacia la política cabría esperar que los sujetos que consideran relevante dicha actividad participaran en mayor grado en las elecciones que aquellos que la creen poco relevante. De este modo, cabría agregar una pregunta que nos entregue información respecto de la conducta de voto de los sujetos (Ej.: ¿ha votado en las últimas elecciones? O a modo de una escala ¿Cuán importante considera UD asistir a votar en los procesos eleccionarios?).

Los datos que pueden ser relevantes para cumplimentar adecuadamente este paso pueden ser de diferente tipo: sociodemográficos (edad, sexo, estado civil, etc.), escalas que midan el mismo rasgo o actitud (para confirmar si la correlación entre dos instrumentos similares es alta y positiva) y preguntas que apunten a medir un rasgo relacionado.

7.4.6. La prueba piloto

De lo que se trata a esta altura es de utilizar nuestra versión inicial de la escala para recoger los datos y hacer los análisis pertinentes.

Para ello debemos hacer que sujetos similares a aquellos a los que está destinada la escala contesten nuestro cuestionario. Respecto del número de participantes existen diferentes criterios: a) el número de sujetos debe ser al menos el doble del de ítems de la escala (Morales, Urosa y Blanco, 2003); b) el número de sujetos debe ser de al menos cinco sujetos por ítem (Nunnally, 1978); c) cualquier muestra sobre 100 sujetos sería suficiente (Kline, 1994). En términos generales, pensar en contar con a lo menos dos sujetos por ítem nos parece un criterio apropiado. Además, procediendo de este modo estaremos rondando el criterio de contar a lo menos con 100 sujetos (recuérdese que habíamos estimado en cerca de cuarenta el número de ítems que debían construirse para quedarnos con aproximadamente la mitad. Obviamente este criterio dependerá del tipo rasgo que estemos midiendo y de las dimensiones asociadas al mismo).

Por otra parte, se debe resguardar que las personas de la muestra tengan características similares. Si la muestra es heterogénea obtendremos mayores niveles de confiabilidad (debido a que existirá mayor dispersión de respuestas), pero se forzará demasiado el instrumento y no se logrará comprobar si funciona o no en la muestra para la cual fue diseñado.

7.5. Cálculo de confiabilidad y de las correlaciones ítem-total

Una vez que se ha aplicado la escala a la muestra debemos proceder a introducir los datos en el programa estadístico que utilizaremos para realizar los cálculos de confiabilidad. Debido al espacio que tenemos para exponer esta etapa deberemos dar por supuesta la familiaridad del lector con el programa estadístico que utilizaremos. En todo caso y para una introducción a SPSS el lector encontrará una muy clara y sintética exposición en el capítulo de "Introducción al programa estadístico SPSS" de este mismo manual o en una serie de trabajos dedicados al análisis

de datos con dicho programa (Pardo y Ruiz, 2002). Los datos incorporados deberán ser recodificados en el caso de los ítems inversos para posteriormente realizar el análisis de fiabilidad.

Previamente a ello realizaremos un análisis sobre los ítems, el cual nos indicará si todos los reactivos contribuyen a medir el rasgo sobre el que estamos trabajando y si contribuyen a diferencias adecuadamente a los sujetos (si son discriminantes y si las respuestas de los sujetos son coherentes). Dos son los métodos disponibles para esta comprobación: la correlación ítem total (correlación del ítem con el total menos el ítem) o el contraste de medias (pruebas t) en cada ítem para los grupos formados por las puntuaciones más altas y más bajas respectivamente. Nosotros utilizaremos aquí el cálculo de la correlación ítem total, el cual nos permitirá avanzar progresivamente en la eliminación de ítems de modo de mejorar nuestra fiabilidad y dejar solo aquellos ítems que funcionan adecuadamente.

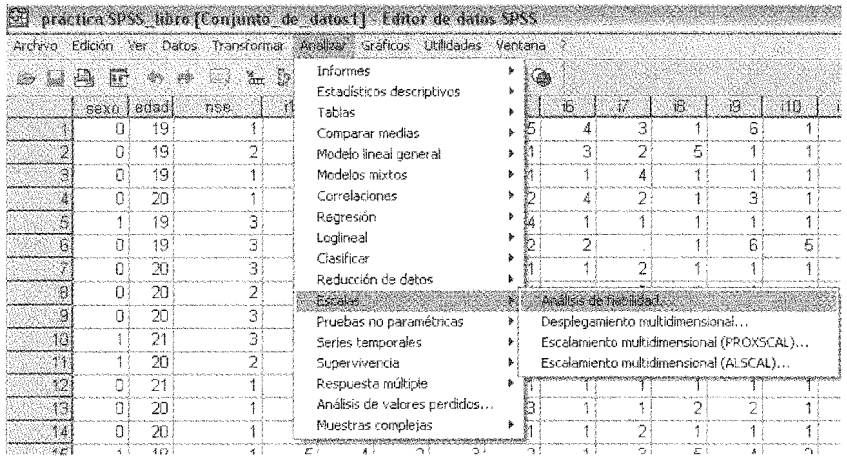
Cuando realizamos un análisis de los ítems aquellos que tienen mayor correlación indicarían que miden lo mismo que los demás y la eliminación de aquellos de más baja correlación con el total (bajo .25) permitirán alcanzar un coeficiente de fiabilidad de magnitud óptima. Existen amplias divergencias respecto del valor considerado adecuado para el coeficiente alfa de Cronbach. Así, existen autores que consideran que desde .50 sería aceptable en investigaciones de nivel básico (Guilford, 1954) y otros que apuntan a coeficientes superiores a .85 (Pfeiffer, Heslin y Jones, 1976). Para los efectos de este manual consideraremos adecuados los valores superiores a .70; meritorios, aquellos superiores a .80 y excelentes los valores sobre .90 (Nunnally, 1978; Pardo y Ruiz, 2002).

Este método de análisis de ítems funciona por pasos progresivos, por lo que una vez realizado el análisis de los reactivos y eliminados aquellos que no se comportan adecuadamente, debemos volver a realizar el análisis de ítems y el cálculo de

confiabilidad. Si nuestra escala posee más de una dimensión, procederemos a obtener un coeficiente alfa para cada una de ellas. En todo caso, cabe señalar que la eliminación de un ítem deberá tener también en consideración elementos racionales y no simplemente los de carácter automático (resultados arrojados por la computadora).

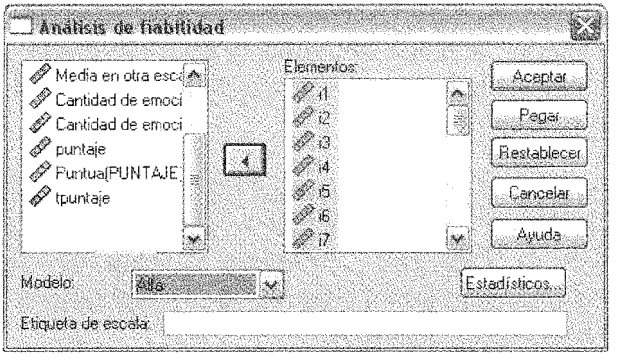
Para realizar el procedimiento en SPSS debemos pulsar la opción "Escalas" del menú "Analizar" y seleccionar el procedimiento "Análisis de fiabilidad..." (Figura 1). Una vez realizado esta selección se desplegará ante nosotros la ventana correspondiente a dicho análisis (Figura 2).

Figura 1. *Menú Analizar: Escalas: Análisis de fiabilidad.*



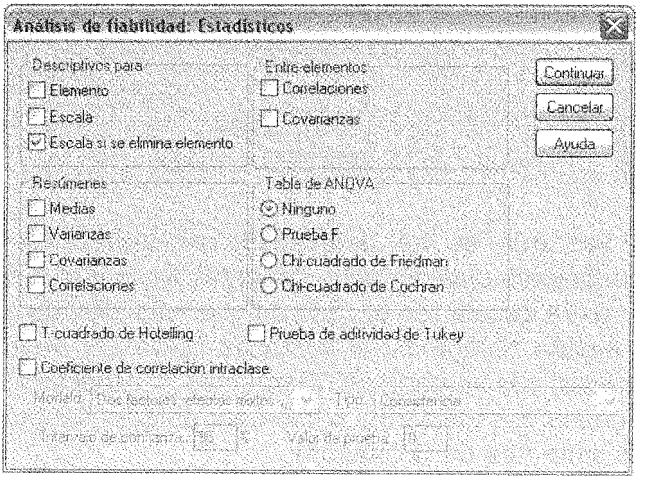
Una vez que se ha desplegado el cuadro de diálogo del procedimiento especificado, debemos traspasar todos los ítems de nuestra escala (en el caso del ejemplo referido a una escala que evalúa homofobia o actitudes hacia los homosexuales y lesbianas) a la ventana de "Elementos" y especificar el "Modelo" que deseamos utilizar (la opción por defecto del programa es el coeficiente Alfa de Cronbach).

Figura 2. *Análisis de Fiabilidad*



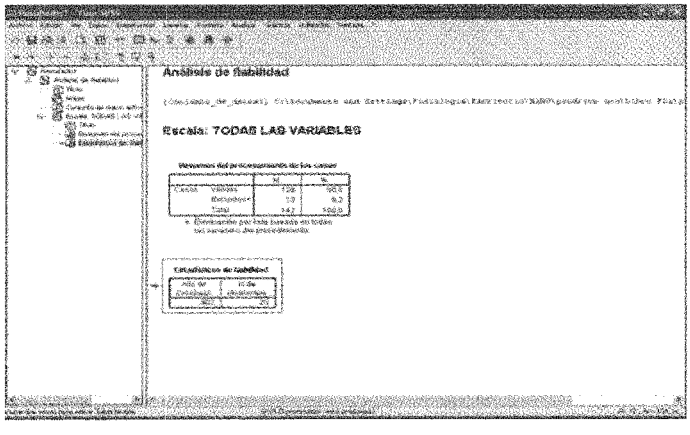
Si pulsamos el botón de estadísticos, se desplegará un nuevo cuadro de diálogo del que podremos obtener una serie de opciones, entre las que se encuentra aquella referida al cálculo de la correlación ítem-total y que se encuentra en el apartado "Descriptivos para" denominada "Escala si se elimina el elemento" (Figura 3). Posteriormente pulsamos "Continuar" y "Aceptar" para que se despliegue el visor de resultados con los datos solicitados.

Figura 3. *Análisis de Fiabilidad: Estadísticos.*



Como podemos apreciar en la Figura 4, los resultados obtenidos nos informan del funcionamiento de la escala que nos sirve de ejemplo, indicándonos un coeficiente de confiabilidad de .9032, lo que podría considerarse como excelente (como se observa, la muestra quedó compuesta por 129 participantes). El análisis de las correlaciones ítem-total nos informa del adecuado funcionamiento de casi todos los ítems de la escala (columna correlación ítem-total corregida), siendo el ítem 9 el único que manifiesta problemas (correlación inferior a .25) e informándonos del aumento del coeficiente alfa si el ítem es eliminado (aumento muy ligero a .9059).

Figura 4. *Análisis de Fiabilidad*



Para el caso de nuestra escala no es necesario repetir el procedimiento, pues ya tenemos la información del aumento del coeficiente fiabilidad si eliminamos el ítem que no funciona del todo bien. En el caso en que más de un reactivo tuviese problemas, deberíamos repetir el procedimiento restando al visor de "Elementos" de la Figura 2 los ítems de baja correlación y volviendo a pulsar en Aceptar. Este proceso lo repetimos cuantas veces sea necesario y hasta alcanzar un coeficiente aceptable y unas correlaciones ítem-total en todos los casos superiores a .25.

Para el caso de validación de escalas no eliminamos los ítems, sino que nos limitamos a informar cómo funciona cada uno de ellos. Si realizamos el cálculo para las dos subescalas o dimensiones contenidas en nuestro ejemplo, podremos apreciar que la confiabilidad para ambas es de .82. Podemos, además, obtener la correlación para ambas subescalas (R=.83), lo que nos aportaría pruebas de que ambas dimensiones estarían relacionadas.

Figura 5. *Análisis de Fiabilidad (correlación ítem-total).*

Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-tot al corregida	Alfa de Cronbach si se eleimina el elemento
i1	36,98	238,304	,585	,897
i2	37,71	235,647	,676	,895
i3	37,82	247,335	,401	,903
i4	38,29	238,647	,668	,895
i5	38,15	243,111	,579	,897
i6	38,50	248,955	,495	,900
i7	38,03	259,671	,221	,906
i8	38,26	244,289	,595	,897
i9	37,42	241,948	,496	,900
i10	38,76	252,090	,615	,898
i11	38,37	243,470	,690	,895
i12	38,31	243,450	,595	,897
i13	38,51	245,299	,511	,899
i14	37,51	240,908	,525	,899
i15	37,49	240,705	,511	,900
i16	38,32	240,687	,716	,894
i17	38,46	252,000	,402	,902
i18	38,78	254,703	,499	,900
i19	37,80	241,943	,521	,899
i20	38,15	241,486	,586	,897

Se debe tener claro que una vez eliminados los ítems defectuosos los cálculos que se realicen deben considerar su supresión. En nuestro caso la versión definitiva de la escala sería de 19 ítems

(una subescala de 9 reactivos y la otra de 10). Estos análisis pueden complementarse con el cálculo de estadísticos descriptivos para cada ítem (media y desviación típica) para el grupo total o para los subgrupos que se quieran formar (por sexo, edad u otro criterio relevante). Corresponde ahora realizar las comprobaciones sobre la validez del instrumento.

7.6. Comprobaciones de validez

Se debe tener claro ante todo, que los análisis de validez no consisten en un cálculo concreto sino más bien en un proceso. A lo largo de este se van aportando indicaciones que nos sirven para verificar si realmente estamos midiendo el constructo que afirmamos medir. En este sentido, validar consiste en investigar y aportar datos que hagan verosímil la utilidad de nuestra escala para acercarse al constructo que nos interesa.

Respecto de la validez, nos interesa confirmar que estamos midiendo el rasgo previsto en el constructo teórico (y para el cual elaboramos la tabla de especificaciones) y para ofrecer un análisis de las relaciones esperadas con otras medidas, ya que se trata de buscar la convergencia con indicadores que podamos interpretar como evidencia de que el rasgo analizado es lo que se está midiendo y no otra cosa. Así, si lo que medimos es una prueba de inteligencia, y encontramos una fuerte concordancia entre los resultados de nuestra prueba y los de otras que se suponen miden el mismo rasgo (otras pruebas de inteligencia), o si realizamos algunas predicciones (por ejemplo sobre el elevado rendimiento académico de una persona que puntúa alto en nuestra prueba de inteligencia), podremos acumular elementos que nos permitan ofrecer una interpretación persuasiva respecto de la validez de nuestro constructo teórico y su valor práctico.

Respecto de la *validez ajustada a criterio* ya hemos afirmado que en ella se relaciona la puntuación en la escala con el estado en alguna otra variable. Es decir, se realiza una predicción de lo que

se espera del sujeto en la segunda variable (que sirve como criterio). De lo que se trata es de aportar datos (y de allí que lo más dificultoso en este momento de la validación es encontrar algún criterio que resulte pertinente y relevante) que permitan mostrar que nuestra escala es capaz de predecir satisfactoriamente en alguna de las áreas de conducta afines a las que mide la escala. Por ejemplo, es esperable que las personas prejuiciosas con las mujeres (sexistas) lo sean también en otros ámbitos vinculados al prejuicio (homofobia). Pero más aún, es esperable que personas sexistas estén de acuerdo con restringir aún más los derechos de las mujeres. Por contrapartida, es esperable que personas igualitarias (que puntúan bajo en una escala que mide prejuicio) manifiesten su acuerdo con la ampliación de derechos a las mujeres.

El procedimiento adecuado para confirmar la relación entre dos puntuaciones es el coeficiente de correlación. Es decir, los sujetos sexistas deberían obtener un coeficiente de correlación alto y positivo con las puntuaciones de otra escala que mida otras formas de prejuicio (también podríamos realizar comparaciones de medias entre los sujetos sexistas y no sexistas en los puntajes para la pregunta de criterio). Estos procedimientos serán revisados con detalle en el capítulo dedicado al análisis de datos, por lo que ahora simplemente destacaremos su uso para este tipo de procedimiento.

Para controlar adecuadamente el *contenido de los test y escalas* de modo que estos resulten válidos, no existe un procedimiento estadístico asociado. Lo que se pretende aquí es más bien de asegurar una adecuada planificación del test y una rigurosa selección de los contenidos que se incluyen en este. Lo anterior implica poseer una visión clara de lo que el test o la escala pretende medir, así como de las formas en que esto se ha venido haciendo (revisión de las escalas disponibles para medir el mismo constructo). Definir adecuadamente el dominio designado por la escala implica distinguir un rango apropiado de tareas, estímulos

y situaciones, así como la clase de respuestas que se registrará y las instrucciones que se le entregarán al examinado (Cronbach, 1998). Otra estrategia posible consiste en dividir la variable en dominios y en asignar un número de reactivos a cada dominio (o dimensión de la variable), tal como lo mostramos en nuestro ejemplo sobre la tabla de especificaciones, el que debe ser rigurosamente revisado para apreciar si cada reactivo se ajusta a lo que se desea medir en cada dominio (por lo que la adecuada definición de cada dimensión debe ser presentada). Así, la forma de asegurar la validez de los contenidos de nuestro instrumento queda asegurada por una cuidadosa planificación y por una rigurosa revisión de los contenidos que se incluyen, ya que un concepto solo tiene sentido dentro de una teoría que lo define y delimita.

En el caso de la *validez de constructo* los antecedentes aportados pueden ser de diferente tipo: correlaciones, comparaciones de grupos o procedimientos de agrupación de datos.

Para el caso de los análisis de correlaciones y las comparaciones entre grupos estas serán ajustadas para indicar validez de constructo cuando lo pongan en relación con otros modos de medir el mismo rasgo (convergencia), para probar relaciones esperadas con otros rasgos (predicción), para comprobar que no exista relación donde esperamos que esta no se presente (divergencia) y para mostrar diferencias entre distintos grupos en el grado en que posee la variable analizada o variables relacionadas.

En el caso de los procedimientos de agrupación de datos es bastante extendido el uso de los análisis factoriales como método de comprobación de la validez de constructo. Ahora bien, no es posible afirmar que se ha validado un instrumento por el solo hecho de realizar un análisis factorial exploratorio, lo único que podemos hacer es obtener indicaciones sobre esta gracias a que nos facilita el proceso de dilucidación de los aspectos comunes que subyacen a un conjunto de reactivos y sobre los ítems que

conforman cada dimensión. Es decir, nos permite clarificar el constructo y compararlo con aquello que en términos teóricos habíamos definido como esencial de la estructura hipotética de nuestro constructo. En este sentido, el análisis factorial (que revisaremos detalladamente en el capítulo dedicado al análisis de los datos) permite indagar en al lógica interna y la estructura del constructo e intenta verificar la dimensionalidad del mismo.

Por otra parte, disponemos de modelos de análisis factorial que nos permiten comprobar la estructura de un instrumento, lo que resulta sumamente útil a la hora de realizar una adaptación de escalas. De este modo, lo que se confirma es la semejanza entre la estructura factorial del instrumento original y de su réplica. Estos métodos confirmatorios permiten una mayor concreción de las hipótesis que deben ser contrastadas (dado que sabemos de antemano en una adaptación qué ítems deberían contener cada dimensión), permitiendo la asignación de indicadores específicos a dimensiones concretas (Batista-Foguet, Coenders y Alonso, 2004) y superar las limitaciones que poseen los análisis factoriales exploratorio (no asumen ningún modelo para lo ítems y se limitan solo a reducir la dimensionalidad, no se explicita el número de factores y se obvia su contraste estadístico, etc.). Pese a todo lo anterior, nos parece que las mejores indicaciones de validez se obtienen del análisis factorial de los ítems de una escala, toda vez que resultan evidentes las ventajas de aquellos análisis que confirman la estructura de una escala respecto de aquellos que se limitan a explorarla.

7.7. Puntuaciones, conversión de puntuaciones y Normas

Es evidente que la puntuación directa obtenida por una persona en un test, una escala o una prueba no nos dice mucho respecto de quien lo ha obtenido. Por ejemplo, afirmar que alguien ha obtenido 40 puntos en una prueba que evalúa inteligencia no nos dice mucho (aun sabiendo el mínimo y el máximo de puntos que podrían obtenerse en ella) respecto de su posición y por ello del

grado en que posee la variable. En este sentido, debemos afirmar que las puntuaciones deben interpretarse siempre desde un marco de referencia claro y en relación a un parámetro con el cual comparar la puntuación. A este parámetro se le denomina habitualmente norma y refiere a la estandarización que se hace de un desempeño objetivo de una muestra representativa en la prueba. En este sentido, la norma es la que nos permite situar el desempeño del sujeto en relación con un grupo de referencia, de modo de saber si su rendimiento es similar al del promedio, si está ligeramente por debajo o por arriba de este, o si se encuentra cerca de alguno de los extremos de la distribución. Para ello debemos transformar la puntuación del sujeto en una medida relativa, la que nos permitirá comparar directamente diferentes individuos (Anastasi y Urbina, 1998).

En términos generales, pueden distinguirse dos tipos de normas: aquellas que comparan al sujeto con un patrón de desarrollo "normal" (normas de desarrollo), y las que lo evalúan en términos del desempeño de su grupo (normas intragrupo).

- *Normas de desarrollo.* Para el caso de este tipo de normas, las puntuaciones de los sujetos adquieren sentido al indicarnos cuánto ha progresado la persona en referencia al patrón de desarrollo normal para su etapa evolutiva. Son típicas de este tipo de normas las pruebas que miden inteligencia (ej. concepto de edad mental) y aquellas que se centran en el cumplimiento de las etapas del desarrollo evolutivo de los sujetos (ej. pruebas de conservación).

- *Normas intragrupo.* Casi todas las formas de estandarización se centran en este tipo de normas, incluso aquellas consideradas como normas de desarrollo tienen como referencia la comparación con ciertos patrones de desarrollo considerados normales para un grupo determinado. Aquí la descripción del desempeño del sujeto se realiza comparándolo con el desempeño medio de su grupo de

referencia y ubicándolo dentro del continuo definido por dicha muestra normativa.

Como ya hemos mencionado, todas las interpretaciones respecto del desempeño de un sujeto o sobre sus puntuaciones son sumamente relativas, en tanto siempre deben recurrir a una norma externa que permita asignarles un valor. Dos puntuaciones solo resultan comparables cuando explicitamos un criterio de comparación y, no está de más decirlo, esos criterios siempre pueden resultar arbitrarios. Por otra parte, cualquier norma siempre se restringe a la población normativa particular de la que se tomó la muestra (Anastasi y Urbina, 1998). Las muestras no resultan comparables si los grupos normativos pertenecen a poblaciones diferentes.

7.7.1. Conversión de las puntuaciones

La forma más sencilla de comparar puntuaciones es ordenándolas de modo de poder precisar el lugar del sujeto dentro del grupo que sirve como referencia. De este modo, la expresión en términos de porcentaje de sujetos que quedan sobre y bajo la puntuación obtenida por un determinado sujeto es una de las primeras y más utilizadas formas de conversión de la puntuación. Se denomina *percentiles* a la cifra que nos indica qué proporción del grupo se sitúa por debajo de una persona. Se trata de una medida de posición que tiene 99 valores de variable y que divide la distribución en 100 partes (donde cada parte contiene el 1% de las observaciones). De este modo un percentil 28 indica que el 28% de los sujetos se encuentra bajo de dicha observación y 72% sobre ella. Hay que tener en consideración que las puntuaciones directas se distribuyen de forma distinta que los percentiles, toda vez que los percentiles de dos grupos distintos, o de dos pruebas distintas, no son comparables.

Las puntuaciones transformadas permiten comparar puntuaciones en diferentes pruebas, aunque nunca nos indican

que las actuaciones sean equivalentes. Si una prueba es difícil, entonces que un sujeto se ubique en el percentil 50 puede reflejar un buen rendimiento. Si la prueba es fácil, entonces reflejará un rendimiento mediocre.

Otra forma de conversión común de puntajes tiene que ver con el uso de las medidas de tendencia central y las de dispersión (media y desviación típica). La media se define como la suma de los valores observados dividida por el número de ellos (Botella, León, San Martín y Barriopedro, 2001). La desviación típica refiere a la raíz cuadrada de la varianza, la que a su vez consiste en el cálculo promedio de las desviaciones cuadráticas respecto de la media. De este modo la desviación típica es un cálculo de las diferencias entre las puntuaciones y la media del grupo (Cronbach, 1998). Estas medidas son utilizadas para crear una escala de puntuaciones típicas (puntuaciones z), la que nos informa cuántas desviaciones típicas está la persona, ya sea por encima o por debajo de la media. Esta transformación no altera la forma de la distribución de frecuencias. En todo caso, y aun sin convertir los puntajes podemos utilizar la media y la desviación típica para generar puntuaciones de corte que nos permitan distinguir a los sujetos que se encuentran entre el 15% más alto y más bajo del grupo (que es el porcentaje que suele estar sobre el valor de la media sumada o restada a una desviación típica).

Lo que se busca con la transformación a puntuaciones típicas es hacer comparables puntuaciones obtenidas en grupos diferentes, ya que cada una nos entrega información sobre la variabilidad respecto del grupo de referencia. Otra forma de transformación de los puntajes crudos consiste en tipificarlos y transformarlos en una escala derivada (puntuaciones T), las que poseen la ventaja de tener siempre el valor 50 como media y la desviación típica de 10. La conversión de puntajes se realiza multiplicando la puntuación tipificada z por una constante (10) y sumarle otra (50). Estas constantes son arbitrarias, pero tienen

el mérito de ser las más utilizadas debido a las facilidades que nos entregan para situar los datos que arrojan.

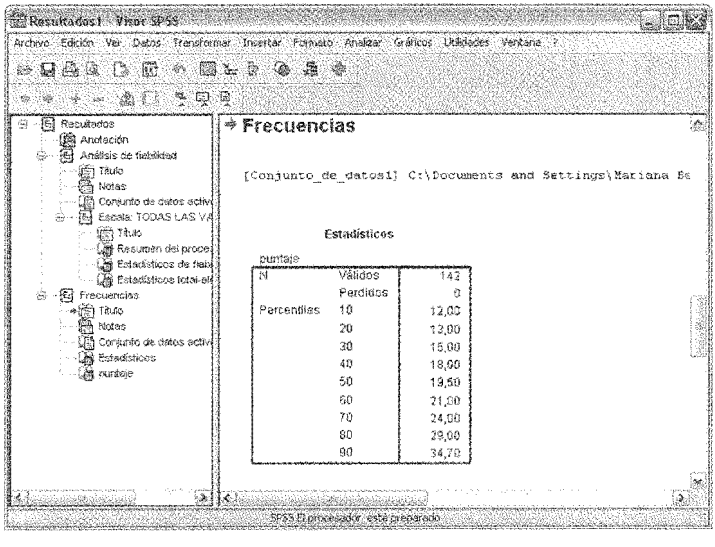
En SPSS todos estos cálculos son bastante sencillos. Así, para calcular los valores asociados a cada percentil debemos pulsar la opción "Frecuencias" ubicada dentro del submenú "Estadísticos descriptivos" del menú "Analizar", con lo cual se desplegará el cuadro de diálogo de Frecuencia (Figura 6). Una vez allí, traspasamos a la ventana de variables aquella que contiene el puntaje del sujeto en la escala y pulsamos el botón de "Estadísticos" y marcamos la opción percentiles (Figura 7).

Figura 6. *Frecuencias: Cálculo de percentiles*

Figura 7. *Frecuencias: Estadísticos: Cálculo de percentiles*

Una vez allí debemos introducir el o los percentiles que deseamos que se calculen. En el caso del ejemplo hemos introducido desde el 10 hasta el 100 en intervalos que van de diez en diez. Una vez concluida esta operación, pulsamos en "Continuar" y aceptamos. En el visor de resultados podremos apreciar los valores asociados a los percentiles calculados por el programa (Figura 8). El ejemplo está calculado sobre una escala de 10 ítems en formato Likert que tiene como opciones de respuesta valores que van desde 1 (Totalmente en desacuerdo) hasta 6 (Totalmente de acuerdo). Así, un sujeto que ha obtenido una puntuación de 29, se ubica en el percentil 80, lo que indicaría que el 80% de los sujetos han puntuado más bajo que él y solo el 20% ha puntuado de forma más alta.

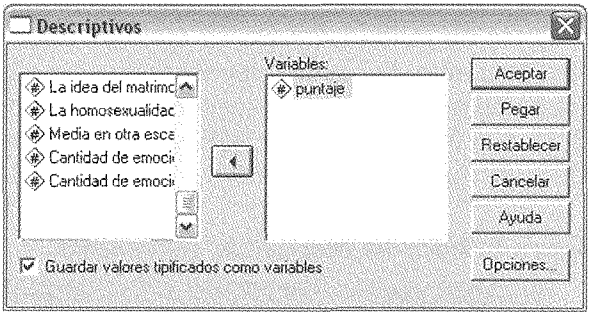
Figura 8. *Visor de resultados: percentiles*



El cálculo de las puntuaciones z es igualmente simple. Basta con escoger la opción "Descriptivos", ubicada dentro del submenú "Estadísticos descriptivos" del menú "Analizar", para que se despliegue el cuadro de diálogo respectivo. Como vemos en la Figura 9, solo debemos llevar la variable que contiene

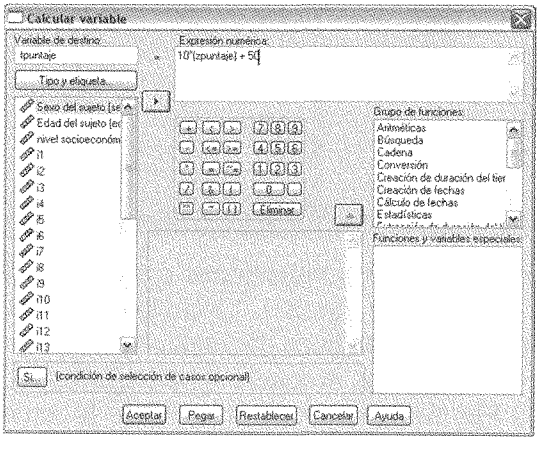
las puntuaciones de los sujetos en la escala a la lista de variables y marcar la opción "Guardar valores tipificados como variables". Si pulsamos en "Aceptar" de forma automática se genera la variable con las puntuaciones z en nuestra planilla de datos (Figura 11).

Figura 9. *Descriptivos: Cálculo de puntuaciones z*



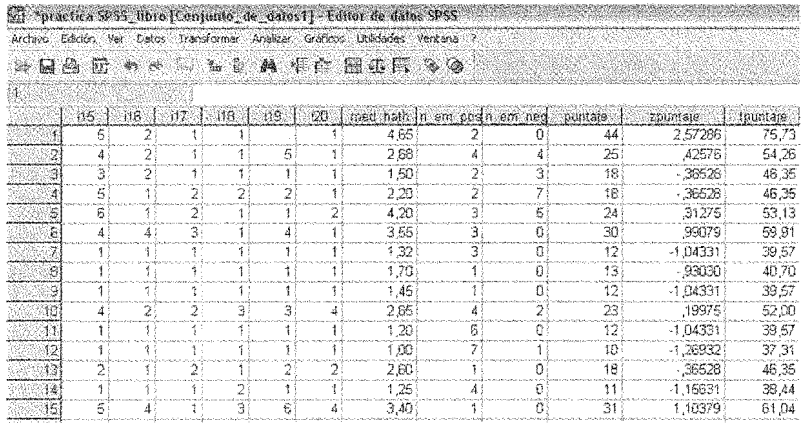
Como podemos apreciar en la Figura 10, al sujeto 2 (al que correspondía un puntaje bruto de 25 puntos) le corresponde una puntuación z de 0,425. Esta puntuación típica nos indicaría que el sujeto se encuentra a menos de media desviación típica de la media.

Figura 10. *Calcular: Puntuaciones T*



La obtención de las puntuaciones derivadas T se calcula, como ya dijimos, multiplicando la puntuación z del sujeto por una constante y sumándole otra (que como ya dijimos serán de 10 y 50 respectivamente). Para realizar este cálculo, debemos pulsar la opción "Calcular" del menú "Transformar" y se desplegará automáticamente el cuadro de diálogo que vemos en la Figura 9. Allí debemos asignar un nombre a la variable que vamos a crear (denominada en el ejemplo "tpuntaje") y multiplicar la variable que contiene las puntuaciones z por la constante 10, para luego sumarle la constante 50 (todo ello dentro de la ventana "Expresión numérica").

Figura 11. *Editor de datos: Puntuaciones z y T*



	i15	i16	i17	i18	i19	i20	med_hath	n_em_pos	n_em_neg	puntaje	zpuntaje	tpuntaje
1	5	2	1	1		1	4,65	2	0	44	2,57286	75,73
2	4	2	1	1	5	1	2,68	4	4	25	,42576	54,26
3	3	2	1	1	1	1	1,50	2	3	18	-,38528	46,35
4	5	1	2	2	2	1	2,20	2	7	18	-,38528	46,35
5	6	1	2	1	1	2	4,20	3	6	24	,31275	53,13
6	4	4	3	1	4	1	3,55	3	0	30	,99079	59,91
7	1	1	1	1	1	1	1,32	3	0	12	-1,04331	39,57
8	1	1	1	1	1	1	1,70	1	0	13	-,93030	40,70
9	1	1	1	1	1	1	1,45	1	0	12	-1,04331	39,57
10	4	2	2	3	3	4	2,85	4	2	23	,19975	52,00
11	1	1	1	1	1	1	1,20	6	0	12	-1,04331	39,57
12	1	1	1	1	1	1	1,00	7	1	10	-1,26932	37,31
13	2	1	2	1	2	2	2,60	1	0	18	,38528	46,35
14	1	1	1	2	1	1	1,25	4	0	11	-1,15631	38,44
15	5	4	1	3	6	4	3,40	1	0	31	1,10379	61,04

Como podemos apreciar en la figura anterior, al sujeto 10 (que obtuvo una puntuación bruta de 23) le corresponde una puntuación z de .199 y una puntuación T de 52. Lo que nos indicaría que está muy cercano a la media del grupo.

7.2.2. Comparación entre los diferentes sistemas de conversión

La pregunta que puede surgir al lector a estas alturas refiere a las bondades y desventajas de cada tipo de puntuación, de modo de poder decidir con propiedad cuál utilizar.

En términos generales, la ventaja de utilizar los percentiles como medida de normalización consiste en que es más fácil su comprensión de modo intuitivo y que pueden interpretarse de forma directa, independientemente de la distribución. Entre los problemas asociados se ha reseñado que magnifican las diferencias cerca de la media y reducen el tamaño de las diferencias grandes en las colas de distribución (Cronbach, 1998), es decir, la distancia de los centiles intermedios serán menores que la distancia de los centiles extremos, debido a que suelen encontrase más valores intermedios que extremos (Botella et al., 2001).

Entre las ventajas de las puntuaciones típicas cabe destacar que las diferencias entre ellas son equivalentes a las encontradas en las puntuaciones directas, de modo que se pueden realizar algunos cálculos estadísticos con ellas (correlaciones) con igual resultado que si utilizamos las puntuaciones directas. Entre sus desventajas encontramos el hecho de que su interpretación directa no es intuitiva y que sus puntuaciones no pueden ser interpretadas adecuadamente cuando la distribución está sesgada (se acumulan valores hacia una de las colas de la distribución debido a la presencia de puntajes altos o bajos).

7.8. Bibliografía

- Anastasi, A. & Urbina, S. (1998). Test psicológicos. México DF: Prentice Hall.
- Batista-Foguet, J. M., Coenders, G. & Alonso, J. (2004). Análisis factorial confirmatorio. Su utilidad en la validación de cuestionarios relacionados con la salud. Med Clin, 122 (1), 21-27.
- Botella, J., León, O., San Martín, R. & Barriopedro, M. I. (2001). Análisis de datos en psicología I. teoría y ejercicios. Madrid: Pirámide.

- Cárdenas, M. (2006). El análisis multivariado de las representaciones sociales. Antofagasta: Editorial Universidad Católica del Norte.

- Cronbach, L. J. (1998). Fundamentos de los test psicológicos. Aplicaciones a las organizaciones, la educación y la clínica. Madrid: Biblioteca Nueva.

- Dovidio, J. F. & Gaertner, S. L. (1986). Prejudice, Discrimination and Racism. San Diego: Academia Press.

- Guilford, J. P. (1954). Psychometric methods. New York: McGraw-Hill.

- Guttman, L. (1944). A basis for scaling qualitative data. American Sociological Review, 9, 139-150.

- Hernández, R., Fernández, C. & Baptista, P. (1998). Metodología de la investigación. México DF: McGraw-Hill.

- Kline, P. (1993). The Handbook of psychological testing. London: Routledge.

- Kline, P. (1994). An easy guide to factor analysis. Newbury Park: Sage.

- Likert, R. (1932). A technique for the measurement of attitudes. Archives of psychology, 140, 44-53.

- Morales, P., Urosa, B. & Blanco, A. (2003). Construcción de escalas de actitudes tipo likert. Cuadernos de Estadística, 26. Madrid: La Muralla.

- Nunnally, J. C. (1978). Psycometric Theory. New Cork: McGraw-Hill.

- Osgood, C. E., Suci, G. J. & Tanembaum, P, H. (1976). La medida del significado. Madrid: Gredos.

- Pardo, A. & San Martín, R. (2004). Análisis de datos en psicología II. Madrid: Pirámide.

- Pardo, A. y Ruiz, M. A. (2002). SPSS 11. Guía para el análisis de datos. Madrid: McGraw-Hill.

- Pettigrew, T. F. & Merteens, R. W. (1995). Subtle and blatant prejudice in western Europe. European Journal of Social Psychology, 25, 57-75.

- Pfeiffer, J. W., Heslin, R. & Jones, J. E. (1976). Instrumentation in human relations training. California: University associates.

- Stevens, S. S. (1951). Mathematics, measurement and psychophysics. En S. S. Stevens (Ed), Handbook of experimental psychology. New York: Wiley.

- Thurstone, L. L. (1928). Attitudes can be measured. American Journal of Sociology, 33, 529-554.