GUÍA PARA CONSTRUIR CUESTIONARIOS y ESCALAS DE ACTITUDES

Publicado en Guatemala: Universidad Rafael Landívar (2011)
Disponible en http://www.upcomillas.es/personal/peter/otrosdocumentos/Guiaparaconstruirescalasdeactitudes.pdf

ÍNDICE

1. Cuestionarios y escalas

En primer lugar hay que aclarar qué entendemos *aquí* por *cuestionario* y qué entendemos por *escala* o *test.* Un *cuestionario* según, el diccionario, es una *lista de preguntas que se proponen con cualquier fin;* los cuestionarios sociológicos, de evaluación, y en general los *sondeos de opinión* son ejemplos típicos.

En los cuestionarios convencionales, los más habituales, las respuestas *se analizan de manera independiente*. Un *test* o una *escala de actitudes* son también cuestionarios pero con estas características:

1) Todas las preguntas (ítems) son *indicadores del mismo rasgo o actitud*,
2) Las respuestas de cada sujeto se van a *sumar en un total* que indica dónde se encuentra o *cuánto tiene* de la variable o característica que pretendemos medir.

Es lo mismo que sucede en los exámenes convencionales de conocimientos (como las pruebas objetivas *tipo test*): a cada sujeto se le suman sus respuestas correctas y este total es el dato que se tiene en cuenta para calificar

También suelen denominarse *escalas* las preguntas con *respuestas graduadas* (como *mucho, bastante, poco, nada*, o en términos de *frecuencia, importancia*, etc.), frecuentes en todo tipo de cuestionarios, aunque las respuestas no se vayan a sumar en un único total porque cada pregunta mide algo distinto.

Independientemente de los términos que utilicemos, las diferencias entre lo que denominamos *cuestionarios* y *escalas* (o tests) están esquematizadas en la figura 1.

	Cuestionarios	**Tests y escalas**
Naturaleza de los ítems	**Con cada uno se obtiene *información distinta***	**Todos los ítems expresan *el mismo rasgo***
Selección de los ítems	**Criterios lógicos, *se pregunta lo que se desea conocer***	***Análisis estadístico de cada ítem* antes de ser considerado definitivo**
Datos interpretados y utilizados	**Las respuestas de *cada* pregunta o ítem por separado**	**La *suma* de todas las respuestas**

Figura 1

Construimos *escalas de actitudes* (o tests de personalidad e instrumentos semejantes) para *medir* determinados rasgos. Aquí entendemos por *medir*, de una manera muy genérica, apreciar *cuantitativamente* si un sujeto *tiene poco o mucho* del rasgo en cuestión, ver dónde se sitúa cada sujeto en un *continuo* de menos a más[1].

Aunque no hay consistencia en el uso de ninguno de estos términos, habitualmente se emplea el término *escala* cuando se trata de medir *actitudes*, y el término *test* cuando se trata de medir otros rasgos psicológicos (como *inteligencia* o *personalidad*); también se utiliza el término *test* en la medición de *conocimientos, habilidades* e *intereses*. Lo que tienen en común estos términos (*test, escala de actitudes*) es que para medir un único rasgo utilizamos varias preguntas cuya respuestas (*traducidas* en números) se suman para cada sujeto en una puntuación total que es el dato individual que se utiliza e interpreta, se calculan medias de grupos, etc.

[1] Este *medir* hay que entenderlo en un sentido analógico; en sentido propio no medimos nada porque carecemos de una *unidad* propiamente dicha, sin embargo estos procedimientos funcionan razonablemente bien; una exposición y justificación más amplia de la medición en psicología puede verse en Morales (2006, Cap. I).

Las *escalas de actitudes*, de las que vamos a tratar de manera más específica, también suelen presentarse con el término general de *cuestionario*. El término *instrumento* es muy genérico, equivale a *cuestionario,* y puede referirse tanto a escalas y tests específicos como a cuestionarios convencionales, lo más frecuente es denominar instrumento a todo el *instrumento de obtención de datos* que suele incluir dos tipos de preguntas:

a) Preguntas de *identificación personal* (sexo, edad, y cualquier otra información útil) que por lo general se colocan al comienzo.

b) Preguntas específicas sobre el *objeto de la investigación* (*variable dependiente*), que pueden ser o preguntas independientes o escalas y tests.

1.1. Cautelas iniciales en la construcción de cuestionarios y escalas

Lo que vamos a exponer sobre la redacción de los ítems (como *tipos de preguntas* y *tipos de respuestas*) es igualmente válido tanto si se trata de un *cuestionario convencional* en el que cada pregunta aporta una información distinta, como si se trata de construir *escalas de actitudes* y tests en general.

Algunas cautelas y recomendaciones iniciales.

1) *Disponer de otras fuentes de información*

Es muy conveniente disponer de otras fuentes de información sobre construcción de cuestionarios y escalas ya que podemos encontrar otros modelos de preguntas, formas de presentar cuestionarios e incluso ejemplos de instrucciones a los sujetos. Esta información se localiza con facilidad en textos de métodos de investigación en las ciencias sociales y también disponemos de ejemplos en cuestionarios ya utilizados en otras investigaciones[2].

2) *Evitar cuestionarios muy largos*

Hay que procurar hacer cuestionarios de una *longitud razonable,* teniendo en cuenta *quiénes y en qué situación* los van a responder. Los cuestionarios muy largos cansan, se responden descuidadamente y muchos sujetos dejan preguntas sin responder. Hay que limitarse a obtener la información que realmente interesa sin caer en la tentación de *aprovechar la oportunidad* para preguntar todo lo que *quizás podría* ser de utilidad.

Algunos cuestionarios dan la impresión de que el autor no ha querido *dejarse nada en el tintero* que pueda tener que ver con su área de estudio, incluyendo preguntas de un interés muy marginal. En caso de duda hay que *evaluar* la necesidad de hacer determinadas preguntas.

Esta observación sobre la *longitud* se refiere sobre todo a los cuestionarios más convencionales; si se trata de una *escala de actitudes* ya veremos que es importante partir de un número más bien grande de ítems[3].

Tampoco hay que olvidar el trabajo y tiempo posterior que supone introducir todos los datos en un programa informático (en una hoja EXCEL) cuando no se dispone de hojas de *lectura óptica*.

---

[2] No hay que olvidar que no se trata de una *ciencia exacta*. En Internet podemos encontrar buenas orientaciones para construir cuestionarios, como Frary (1996), Marshall (1998) y Fanning (2005); también disponemos de buenos textos con modelos de cuestionarios y preguntas de muy diverso tipo (como Hernández Sampieri, Fernández Collado y Baptista Lucio, 2010). Una guía sobre *construcción de escalas* que incluye cómo utilizar el programa SPSS en la construcción de escalas en Morales, Urosa y Blanco (2003); en Morales (2006) un tratamiento más extenso sobre cuestiones conceptuales y metodológicas en relación con la construcción de escalas (referencias completas en la bibliografía). En la bibliografía indicamos otras publicaciones que orientan sobre la construcción de cuestionarios, escalas de actitudes y de tests en general, y también se enumeran una serie de obras en las que se reproducen muchas escalas e instrumentos semejantes.

[3] Lo comentamos en al apartado 8.1

3) *Evitar preguntas redundantes*

Hay que evitar preguntar básicamente *lo mismo* más de una vez, por ejemplo *edad* y *curso* (o *grado*) en muestras escolares cuando todos los del mismo curso son casi de la misma edad (aunque esto no es siempre así). Tampoco hay que preguntar *lo que ya se sabe* (por ejemplo el sexo si todos los sujetos de la muestra son niños o niñas)

4) *Evitar preguntas que complican los análisis*

Hay modos de preguntar que por una razón o por otra complican los análisis, como son las preguntas que requieren *ordenar* y las preguntas de *respuesta abierta*. Esto es sólo una recomendación pues ambos tipos de preguntas pueden tener su interés.[4]

5) *Tener un plan inicial claro de la información que interesa recoger*

Un plan inicial claro facilita el centrarse en las preguntas que realmente interesan; hay que evitar hacer *preguntas inútiles*. Cuando el investigador no ha definido de manera *precisa* qué información le interesa obtener y se redactan muchas preguntas, hay que hacer después una *revisión tranquila* para seleccionar (o reformular) las preguntas que realmente aportan información de interés. En cualquier caso son recomendables estas segundas revisiones para perfilar mejor el instrumento.

Si se trata de hacer una escala de actitudes, dedicamos un apartado específico (el nº 5) para comentar los diversos *puntos de partida para construir una escala de actitudes*.

1.2. Preguntas de identificación personal en cuestionarios y escalas

Estos datos de identificación personal suelen ser edad, sexo, ocupación, antigüedad o curso, estado civil, grupo de pertenencia, etc. Estas preguntas con información sobre cada sujeto suelen ir siempre al comienzo de cualquier tipo de cuestionario o escala.

1.2.1. Orientaciones generales

Este tipo de información también se codifica con números. Conviene buscar y ver modelos, pero en principio y como orientación podemos distinguir tres tipos de datos:

1) *Datos dicotómicos* (sólo *dos categorías de respuesta*, como *sí* o *no*, sexo, etc., que se excluyen mutuamente); estos datos se codifican con *unos* y *ceros* (es lo habitual, aunque también se utiliza *uno* y *dos*).

2) *Datos continuos* que expresan una *cantidad* o al menos un *orden*, se codifican con los números originales, como pueden ser *curso* y *edad*, y también los números que expresan *grado* de acuerdo, frecuencia, etc. Datos como *edad, antigüedad en la institución*, etc., es frecuente agruparlos en *intervalos* que se numeran correlativamente (1, 2, etc., de menos a más).

3) *Datos nominales* o *cualitativos* con más de dos respuestas para escoger una; en realidad son *categorías de clasificación* (como *profesión, carrera, grupo étnico, lugar de procedencia*, etc.). También se asigna un número a cada categoría de respuesta[5], aunque estos números tendrán en los análisis un tratamiento diferente porque no expresan ni orden ni cantidad; no son números en sentido propio sino una manera de categorizar conceptos.

En este apartado sobre información personal conviene:

---

[4] Ampliamos la información en los apartados correspondientes, 2.7 *listas de ordenamiento* y 2.9 *preguntas abiertas.*

[5] Programas informáticos como EXCEL o el SPSS sólo admiten *números*.

1) Mantener el *anonimato* del que responde para garantizar la sinceridad en las respuestas. Como por diversas razones nos puede interesar conocer *quién es quién*, tratamos esta cuestión en el apartado siguiente (1.2.2.).

2) No hacer más preguntas de las necesarias.

3) Evitar la opción *otras* (como cuando se pregunta por *carrera, profesión, motivación,* etc.).

En algunos cuestionarios a *otras* se añade *especifique;* estas respuestas habría que leerlas y sistematizarlas, complican la descripción de la muestra y con frecuencia no se hace ningún caso a estas respuestas en los análisis. Naturalmente se trata de una *recomendación*, no una *norma,* pero como criterio general el investigador no debe complicar su tarea más de lo necesario.

Un ejemplo sobre cómo presentar en un cuestionario y cómo codificar con números este tipo de datos lo tenemos en la figura 2.

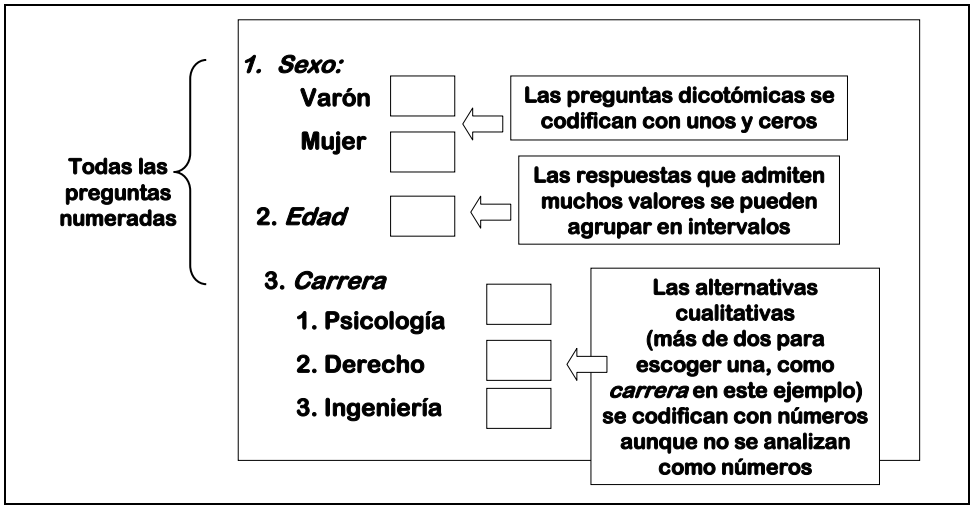


Figura 2

Estos datos de *información personal* servirán para *describir* la muestra y también para hacer análisis adicionales, como pueden ser.

a) Exponer tablas y gráficos *descriptivos*, por sexos, subgrupos, etc.
b) *Comparar* subgrupos en la variable o variables medidas por nuestro instrumento
c) Verificar *relaciones* entre la escala o preguntas el cuestionario y este tipo de datos (edad, sexo, etc.).
d) En el caso de tests y escalas se pueden preparar *normas* de interpretación individual (también denominadas *baremos*) como los *percentiles*, para los distintos subgrupos (se pueden calcular percentiles según edades, sexos, etc.)[6].

1.2.2. El anonimato en los cuestionarios y escalas

Para garantizar la *sinceridad* en las respuestas a este tipo de instrumentos es importante el *anonimato*. Por esta razón, y como criterio general, en el caso frecuente en el que son alumnos los que responden a un cuestionario o escala de actitudes, no es aconsejable pedirles un *número* (el de inscripción o cualquier otro dato) que permita su identificación. Es más en las mismas instrucciones para responder, al inicio del cuestionario o escala, se puede indicar

[6] Los *percentiles* (porcentaje de sujetos por debajo de cada puntuación; expresan la *posición relativa* del sujeto en el grupo) son un método habitual para interpretar resultados individuales aunque hay otros tipos de puntuaciones que pueden resultar útiles (explicados en Morales, 2008, cap. 3).

expresamente *este cuestionario es anónimo, por favor responda con sinceridad a todas las preguntas.* Cuando se pasa solamente un cuestionario, *cada respuesta es una firma para todas las demás* y no hace falta saber quién es quién.

Esta recomendación sobre el anonimato no puede considerarse como una norma absoluta; dependiendo de la *naturaleza* de lo que se pregunta (pueden ser preguntas muy *inocentes*), la *edad* de los alumnos, *situación* y *clima*, se puede valorar si el que el alumno se identifique con su nombre (o una clave que permite identificarle) puede restar sinceridad a sus respuestas. A veces estos cuestionarios (o tests o escalas) se responden y entregan junto con un examen en el que obviamente figura el nombre del sujeto; en estos casos siempre hay que calibrar si la falta de anonimato puede poner en peligro la falta de sinceridad[7].

1. Como el saber *quién es quién* nos puede interesar por distintas razones, una alternativa es pedir a los alumnos que pongan una clave que *solamente ellos van a conocer* y que puedan recordar con facilidad, como pueden ser los *cuatro últimos números* de su teléfono o de un documento de identidad.

Por ejemplo, cuando los sujetos van a responder *a lo largo del tiempo* (durante un curso) a varios de estos instrumentos *anónimos* (escalas, cuestionarios diversos) o al mismo instrumento en tiempos distintos *(¿sube o baja el interés por la asignatura?*), nos interesa disponer de *toda la información de cada alumno obtenida en momentos distintos* para hacer determinados análisis; por ejemplo para verificar *relaciones, cambos y tendencias*. Es suficiente que cada uno tenga su contraseña, siempre la misma y que el profesor no necesita conocer, que le dé seguridad de que el anonimato está garantizado.

2. En algunos casos el saber *quién es quién* es necesario poder *dar a todos una información individual* sobre los resultados de un test, examen, etc., de manera que cada uno conozca su propio resultado pero no el de sus compañeros. Cada uno puede contar de sí mismo lo que quiera y a quien quiera, pero un profesor debe ser prudente al hacer pública esta información. En estos casos el profesor puede utilizar cualquier clave que tenga disponible, como un número de inscripción; en estos casos el profesor conoce la clave de todos, pero ninguno conoce la clave de sus compañeros. Naturalmente esta clave debe ser distinta a la que el alumno puede utilizar en un cuestionario anónimo y que sólo el conoce.

3. Otra razón para saber *quién es quién* es el poder buscar información adicional sobre los mismos sujetos; frecuentemente se trata de tener acceso a calificaciones escolares para ampliar algún estudio (como correlación entre determinadas preguntas o escalas y rendimiento académico). Como el rendimiento académico es una variable de interés en muchas investigaciones, lo tratamos en un apartado diferenciado; hay *procedimientos indirectos* que nos pueden dar una información suficientemente válida sin necesidad de buscar las calificaciones de cada sujeto.

1.2.3. Cuando interesa disponer de información sobre el rendimiento académico

En muestras escolares la información sobre el rendimiento de los alumnos interesa con frecuencia para verificar su relación con otras variables (como actitudes y motivaciones) que medimos mediante cuestionarios y escalas.

---

[7] Un ejemplo interesante lo tenemos en Tian (2007) que utiliza un conocido cuestionario para medir *enfoques en el aprendizaje* (*superficial*, más memorístico o *profundo*, más orientado a la comprensión) y ver su relación con las calificaciones obtenidas en dos grupos distintos de alumnos universitarios, en uno la evaluación se hace con exámenes tradicionales y el otro con trabajos escritos para hacer en casa; en este caso es necesario conocer de cada alumno su puntuación en los dos enfoques de aprendizaje y las calificaciones que ha obtenido pues se trata precisamente de conocer si un determinado enfoque en el estudio favorece el obtener mejores calificaciones según cómo sea el tipo de examen. El test de *enfoques de aprendizaje* (20 ítems, 10 para medir cada enfoque) es de Biggs, Kember, y Leung (2001) (la versión en español, de Blanco, Prieto, Torre, y García, 2009, está disponible en Morales, 2011, *cuestionarios y escalas*).

En este punto hay que hacer dos observaciones para situaciones distintas.

1. Cuando por alguna razón el cuestionario no es anónimo, sí podemos tener acceso a esta información. Aunque el anonimato debe ser la norma, puede haber situaciones en que sí sabemos *quién es quién*. En estos casos lo habitual es buscar la *nota media* de cada alumno (en una asignatura, de final de carrera, etc.).

En estos casos las *notas medias* no son necesariamente la mejor opción; suelen parecerse demasiado unas a otras y las correlaciones con otras variables suelen ser bajas porque dependen de las diferencias entre los sujetos. Una mejor opción (*además* de notas medias) pueden ser las notas en cada materia o en algunas materias concretas, por su *dificultad* (como suele suceder con las *matemáticas*) o por cualquier otra razón.

Las calificaciones tampoco son siempre el mejor o más útil indicador de rendimiento; puede ser también el número de respuestas correctas en un determinado examen o los resultados en ejercicios específicos (como *prácticas,* trabajos hechos en casa, etc.). En cualquier caso es útil disponer de *varios criterios* de rendimiento y no de uno solo.

2. Si el cuestionario es anónimo hay métodos indirectos para obtener información sobre el rendimiento salvaguardando el anonimato con preguntas de este estilo:

- *Nota esperada en la asignatura más fácil*
- *Nota esperada en la asignatura más difícil (se puede especificar la asignatura)*
- *Número probable de notas altas*
- *Número probable de notas bajas*
- *Puesto relativo que ocupo en la clase (de los primeros, en la mitad superior, promedio, debajo del promedio pero no muy bajo, entre los últimos, o algo similar).*

También caben preguntas no relacionadas directamente con calificaciones sino con otras variables que probablemente sí lo están, como por ejemplo (Cheang, 2009):

Preparo cada clase: Siempre--con frecuencia—ocasionalmente--nunca[8]

Estas preguntas y otras semejantes[9] pueden verse influidas por la *deseabilidad social* y los sujetos pueden valorarse con mucha benevolencia, pero en la medida en que esta *deseabilidad social* afecta más o menos a todos, se mantienen las diferencias entre los sujetos que es lo que importa para detectar relaciones; los *extremos* suelen estar claros. Las respuestas a estas preguntas siempre pueden interpretarse en términos relativos.

Dos o tres *indicadores indirectos* de rendimiento (que pueden incluso sumarse) pueden darnos una información útil, menos exacta obviamente que el disponer de resultados reales, pero que garantizan mejor la sinceridad de los sujetos en el resto del cuestionario; además este tipo de preguntas tienen interés en sí mismas y son suficientes para responder a nuestras preguntas de investigación. Además diversos estudios experimentales muestran que estos indicadores de rendimiento académico *funcionan* bien como sustitutos de las calificaciones[10].

1.3. Los cuestionarios: ¿Podemos 'medir' actitudes con una sola pregunta?

En otros apartados expondremos en detalle cómo se construye una *escala de actitudes,* pero también hay que dejar claro que se pueden medir actitudes *con una sola pregunta;* esto es lo habitual en los cuestionarios sociológicos. A un sujeto se le puede preguntar que se sitúe en una escala de 1 a 6 (de *nada* a *mucho)* en una serie de rasgos o actitudes personales:

---

[8] Sugerencias en la misma línea en el apartado 13.2, *comprobación de relaciones esperadas con otras variables.*

[9] En Morales (2011, *cuestionarios y escalas*) tenemos ejemplos de *indicadores de rendimiento académico.*

[10] Por ejemplo, Laird, Shoup, y Kuh (2005) y Olsen, Schilling, KM, Schilling, K., Connolly, y Vesper (1998).

conservador, extravertido, religioso, partidario de una determinada opción política, aficionado a la ópera, etc. Así se miden actitudes con frecuencia, tenemos numerosos ejemplos y está bien hecho; no hay que *hacer escalas* siempre que se piensa en medir actitudes.

El medir (*evaluar, hacer una investigación sobre*) actitudes (o *cada* actitud) con *una única pregunta* puede estar más indicado con alguna de las finalidades siguientes:

a) Cuando no se pretende obtener una información muy precisa sobre *cada sujeto* en particular (como sería necesario, por ejemplo, para hacer un *diagnóstico individual*) sino más bien conocer la *tónica* de un grupo representada por su media o por unos porcentajes[11] (como suele ser el caso en los cuestionarios sociológicos).

b) Cuando no se trata de medir con precisión *una* actitud conceptualizada con cierta complejidad sino las *actitudes o valoraciones generales* hacia una pluralidad de elementos del mismo ámbito (asignaturas, profesiones, actividades, estilos, etc.).

c) Cuando se trata de medir actitudes (u otros rasgos psicológicos o de otro tipo) de una manera sencilla y económica como *medida complementaria* de instrumentos más elaborados[12].

Lo que iremos exponiendo más adelante sobre la *redacción de los ítems* de una escala o test es también válido para preparar las preguntas o ítems de un *cuestionario convencional;* son modelos de preguntas (sobre actitudes y conceptos afines) independientemente de que formen parte de una escala o test o de que figuren en un cuestionario como preguntas independientes.

Los *cuestionarios convencionales* (no las escalas de actitudes) son el instrumento habitual en muchas investigaciones en las que interesa obtener información sobre *aspectos distintos,* que pueden ser muchos, de la misma situación o de la misma población

1.4. La validez de los *cuestionarios*

En psicometría el desarrollo del concepto de *validez* está asociado a la creación y uso de escalas y tests con los que pretendemos medir *un único rasgo*; por eso sumamos las respuestas a todos los ítems.

Si se trata de un *cuestionario* en el que *cada pregunta mide algo distinto,* el planteamiento de la validez no es exactamente el mismo; la validez de un cuestionario se confirma respondiendo a preguntas como estas:

Las preguntas del cuestionario:

- ¿Son *relevantes* para la finalidad que se pretende?
- ¿Hay preguntas *innecesarias* o repetitivas? ¿Se podría *acortar* el cuestionario?
- ¿*Falta* alguna pregunta que aporte información importante para la finalidad del cuestionario?
- ¿Están redactadas con *corrección* gramatical y sintáctica?
- ¿Son *claras* y previsiblemente las van a *entender sin ambigüedad* los sujetos que las van a responder?
- ¿Hay preguntas que incluyen *más de una idea*?
- ¿Son de respuesta *fácil*?

---

[11] La *satisfacción laboral medida con un solo ítem* está especialmente investigada; por ejemplo Gardner, Cummings, Dunham y Pierce, 1998 (citan además otros ejemplos); Wanous, Reichers y Hudy (1997) tienen un meta-análisis en el que la correlación media entre un solo ítem de *satisfacción laboral* y escalas más amplias es de .63 (síntesis de 28 coeficientes de correlación de 17 estudios que comprenden un total de 7682 sujetos); Nagy (2002) encuentra correlaciones entre .60 y .72 entre un solo ítem de *cada variable* de *satisfacción laboral* y otras medidas (con N = 207).

[12] Podemos ver ejemplos en el apartado 13.2, *Sugerencias para obtener datos adicionales que faciliten la validación de la escala.*

En definitiva se trata de confirmar:

1) Que el cuestionario recoge la información de interés pretendida en función del objetivo de la investigación,

2) Que el cuestionario está *bien hecho,* las preguntas y las respuestas son relevantes y claras.

La llamada *validación de expertos* puede tener aquí sentido; no es un *trámite necesario* (porque el que prepara el cuestionario puede ser ya el *experto*) pero puede ser conveniente que alguien más revise el cuestionario[13]. Los *expertos* pueden en su caso ser de dos tipos (puede tratarse de la misma persona):

- Los que tienen ya cierta práctica o conocimiento sobre cómo hacer un cuestionario.
- Los conocedores de la situación, finalidad y contexto en el que se va a aplicar el cuestionario.

2. Los ítems o preguntas de cuestionarios y escalas

Las diversas formas de redactar los ítems (y sus respuestas) son válidas para construir tests o escalas y también para redactar las preguntas de un cuestionario convencional en el que además pueden caber otros tipos de preguntas porque en definitiva preguntamos por el tipo de información que nos interesa conocer.

Diferenciamos en apartados distintos la formulación de los *ítems* (lo que es propiamente la *pregunta*, como quiera que se redacte) y la formulación de las *respuestas*, aunque habitualmente denominamos *ítems* al conjunto *pregunta-respuestas*. Algunas modalidades *piden* que los ejemplos incluyan la pregunta y las respuestas (como en los *ítems bipolares*)[14].

Los ítems o preguntas de una *escala* o de un *cuestionario* se pueden redactar de varias maneras y estilos; aquí exponemos los más comunes.

Según el *tipo de respuesta* se pueden establecer cuatro grandes tipos de preguntas o ítems que iremos exponiendo.

a) Preguntas con *varias respuestas*

Las respuestas (dos al menos) suelen ser en términos de *grado de acuerdo, de frecuencia, de importancia, de intensidad*… Es el tipo de formulación apropiado en el caso de los ítems de las escalas de actitudes.

b) *Listas de ordenamiento*.

A los sujetos se les pide que *ordenen* una serie de elementos o ítems según su preferencia personal; no hay por lo tanto una respuesta independiente a cada uno de los elementos que tienen que ordenar. Estas preguntas que requieren *ordenar* son frecuentes en cuestionarios pero no son adecuadas como ítems en escalas de actitudes.

---

[13] Sobre la *validación de expertos* volvemos a tratar en el apartado 10.4 sobre *Las pruebas piloto y la validación de expertos*

[14] Amplios resúmenes y conclusiones experimentales sobre las diversas modalidades en la formulación de los ítems y de las respuestas en Yorke (2009) y Stapleton, Cafarelli, Almario, y Ching (2010); estos últimos autores se refieren de manera específica a instrumentos (escalas, cuestionarios) para niños y adolescentes.

c) *Preguntas con varias respuestas para escoger todas las que quieran*

A los sujetos se les pide que escojan todas las alternativas que quieran de la lista ofrecida (profesiones, deportes, etc.). Ya veremos que hay modos mejores de presentar estas preguntas.

d) Preguntas de *respuesta abierta*.

También son frecuentes en muchos cuestionarios, a veces para que el sujeto aclare su respuesta a una pregunta previa; también se incluyen ocasionalmente al final de las escalas de actitudes aunque no como ítems de la escala.

Las *listas de ordenamiento, las preguntas con varias respuestas* y las *preguntas de respuesta abierta* no son apropiadas como ítems en las escalas de actitudes. Por lo que respecta a los cuestionarios, vimos en un apartado anterior (1.1) sobre *cautelas iniciales en la construcción de cuestionarios y escalas* que las *listas de ordenamiento* y las *preguntas de respuesta abierta* se desaconsejan como criterio general para no complicar los análisis, pero naturalmente también pueden *tener su lugar*[15].

2.1. Ítems en forma de opiniones

Redactar los ítems en forma de *opiniones* es lo más frecuente cuando se trata de medir *actitudes*, pero caben también otros tipos de redacción. Además el proceso para construir tests para medir *rasgos de personalidad* y otros tipos de variables es el mismo (aunque hay también otros procedimientos) independientemente de que los ítems se redacten en forma de opiniones.

Si construimos un test o escala, las características de los ítems que exponemos más adelante (apartado 6; *relevancia*, *claridad*, *discriminación*) deben mantenerse siempre cualquiera que sea el estilo en que se formulan. Si se trata de un cuestionario (en el que *no* sumamos las respuestas a las diversas preguntas) la relevancia y la claridad son siempre importantes pero no lo es la discriminación, la pregunta puede ser de interés aunque supongamos que todos van a responder casi lo mismo.

2.2. Ítems en forma de conductas o casos

En muchos tests de personalidad y escalas o cuestionarios para evaluar actitudes se incluyen dos tipos de conductas, o propias o ajenas.

a) Conductas o hábitos personales

Por ejemplo, en una escala de *asertividad* (Gismero, 1996):

Muchas veces prefiero ceder, callarme o 'quitarme de en medio' para evitar problemas con otras personas.

En este ítem el *máximo acuerdo* indicaría una *menor asertividad*.

En un sencillo cuestionario para medir los *enfoques superficial y profundo de aprendizaje*[16] tenemos tres *conductas* o *hábitos* referidos al modo de estudiar:

-Yo suelo estudiar subrayando lo más importante
-Aprendo de memoria lo que no entiendo
-Yo no estudio lo que sé o sospecho que el profesor no va a preguntar,

---

15 Ambos tipos de preguntas los comentamos en sus apartados correspondientes (2.7 y 2.9).

16 De Simons, Dewitte, y Lens, (2004); este cuestionario (12 ítems) está traducido al español en Morales (2011, *cuestionarios y escalas*); el primer ítem es un indicador del enfoque profundo y los otros dos reflejan un enfoque superficial.

En este caso las respuestas son en términos de frecuencia (*6 = siempre y 1 = nunca* en los dos primeros ítems y *1= nunca, 6 = siempre* en al tercero).

b) Conductas ajenas o casos

También cabe presentar determinadas situaciones o breves casos ante los que se puede reaccionar de distinta manera, reflejando así actitudes y valoraciones personales, por ejemplo:

-*Un padre falsifica la edad de su hijo para obtener un billete a precio reducido en el transporte público.*
-*Un comprador se queda con el exceso de cambio que por error le han dado en unos almacenes.*

En este ejemplo las respuestas medirían el *nivel ético* del que responde y se pueden expresar en forma de:

Valoraciones, *esta conducta me parece:* *muy bien, regular, mal, muy mal*
Probables conductas personales, *yo lo haría:* *habitualmente, ocasionalmente, nunca*

2.3. Ítems bipolares

Otra forma de redactar los ítems de un cuestionario o de una escala o un test es *describir las dos respuestas extremas* (la más favorable y la más desfavorable). En estos ítems bipolares caben varias modalidades o estilos según estas descripciones extremas estén más o menos elaboradas. Un ejemplo de descripción más elaborada lo tenemos en la figura 5, con dos ítems adaptados de un test que mide el *ver sentido a la vida*[17].

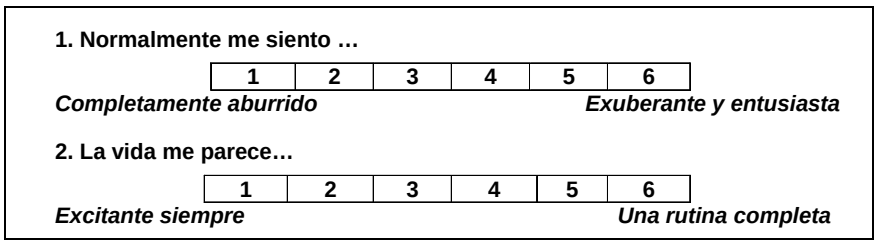


Figura 5

En el *Diferencial Semántico* de Osgood se sigue el mismo procedimiento utilizando pares de adjetivos con significado opuesto, con unos cinco o siete intervalos entre los dos (figura 6):

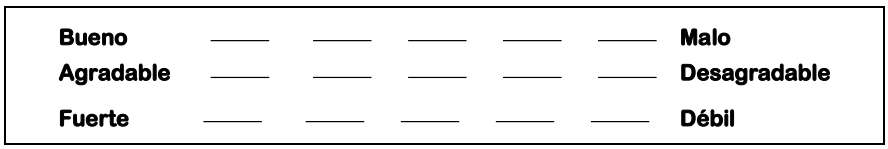


Figura 6

Se utiliza para *valorar* o medir a*ctitudes hacia* cualquier objeto posible de una actitud. Aunque los adjetivos no siempre parecen los más adecuados siempre hay *significados connotativos* que expresan sentimientos valorativos. Pueden localizarse con facilidad listas de pares de adjetivos que pueden servir de modelo[18]. Tres pares de adjetivos utilizados con

[17] *Purpose in Life Test*, de Crumbaugh y Maholic (1969), consta de 20 ítems y está traducido al español en Morales (2011, *cuestionarios y escalas*).

[18] Por ejemplo, en Morales, Urosa y Blanco (2003, pp. 34-39) pueden verse unos 20 pares de adjetivos con significados opuestos y una explicación más amplia del *semántico diferencial*; en Morales (2006, pág. 601) se utiliza un diferencial semántico de 12 pares de adjetivos para valorar el sistema democrático de gobierno; Burden (2008) utiliza 30 pares de adjetivos con significados opuestos para evaluar el *clima* (*'ethos'*) de la Universidad (210 alumnos de primer curso, con bastantes diferencias estadísticamente significativas entre los sexos).

frecuencia para una evaluación rápida de asignaturas o actividades son *fácil-difícil, entretenida-aburrida* y *útil-inútil.*

Este tipo de ítems se utilizan también en planteamientos de *evaluación* en los que no se trata de medir actitudes en sentido propio (figura 7).

1. *El trabajar en este proyecto ha sido una experiencia …*	**5**	**4**	**3**	**2**	**1**	
muy agradable						**nada agradable**
estimulante						**aburrida**
fácil						**difícil**
satisfactoria						**frustrante**
de buen aprendizaje						**pobre de aprendizaje**
muy creativa						**nada creativa**

Figura 7

En el ejemplo de la figura 7 se trata de un cuestionario para evaluar una experiencia de *trabajo en equipo*[19]. En este caso concreto no se sumarían todas las respuestas en una puntuación total; cada par de adjetivos se analiza por separado porque mide características distintas (se puede juzgar que ese proyecto grupal es *fácil* pero también que es *aburrido*) y lo que se pretende es ver cómo los alumnos evalúan este proyecto hecho en equipo *en cada característica.*

Un ejemplo como el de la figura 7 también podría concebirse no como un cuestionario de preguntas independientes sino como una *escala de actitudes hacia los proyectos grupales* si lo que se evalúa no es un proyecto grupal sino cómo se cree que son los trabajos grupales *en general;* en ese caso habría que matizar la introducción a los ítems para que no se refiera a un proyecto concreto sino a todos en general (*el trabajar en proyectos colaborativos es una experiencia*…) y analizar los ítems siguiendo el proceso normal de construcción de escalas de actitudes que veremos en su lugar.

Otra forma de redactar los ítems bipolares consiste en presentar los dos polos como dos alternativas con dos respuestas:

1. En primer lugar hay que *escoger una* de las dos alternativas;
2. En segundo lugar hay que indicar el *nivel de seguridad* en la elección[20].

Un ejemplo lo tenemos en los dos ítems de la figura 8 con los que se pretende medir la *atribución externa o interna del éxito* en los exámenes.

[19] Ítems del cuestionario de Bourner, Hughes, y Bourner (2001). Este cuestionario consta de 17 preguntas, unas cerradas y otras abiertas; se trata de evaluar un proyecto de biología hecho en equipo por alumnos de un primer curso de universidad; el análisis hecho es descriptivo (distribución de frecuencias y porcentajes). Está en español en Morales (2011, *cuestionarios y escalas*). Básicamente el mismo cuestionario para evaluar trabajos de grupo se encuentra en Garvin, Stefani, Lewis, Blumsom, Govier y Hill, (1995) y en Mills y Woodall (2004).

[20] Variantes de este tipo de ítems en Morales (2006, Cap.8).

	Si he hecho un buen examen creo que se debe
A. He estudiado mucho y con constancia **B. He tenido suerte y las preguntas que ha puesto el profesor eran precisamente las que mejor sabía**	**[ ] Casi seguramente a A** **[ ] Probablemente a A** **[ ] Probablemente a B** **[ ] Casi seguramente a B**
A. El examen ha sido en conjunto un examen fácil **B. He puesto interés en estudiar a fondo las cosas, incluso las muy difíciles**	**[ ] Casi seguramente a A** **[ ] Probablemente a A** **[ ] Probablemente a B** **[ ] Casi seguramente a B**

Figura 8

En este ejemplo cada ítem consta de dos formulaciones (A y B), una formulación representa una *motivación interna* (A en la primera, y B en la segunda) y la otra una *motivación externa*. En este caso la clave de corrección va de 4 (*casi seguramente motivación interna*) a 1 (*casi seguramente motivación externa*). Puede haber pares de afirmaciones que no puntúan, por ejemplo si ponemos juntas dos motivaciones externas que se incluyen en el test o cuestionario para disimular algo lo que se pretende medir y facilitar así respuestas sinceras. En el test original se combinan dos atribuciones internas (*estudiar mucho y con constancia* e *interés en estudiar incluso lo difícil*) con tres atribuciones externas (*suerte, examen fácil, profesor benévolo*)[21].

En la figura 9 tenemos otro ejemplo parecido; lo que se pretendería medir es *actitudes éticas* o una actitud de *sinceridad* y *altruismo.*

La situación que se presenta a los sujetos es la de *una entrevista para conseguir un puesto de trabajo;* de las dos afirmaciones de cada ítem, una refleja una postura *sincera y altruista* y la otra *insincera y egoísta.* El sujeto debe indicar 1º qué alternativa es más probable en él y 2º si está muy seguro de su elección[22].

	1º Entre A y B escogería:	**2º de mi elección estoy:**
1. A: Halagar e intentar hacerme amigo de quien me puede ayudar **B: Gastar tiempo y energías en ayudar a personas que no pueden favorecerme en nada**	**[ ] A** **[ ] B**	**[ ] muy seguro** **[ ] poco seguro**
2. A: Decir siempre toda la verdad aunque eso me pueda perjudicar **B: Callar cosas que son verdaderas pero que podrían dar ventaja a un competidos**	**[ ] A** **[ ] B**	**[ ] muy seguro** **[ ] poco seguro**

Figura 9

En la clave de corrección la puntuación máxima (= 4) la tiene la respuesta *altruista y muy segura*, y la más baja (= 1) la respuesta más *egoísta y muy segura*. Con este formato el escoger la alternativa *negativa* pero con *poca seguridad* puede contribuir *a salvar la propia imagen* y constituir a la vez una respuesta suficientemente discriminante.

En la versión original las conductas positivas son dos:

---

21 La versión original consta de nueve pares de ítems de los que sólo seis puntúan; con N = 150 (alumnos de formación profesional) el coeficiente de fiabilidad es .854. Versión completa en Morales (2006, 565-568; 99-100)

22 La escala consta de ocho ítems con una fiabilidad de .75 (con N = 150) y .82 (N = 50) y una correlación de .221 ($p<.01$) con una breve escala de relativismo religioso (N = 150) (Morales, 2006: 581-587).

-Decir siempre toda la verdad aunque eso me pueda perjudicar
-Gastar tiempo y energías en ayudar a personas que no pueden favorecerme en nada,

Estas dos conductas se combinan con cuatro conductas más egoístas:

-Halagar e intentar hacerme amigo de quien me puede ayudar,
-Ocultar verdades que pueden perjudicarme,
-Callar cosas que son verdaderas pero que podrían dar ventaja a un competidor,
-Decir alguna mentira que no perjudique a nadie pero que me beneficia a mí[23].

En estos ejemplos (figuras 8 y 9) las formulaciones de los *dos polos* también se podrían proponer como ítems independientes con los que se podría estar más o menos de acuerdo, pero sería ya un test distinto; el tener que *elegir uno de los dos* puede manifestar mejor la verdadera actitud.

2.4. Listas de adjetivos

Un sistema sencillo de 'medir' tanto rasgos de personalidad como actitudes es preparar una *lista de adjetivos* (o *frases cortas*) que pueden ser más o menos aplicables a grupos, a uno mismo, a determinadas experiencias o actividades, etc. Estas listas se han utilizado con frecuencia para medir actitudes hacia otros grupos (como los *prejuicios*)[24].

Cualquiera que sea el uso que se haga de estas listas de adjetivos:

a) Las respuestas pueden ser dicotómicas (dos alternativas) como *sí* (= 1) o *no* (= 0); *aplicable* (= 1) o *no aplicable* (= 0)*; yo soy así* (= 1) o *yo no soy así* (= 0), etc.

b) Pueden también admitir una gradación en las respuestas (*mucho, bastante, poco, nada*).

Si se trata de escalas o tests formados por adjetivos (o breves autodescripciones), el poner más de dos respuestas es preferible porque casi siempre sube la fiabilidad de todo el instrumento (los sujetos quedan mejor diferenciados).

Estas listas de adjetivos pueden referirse: a uno mismo (*cómo soy*).
a otras personas, situaciones, objetos, etc.

En el uso de adjetivos hay que distinguir dos variantes que explicamos en los dos apartados siguientes:

a) Todos los adjetivos tienen el mismo significado básico.
b) Cada adjetivo tiene un significado distinto

a) Varios adjetivos expresan el mismo rasgo o actitud.

Si los adjetivos van a ser *ítems* cuyas respuestas se van a sumar, todos deben ser indicadores de la misma variable, como en estas dos posibilidades que equivalen o a un *test de personalidad* o a una *escala de actitudes.* En estos casos todos los ítems expresan el mismo rasgo o actitud *o su contrario* (la distinta dirección de los adjetivos se controla con la clave de corrección).

---

[23] En la versión original se presentan 15 ítems de los que solamente 8 puntúan; con N = 150 la fiabilidad es de .750; en otra muestra con N = 50, la fiabilidad es de .82; en muestras distintas se han encontrado correlaciones pequeñas pero significativas; negativas con *edad* (r = -.199, los mayores son menos altruistas), y positivas con *integración familiar* (r = .182). Versión completa en Morales (2006:581-587).

[24] Más información y bibliografía sobre instrumentos construidos con listas de adjetivos en Morales (2006, 57-60)

a) *Test de personalidad*.

Todos los ítems expresan *el mismo rasgo* o su *contrario;* los adjetivos son *autodescriptivos* y pueden equivaler a los ítems de un *test de personalidad.*

Por ejemplo *¿Cómo se describiría Vd.?*

	Nada			*Mucho*
Autoritario	____	____	____	____
Dominante	____	____	____	____
Dócil	____	____	____	____

En un caso como éste, las respuestas irían d 1 a 4; para poder sumar las respuestas hay que ajustar la clave; la respuesta *mucho* a *autoritario* y *dominante* valdría 4, y la respuesta *nada* a *dócil* también valdría 4.

Cuando se trata de medir rasgos de personalidad con series (por lo general breves) de adjetivos, se suelen medir a la vez varios rasgos y los adjetivos que los expresan van mezclados en una lista única, aunque después a cada sujeto se le suman por separado las respuestas a los adjetivos que corresponden a cada rasgo.

Como tests de personalidad son muy modestos y serían cuestionables para hacer un psicodiagnóstico, pero pueden cumplir bien su función con otras finalidades, como sería ver la relación entre determinados rasgos de personalidad y algunas actitudes, valoraciones, etc.[25].

b) *Escala de actitudes*.

Todos los ítems expresan una *valoración* positiva o negativa; una actividad, profesión, deporte, etc. puede ser *bonito, útil, aburrido, interesante, pesado,* etc. Los adjetivos son *valorativos* y equivalen a ítems de una *escala de actitudes* hacia el objeto al que se puedan aplican estos adjetivos.

En estos dos casos *cada adjetivo es un ítem;* la clave de corrección permite sumar las respuestas (aunque unos adjetivos sean positivos y otros negativos). Estas listas se analizan de la misma manera que se analiza un test o una escala de actitudes.

b) Cada adjetivo expresa un rasgo distinto

Por ejemplo *trabajador, constante, culto, artista, paciente, tímido, agresivo,* etc. Estos adjetivos se analizan de manera independiente porque no se pretende que todos expresen el mismo rasgo (más bien todo lo contrario); no se suman las respuestas y no se puede por lo tanto hablar con propiedad de un *test* o *escala;* se trata de cuestionarios sencillos y útiles que suelen utilizarse con estas dos finalidades:

a) Como *autodescripción*; el sujeto que responde señala los adjetivos que cree que le describen[26].
b) Para describir grupos (*nacionales, regionales, étnicos, profesionales,* etc.) y *detectar prejuicios;* un ejemplo típico es el de la figura 10[27].

---

[25] Ejemplos de sencillos *tests de personalidad* (como instrumentos complementarios en una investigación) en los que se mide cada rasgo por medio de una serie de adjetivos pueden verse en Gismero (1996) y en Trechera (1997) (este último reproducido en Morales, 2011, un documento sobre *Análisis Factorial,* que suele ser un procedimiento habitual para seleccionar los adjetivos que miden un mismo rasgo).

[26] Un ejemplo muy breve en la tabla 5, apartado 13.2.1

[27] Utilizado para medir o detectar prejuicios o estereotipos de grupos nacionales (estos grupos no se mencionan en el ejemplo de la figura 10).

	Grupo A	Grupo B	Grupo C	Grupo D
Alegres				
Astutos				
Cordiales				
Desconfiados				
Educados				
Emprendedores				
Engreídos				
Flojos				
Religiosos				
Trabajadores				
Valientes				

Figura 10

Los sujetos que responden se limitan a señalar (con una X) los adjetivos aplicables a los distintos grupos que según su *propia percepción* o según lo que piensan que es *creencia común* (esto hay que dejarlo claro en las instrucciones de respuesta).

Los análisis básicos son muy sencillos, pueden limitarse al porcentaje de respuestas de cada grupo en cada adjetivo. El presentar y cuestionar los resultados al grupo que ha respondido puede ser un buen ejercicio de formación.

2.5. Bloques de ítems del mismo ámbito

En un *cuestionario* nos puede interesar que los sujetos *valoren* con las mismas respuestas una serie de elementos o conceptos del mismo ámbito, como podrían ser *profesiones, actividades, deportes, motivaciones, objetivos personales,* etc.; puede ser de alguna manera *un cuestionario dentro de otro cuestionario*, o un complemento a una escala de actitudes e incluso pueden ser ítems apropiados para construir una escala de actitudes.

Por ejemplo, una serie de *programas de televisión* (cada uno es un ítem: *noticias, deportes, películas, naturaleza,* etc.) pueden ser valorados con estas cinco respuestas:

1	2	3	4	5
lo odio	*lo soporto*	*me entretiene*	*me gusta mucho*	*no me lo pierdo*

Una serie de *actividades* podrían ser valoradas con respuestas semejantes desde *no me gusta nada* hasta *me gusta mucho,* como en el ejemplo de la figura 11 en el que las estudiantes de una residencia tienen que valorar las tareas que tienen que hacer en la misma residencia[28].

	No me gusta nada					Me gusta mucho
tareas	1	2	3	4	5	6
Limpiar los baños						
Arreglar la sala de estar						
Arreglar la sala de estudio						
Atender el teléfono						
Barrer los pasillos						
Ayudar en el lavadero						
Ayudar en la cocina						

Figura 11

[28] Cuestionario de un trabajo académico de Sara Lozano (Universidad Pontificia Comillas, 2º de Psicopedagogía, 1999).

En este formato distinguimos dos posibilidades: a) ítems de un *cuestionario* y b) ítems de una *escala de actitudes*.

a) *Cuestionario*. Si se trata de un cuestionario convencional, cada ítem se analiza de manera independiente (como el de la figura 11; no se le *suman* a cada sujeto todas sus respuestas). Es lo más habitual; precisamente la razón de presentar aquí estos *bloques de ítems del mismo ámbito* como una manera más de formular las preguntas de un cuestionario es para destacar lo siguiente:

1º Es un sistema fácil de hacer y presentar preguntas de interés
2º Se prestan a hacer análisis específicos con los que podemos responder a preguntas en las quizás no hemos pensado.

Además de los análisis puramente descriptivos (como las medias y desviaciones de cada tarea) podríamos al menos responder a estas tres preguntas:

1. *¿Hay entre los elementos valorados diferencias mayores de lo que podríamos esperar por azar? ¿Se puede hablar de una 'jerarquía de preferencias'?*[29]
2. *¿Cuál es el grado de acuerdo de los sujetos [su fiabilidad] al diferenciar unos elementos de otros?*
3. *¿Existe alguna relación entre la valoración que se hace de cada elemento con características del sujeto (como sexo, edad o cualquier otra)?*

No entramos aquí en el *cómo* de estos análisis pero sí es conveniente dejar *puertas abiertas* porque con un cuestionario muy breve semejante al de la figura 11 se pueden hacer análisis muy interesantes[30].

b) *Escala de actitudes*. El que un conjunto de ítems de este tipo puedan formar una escala de actitudes (en este caso a cada sujeto se le suman todas las respuestas) depende del tipo de ítems; en una escala de actitudes se pretende que todos *midan lo mismo* y la suma de las respuestas *debe tener sentido*.

En el ejemplo de la figura 11 podría tratarse de medir *actitud de servicio*; y si en los ítems en vez de tareas tuviéramos deportes podría tratarse de *actitud hacia el deporte*. Pero si se va a construir una escala de actitudes de este estilo hay que seguir *todo el* proceso que veremos en apartados posteriores, no basta con formular los ítems; habrá que hacer el *análisis de ítems* y calcular la *fiabilidad*.

En principio lo más fácil y frecuente es considerar estos bloques de ítems como un *cuestionario* y analizar cada tarea (cada ítem) por separado.

Otra posibilidad distinta de presentar *bloques de ítems* consiste en invertir la figura 11 y colocar *los ítems en las columnas y las preguntas y respuestas en las filas*, como en el ejemplo de la figura 12 en el que tenemos dos ítems (las dos asignaturas).

---

29 Según de qué se trate se podría hablar de una *jerarquía de valores*.

30 A la tercera pregunta (que ya puede justificar el hacer un cuestionario de este estilo) se responde con sencillos *análisis correlacionales* (rápidos y fáciles en EXCEL); a las dos primeras preguntas se responde con *un análisis de varianza para muestras relacionadas* (puede verse explicado en el documento *online* (Morales, 2009).

	Matemáticas	Lengua
1. Comparado con los demás soy muy bueno en …	1 2 3 4 5	1 2 3 4 5
2. Yo suelo tener buenas calificaciones en …	1 2 3 4 5	1 2 3 4 5
3. Me resulta fácil estudiar …	1 2 3 4 5	1 2 3 4 5
4. Me siento desanimado, sin esperanza en…	1 2 3 4 5	1 2 3 4 5
5. Yo aprendo rápidamente en …	1 2 3 4 5	1 2 3 4 5
6. Siempre se me ha dado bien…	1 2 3 4 5	1 2 3 4 5

Figura 12

Este ejemplo es una traducción literal de Corbiere y otros (2006)[31] y se trata de escalas de actitudes en sentido propio (dos escalas, una para cada asignatura) porque se pretende medir en todos los ítems el mismo rasgo (*autoconcepto académico* referido a dos asignaturas; en el ítem 4 hay que invertir la numeración al codificar los datos); los análisis son los propios de las escalas de actitudes.

2.6. Cuando el énfasis está en la medición de *valores*

Cuando el interés está en medir o evaluar lo que habitualmente solemos denominar *valores*, los procedimientos para construir *escalas de actitudes* también son útiles, aunque hay otros tipos de preguntas que quizás se ajustan mejor al concepto de valor, como cuando los sujetos tienen que *elegir entre alternativas* (como el ejemplo de la pregunta de la figura 9) o cuando tienen que *ordenar* diversas alternativas según la preferencia personal.

En este caso (*énfasis en los valores*) es preferible especificar en las respuestas grados de *importancia* más que grados de *acuerdo.* Un ejemplo claro y sencillo son estos ítems sobre *valores en el trabajo* o qué se considera importante en el propio *trabajo profesional* (figura 13)[32].

	Muy importante				Nada importante
En qué medida es importante para Vd. …	5	4	3	2	1
1. Ganar lo suficiente para vivir con mucha holgura					
2. Tener aumentos salariales para mejorar mi nivel de vida					
3. Ganar mucho dinero					

Figura 13

El *valor* expresado por estos tres ítems es la *importancia* que se da a las *ganancias económicas en el trabajo;* la *suma* de las respuestas a estos tres ítems sería la puntuación de cada sujeto en este valor. En el test original se miden 15 valores (*altruismo, buenas relaciones con los demás, independencia, tener puestos de dirección, estabilidad,* etc.) expresado cada uno por tres ítems que tal como se presentan en el cuestionario deben ir mezclados de manera que ítems que miden el mismo valor y que tienen formulaciones parecidas estén convenientemente separados unos de otros.

En cuestionarios más sencillos *cada ítem puede representar un valor distinto* (sin sumar las respuestas a varios ítems que representan el mismo valor). En este ejemplo cada ítem expresa *orientaciones o metas que uno puede buscar en su propia vida* (seis respuestas de *nada importante* a *muy importante*):[33]

---

[31] Citado también en el apartado 8.2 sobre *número de ítems y fiabilidad*

[32] Adaptados del *Work Values Inventory* (Super, 1968). En Internet tenemos disponible el test completo (en inglés, 15 valores y 45 ítems) con su clave de corrección (ver Super en la bibliografía); una versión reducida en español en Morales (2011, *cuestionarios y escalas*).

[33] Ítems tomados de Wilding y Bernice (2006); cuestionario reproducido en Morales (2011, *cuestionarios y escalas*).

-Tener una carrera profesional muy gratificante
-Poder hacer una contribución importante a la sociedad
-Tener una gran seguridad económica
-Tener en la vida un compromiso religioso serio
-Ser muy rico en recursos económicos

Aunque las respuestas en términos de *importancia* se prestan a evaluar valores, no siempre que se utilizan se trata de valores en sentido propio; *son los ítems los que deben expresar lo que entendemos por valores*.

2.7. Listas de ordenamiento

Las *listas de ordenamiento* requieren *ordenar* según su importancia una serie de elementos del mismo ámbito (*colores, profesiones, problemas, objetivos, necesidades*, etc.).

a) Constituyen en sí mismas *otro tipo* de cuestionario con análisis distintos a los específicos de una escala de actitudes o de un cuestionario convencional. Aquí no tratamos de estos análisis que por otra parte pueden tener especial relevancia en la evaluación o medición de los *valores* y de las *preferencias* personales más en general[34].

b) Las listas de ordenamiento no son adecuadas como ítems de *las escalas de actitudes* en las que cada ítem debe tener sus propias respuestas y se responde de manera independiente.

c) En los *cuestionarios,* y como recomendación general, es preferible evitar que los sujetos tengan que ordenar una serie de elementos.[35]

Dos razones para evitar que los sujetos tengan que *ordenar* son estas:

1) Para los sujetos no suele ser fácil ordenar, sobre todo más de seis elementos; si hay que ordenar una lista, ésta debe ser corta.

2) Una razón más importante para evitar listas de ordenamiento es que el análisis estadístico de estas preguntas suele resultar confuso y hay análisis útiles como los correlacionales que no se deben hacer con estos datos o no son de fácil interpretación (aunque sí podríamos calcular el *número de orden medio*).

Si en un cuestionario se incluyen preguntas que requieren *ordenar* hay que tener en cuenta que cuando se ordenan una serie de elementos los sujetos expresan *preferencias relativas,* comparando unos conceptos con otros, y no valoraciones absolutas; lo que queda en último lugar puede ser *muy* estimado, pero *menos* que lo demás; esto, naturalmente, hay que tenerlo en cuenta en la interpretación.

Dos alternativas a *ordenar* una serie de elementos que se pueden considerar, haciendo *preguntas independientes*:

a) Como hemos visto en el apartado 2.6 (*énfasis en los valores*) se puede indicar en las respuestas el *grado de importancia* (de *ninguna* a *mucha*) dado a cada elemento. Con estas respuestas se pueden hacer todos los análisis estadísticos habituales (medias, desviaciones, correlaciones).

El *orden de las medias* en *importancia* suele expresar la misma jerarquía que la obtenida con *listas de ordenamiento*.

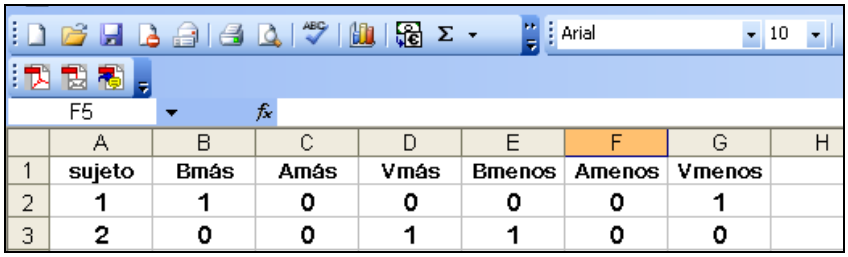
b) Otra alternativa posible es no preguntar por el *orden* sino por *lo más y lo menos* preferido en preguntas distintas; por ejemplo dados varios elementos (tres colores en este ejemplo) se pueden hacer preguntas de este estilo.

	1. *Blanco*	2. *Azul*	3. *Verde*
El que *más* me gusta.....	_____	_____	_____
El que *menos* me gusta.....	_____	_____	_____

[34] De estos análisis tratamos en Morales (2011, *Evaluación de los valores: análisis de listas de ordenamiento*).

[35] Recomendación de Frary (1996, *avoid asking responders to rank responses*); también la experiencia muestra que las preguntas que requieren ordenar complican los análisis; en principio es preferible buscar otras alternativas.

Al copiar las respuestas en la hoja de análisis (EXCEL), cada combinación *color/gusto* es una pregunta distinta, con lo que tendríamos *seis* preguntas (*tres* respondiendo a lo que *más* me gusta y otras *tres* respondiendo a lo que *menos* me gusta) con respuestas *uno* o *cero;* para mayor claridad ponemos un ejemplo en la figura 14.



	A	B	C	D	E	F	G	H
1	sujeto	Bmás	Amás	Vmás	Bmenos	Amenos	Vmenos	
2	1	1	0	0	0	0	1	
3	2	0	0	1	1	0	0	

Figura 14

En este ejemplo al sujeto 1 el color que más le gusta es el *blanco* (Bmás = 1) y el que menos le gusta es el *verde* (Vmenos = 1); al sujeto 2 el color que más le gusta es el *verde* (Vmás = 1) y el que menos le gusta es el *blanco* (Bmenos = 1). Con esta disposición de los datos se pueden calcular fácilmente datos descriptivos (medias y desviaciones), correlaciones entre preferencias por un color y otras variables (sexo, edad, etc.).

2.8. Escoger más de una respuesta

A veces nos interesa que los sujetos escojan todas las alternativas que quieran, por ejemplo, suponemos que esta pregunta es la pregunta número 5 de un cuestionario.

5. *De las carreras puestas a continuación*	Derecho
¿Cuáles piensa que le gustaría estudiar?	Económicas
Señale todas las que quiera.	Educación
	Ingeniería
	Psicología

Estas preguntas, en las que se pueden escoger todas las respuestas que se quiera, es preferible presentarlas de esta otra manera:

5. *De las carreras puestas a continuación ¿Cuáles piensa que le gustaría estudiar?*

	Sí	*No*
5.1. Derecho		
5.2. Económicas		
5.3. Educación		
5.4. Ingeniería		
5.5. Psicología		

Las ventajas de esta presentación son dos:

1) Posiblemente para los sujetos son de respuesta más fácil y rápida;

2) Cada respuesta se convierte en una *pregunta distinta* (con su propia numeración y con respuestas *sí* o *no*): con este formato es más sencillo introducir las respuestas en una hoja EXCEL y hacer los análisis que correspondan.

Las respuestas podrían ser otras (*sí, quizás, no*) pero como alternativa a *escoger todas las que quieran*, las respuestas *sí* y *no* son las más obvias; si se utilizan cuatro respuestas

siempre cabe dicotomizarlas en dos categorías; con más de dos respuestas es más probable encontrar correlaciones significativas con otras preguntas.

Estas preguntas no son apropiadas como ítems en una escala de actitudes pero sí pueden tener interés en cuestionarios más convencionales o como preguntas complementarias en una escala de actitudes.

2.9. Preguntas abiertas

Las preguntas de *respuesta abierta* pueden aportar una información muy valiosa y cualitativamente superior a las preguntas de respuesta cerrada, pero realmente una investigación basada en el análisis de respuestas abiertas es ya *otro tipo* de investigación. Las observaciones hechas aquí se refieren a la inclusión de estas preguntas en los cuestionarios convencionales de respuesta cerrada o puestas como complemento de las escalas de actitudes.

Como *norma general* es preferible evitar estas preguntas de *respuesta abierta*[36]. Las preguntas abiertas, por lo general puestas al final del cuestionario o para que el sujeto matice más alguna de las preguntas cerradas, tienen estos inconvenientes:

1) Es muy normal que muchos sujetos no las respondan o lo hagan de manera muy descuidada; tanto la *habilidad* como la *disposición personal* de los sujetos para responder a estas preguntas puede ser muy distinta; cuando unos responden a estas preguntas y otros no responden, unos se explayan en sus comentarios y otros responden con dos palabras, la información recogida puede estar muy sesgada.

2) El trabajo de leer y sistematizar las respuestas puede ser complicado o excesivo. Es un hecho que con frecuencia estas respuestas abiertas ni se leen, ni se analizan[37].

3) Si se pueden prever las respuestas *más frecuentes* o *importantes* a estas preguntas abiertas, es preferible hacerlas cerradas.

Naturalmente el cómo hacer un cuestionario no es una *ciencia exacta*; por lo menos antes de hacer preguntas abiertas conviene preguntarse si son realmente necesarias. El poner estas preguntas *abiertas* al final de un cuestionario *cerrado* tiene a veces más que ver con la *ansiedad* o *inseguridad* del investigador (que no quiere que *se le escape* nada) que con un planteamiento racional de la investigación.

3. Las respuestas de cuestionarios y escalas

Por lo que respecta a la redacción de las respuestas hay al menos tres temas que hay considerar: las diversas *maneras de redactar* las respuestas, el *número* de respuestas de cada ítem y si se debe poner un número *par o impar* de respuestas.

3.1. Tipos de respuestas

Las respuestas más habituales en las escalas de actitudes suelen expresar el *grado de acuerdo* con el contenido del ítem, sobre todo cuando los ítems expresan *opiniones,* pero puede haber otros tipos de respuestas más adecuadas a la formulación del ítem (como grado de *interés,* de *importancia,* de *frecuencia,* etc.). Otros estilos de formular las respuestas ya los hemos visto al ver otros tipos de ítems como los *bipolares*.

Respuestas típicas en las escalas de actitudes en términos de *grado de acuerdo* y según el número de respuestas que se empleen, pueden ser, por ejemplo, las puestas en la figura 15.

---

[36] Es una recomendación de Frary (1996), sobre todo por la primera razón expuesta Ya está indicado en el apartado 1.1. sobre *cautelas iniciales*.

[37] El *análisis del discurso* o de respuestas abiertas tiene su propia metodología (*investigación cualitativa*).

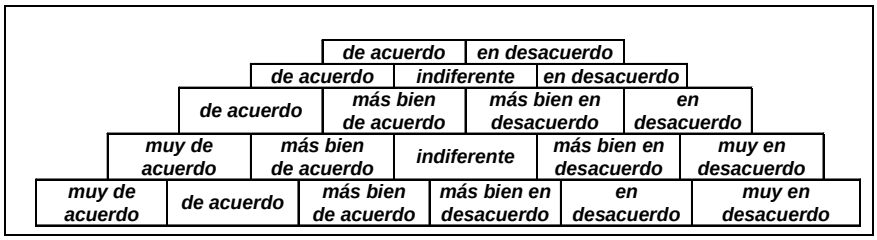


Figura 15

Caben otros *formatos* en las respuestas, como utilizar *números* o *letras* (especificando antes su significado); también es habitual dejar espacios en blanco (recuadros o guiones) indicando con palabras el significado de las respuestas extremas (*muy de acuerdo* y *muy en desacuerdo*) (figura 16).

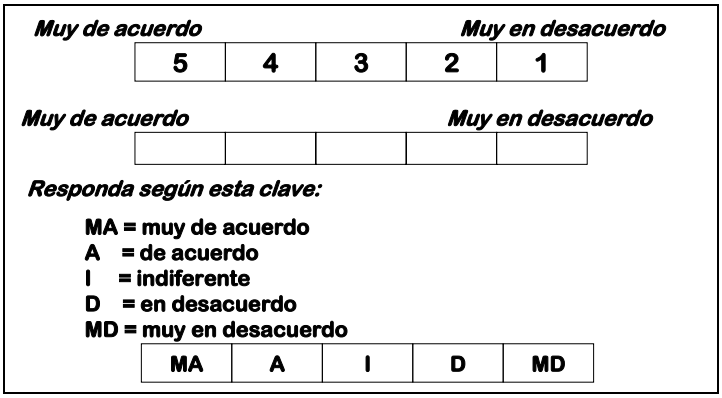


Figura 16

Cuando se utilizan números, todos deben ir en la misma dirección (*muy de acuerdo* siempre tiene el valor máximo), aunque en las escalas de actitudes se inviertan después estos valores en la *clave de corrección* como indicamos más adelante (apartado 7).

Las respuestas en términos de *frecuencia* están especialmente avaladas por la investigación experimental y en concreto se han propuesto las posibles respuestas indicadas en la figura 17[38].

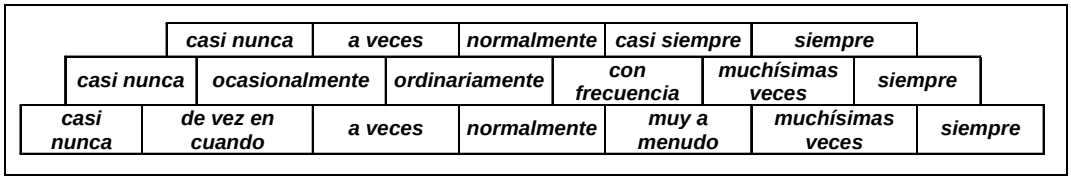


Figura 17

Sobre todo con niños (y siempre que sea *realista* hacerlo) estas respuestas en términos de frecuencia son más claras y contribuyen a aumentar la fiabilidad cuando se expresan de manera más específica, por ejemplo *todos los días, una o dos veces por semana*, etc.; o si se trata de tiempo empleado en ver televisión al día *nunca, menos de una hora, entre una y dos horas, más de dos horas*, etc. En estos casos y para establecer unos *intervalos de tiempo realistas* conviene obtener antes algún tipo de información[39]

[38] La justificación de estas categorías de respuesta, y un listado mayor, puede verse en Cañadas y Sánchez Bruno, (1998).

[39] Stapleton et al. (2010) citan varios estudios que confirman esta mayor fiabilidad cuando las respuestas ofrecidas son más específicas.

Caben otros modos de redactar las respuestas *graduadas de más a menos* que dependerán de cómo estén redactados los ítems; en la figura 18 tenemos cuatro respuestas redactadas con cuatro estilos distintos:

Acuerdo	**Muy de acuerdo**	**Más bien de acuerdo**	**Más bien en desacuerdo**	**En desacuerdo**
Frecuencia	**Siempre**	**Bastantes veces**	**Algunas veces**	**Nunca o casi nunca**
'Cantidad'	**Mucho**	**Bastante**	**Poco**	**Nada**
Seguridad	**Ciertamente sí**	**Más bien sí**	**Más bien no**	**Ciertamente no**

Figura 18

En general las diversas maneras de expresar las respuestas (todas con expresiones verbales, o describiendo solamente las dos respuestas extremas, o utilizando números, etc.) dan resultados semejantes por lo que respecta a la validez y la fiabilidad[40]. Es útil ver *modelos* antes de decidir cómo hacer la presentación definitiva de la escala o cuestionario. A veces el poner las respuestas de una manera u otra depende del espacio disponible. Para el que responde el significado de sus respuestas debe estar muy claro. En general parece preferible utilizar *palabras* en vez de números.

También cabe redactar *respuestas distintas en ítems distintos* aunque formen parte del mismo instrumento o escala, manteniendo siempre en todos los ítems el mismo número de respuestas (un ejemplo en la figura 19)[41]. Sobre todo con niños hay que facilitar el que se identifiquen rápidamente con una respuesta determinada y esto puede requerir respuestas distintas en ítems distintos.

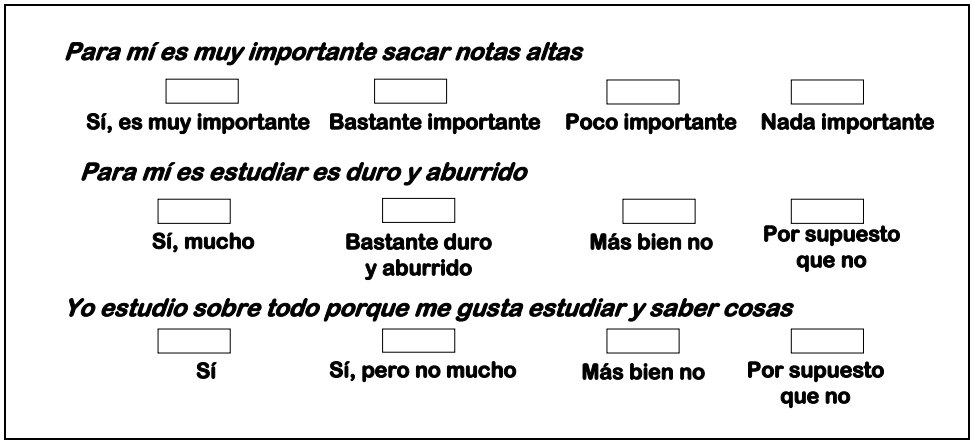


Figura 19

No siempre es fácil encontrar suficientes respuestas verbales bien graduadas y que tengan un significado claro; cabe, como vamos viendo, combinar números y expresiones verbales; e incluso con alguna ayuda de tipo *gráfico*, como en el ejemplo de la figura 20 en el que tenemos un único ítem para evaluar un programa de formación (adaptado de Davies, 2008).

---

[40] Por ejemplo, Chang, 1997, con dos muestras de 173 y 108 sujetos. La fiabilidad *test-retest* (la correlación cuando se responde al mismo test dos veces con un intervalo de al menos una semana) suele ser mayor (mayor estabilidad en las respuestas) cuando todas las categorías de respuesta están expresadas verbalmente (Weng, 2004, con una muestra de 1247 estudiantes universitarios).

[41] En Morales (2006), Anexo III está la escala completa de *actitud hacia el estudio* (otra escala semejante en el Anexo IV).

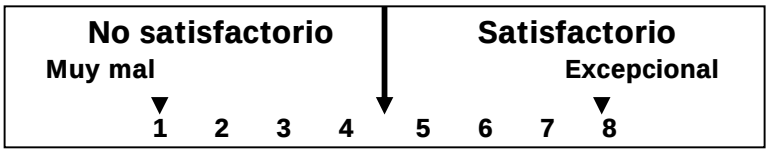


Figura 20

Se definen verbalmente los extremos (*muy mal* y *excepcional*) y para que los sujetos sitúen su respuesta con más claridad, se separan gráficamente las áreas que corresponden a una valoración positiva o negativa (*satisfactorio* y *no satisfactorio*).

3.2. Número de respuestas

En las escalas de actitudes (y en los cuestionarios en general) el número tradicional de respuestas es de *cinco,* pero pueden ser más o pueden ser menos.

Una observación importante ya hecha es que en general, y en el caso de escalas de actitudes, *a mayor número de respuestas en los ítems, suele haber una mayor fiabilidad en toda la escala,* con tal de que el número de respuestas no supere la capacidad de discriminación de los que responden. También con más respuestas se verifican con más facilidad relaciones con otras preguntas porque los sujetos quedan mejor diferenciados.

En las escalas de actitudes resulta más económico aumentar el número de respuestas en vez de aumentar el número de ítems y además se responde en menos tiempo. Seis o siete respuestas suele ser el número máximo habitual; a partir de nueve respuestas ya no se discriminan bien. La práctica más generalizada es poner entre 4 y 6 respuestas.

El número mínimo puede ponerse en tres respuestas; en cualquier caso con tres respuestas suele subir la fiabilidad con respecto a dos nada más. En las escalas de actitudes con dos respuestas (*sí* o *no, de acuerdo* o *en desacuerdo*, etc.) la fiabilidad suele ser menor.

Las dos respuestas (*sí, no, de acuerdo en desacuerdo,* etc.) es preferible dejarlas para ítems de naturaleza dicotómica (*varón/mujer, haber tenido o no una determinada experiencia,* etc.) o cuando lo aconseja la previsible capacidad de comprensión de los sujetos (como puede suceder con niños).

3.3. Número par o impar de respuestas

Una cuestión distinta es si se debe incluir un número *par* o *impar* de respuestas (con o sin una *respuesta central* de indecisión). No hay normas claras sobre este punto; lo más claro es que *son preferibles tres respuestas a dos* (la fiabilidad es casi siempre mayor con tres respuestas que con dos, y además con sólo dos respuestas los que responden pueden sentirse incómodos).

El incluir un número *par* de respuestas (4 ó 6) tiene al menos dos ventajas.

1) En primer lugar siempre cabe la posibilidad de *agruparlas en dos categorías, de acuerdo* y *en desacuerdo,* y esta agrupación en dos categorías puede ser útil para determinados análisis o para presentar los resultados de manera más sucinta.

2) En segundo lugar se elimina la posibilidad de que los sujetos *se evadan* escogiendo la respuesta central (casi nunca hay verdadera indecisión si los ítems son *relevantes* para que los que responden).

Una respuesta central del tipo *indiferente, no sé, indeciso,* puede tener problemas de interpretación y no representar adecuadamente la magnitud o intensidad pretendida (*punto medio* entre las respuestas extremas) porque se puede escoger por razones distintas.

En general esta *respuesta central* funciona mejor si verbalmente se expresa su posición de manera explícita (como *a veces*, entre los extremos *nunca* y *siempre*) (Hernández, Espejo y González Romá, 2006). En cualquier caso no se puede hablar de una norma; una práctica muy habitual es utilizar cinco respuestas.

4. Las escalas de actitudes

En primer lugar exponemos las razones para construir escalas (o tests) y damos una visión preliminar de todo el *proceso* de construcción y de la *estructura* final del cuestionario que responderán los sujetos y en el que está incluida la escala de actitudes.

4.1. Por qué construimos una escala (o un test) en vez de limitarnos a una sola pregunta.

Hay varias razones para construir escalas sin limitarnos a una sola pregunta. Aun en el caso de los cuestionarios sociológicos, que no son escalas de actitudes en sentido propio, puede ser útil y conveniente el disponer de *varios indicadores* de una misma actitud (o de una misma variable, como nivel socioeconómico) que van a ser *sumados* después como indicador de esa actitud o variable.

1º Con una serie de ítems *describimos* y *medimos* mejor constructos relativamente *complejos*.

En la vida cotidiana juzgamos sobre cómo es una persona (si es aficionado a un deporte, si es *más o menos* conservador, religioso o asertivo, si le gustan los temas relacionados con la naturaleza, etc.) en función de de *varios indicadores*, como pueden ser diversas opiniones referidas a la misma actitud o conductas observadas en el mismo sujeto. Con una única pregunta podemos simplificar en exceso el concepto que vamos a medir.

Una *analogía* sobre la conveniencia de hacer *varias preguntas sobre lo mismo* la tenemos en una *consulta médica*. Cuando vamos a hacernos un reconocimiento, el médico suele hacer *varias preguntas*, que no son otra cosa que una *serie de posibles síntomas*, y no es lo mismo de cara a un diagnóstico el responder *sí* a una de sus preguntas que responder *sí* a todas o casi todas sus preguntas; es más seguro disponer de varios indicadores o síntomas de la misma posible patología.

En conjunto una medida compuesta por varios ítems es más *válida* en el sentido de que *expresa mejor la actitud o rasgo* al menos por dos razones:

a) Varios indicadores describen mejor un constructo, rasgo o actitud que uno solo.
b) Una única pregunta puede de hecho ser *poco afortunada*, o equívoca, o ser mal entendida por muchos o algunos sujetos.

Cuando hay varios indicadores de la misma actitud (o rasgo en general) se obvian mejor las limitaciones de cada ítem en particular. Además una única pregunta puede distorsionar la información que el sujeto aporta de sí mismo; por ejemplo uno puede definirse como muy liberal en una única pregunta, porque ésa es la imagen que tiene de sí mismo, pero puede no aparecer tan liberal ante varias cuestiones más específicas.

Como ya hemos indicado al decir qué entendemos por test o escala (*varios ítems miden el mismo rasgo*), podemos pensar en la *medida del rendimiento académico* que quizás es un ejemplo más claro y con el que estamos más familiarizados. Si queremos saber si un alumno sabe química, no le hacemos una sola pregunta porque puede saber esa pregunta pero no otras muchas posibles preguntas, o esa pregunta puede ser de hecho ambigua o muy difícil, etc.; una serie de preguntas sobre el mismo tema o asignatura nos da una idea más certera sobre si sabe *más* o sabe *menos*. Después de todo, nuestra conclusión y nuestro juicio no van a ser sobre si sabe o no sabe unas preguntas concretas, sino sobre si sabe o no sabe *en general*; de

unas pocas preguntas extrapolamos nuestras conclusiones a otras muchas posibles preguntas semejantes.

Algo análogo hacemos con las escalas de actitudes, tests de inteligencia, etc.; una muestra relativamente amplia de preguntas (*ítems*) constituye una mejor base para formarnos un juicio más preciso y fundado sobre *cómo está* una persona (o un grupo representado por su *media*) en un rasgo concreto. Aun así ya veremos que *pocas preguntas* pueden ser suficientes si lo que queremos medir o apreciar está conceptualizado de una manera relativamente simple.

Este tipo de razones para utilizar un test o escala tiene más peso cuando se van a tomar decisiones sobre los sujetos (como por ejemplo admitir o no admitir a un programa de estudios, o a un puesto de trabajo) o interesa hacer un buen *diagnóstico individual.*

2º Cuando hay más ítems aumenta la *fiabilidad* de la medida.

Por *fiabilidad* entendemos ahora lo que significa este término de manera intuitiva; sin entrar en cuestiones de psicometría *fiabilidad* significa *precisión* en la medida, menor margen de error.

Por las razones dichas anteriormente se minimizan las limitaciones de cada ítem en particular; *merecen más confianza* varias preguntas que una sola.

Un solo ítem está más sujeto a los *errores de medición* (respuestas rápidas y distraídas, ítems no bien entendidos); en cambio cuando varios ítems se van a sumar en un total estos *errores de medición* tienden a cancelarse mutuamente o al menos *pesan menos* en ese total.

En un sentido más *psicométrico,* si disponemos de una serie de ítems podemos calcular el coeficiente de fiabilidad como veremos más adelante; los coeficientes de consistencia interna (o de fiabilidad) como el coeficiente *alfa* de Cronbach sólo se pueden calcular si el instrumento consta de varios ítems (dos al menos), no de uno solo. La fiabilidad puede ser alta o baja, pero eso es algo que podemos verificar, de la misma manera que podemos analizar la calidad de cada ítem.

3º Cuando hay más ítems detectamos mejor las *diferencias interindividuales*

Una razón *de peso* para sumar varios indicadores del mismo rasgo es que las diferencias entre los sujetos van a quedar más claras; va ser más fácil ordenar o diferenciar a unos sujetos de otros; en definitiva va a aumentar la *varianza.* De alguna manera *medir es diferenciar:* un test de inteligencia que no diferencia a los más inteligentes de los no tan inteligentes no nos sirve para nada.

4º Cuando hay más ítems detectamos con más facilidad *relaciones entre variables*

El *detectar diferencias* es de interés en cualquier estudio o investigación porque sin diferencias claras en los sujetos es más difícil encontrar *relaciones* entre variables (es decir, si el estar alto o bajo en una variable coincide de hecho con estar alto o bajo en otra variable). Muchas investigaciones se limitan casi exclusivamente a analizar relaciones entre variables.

Por la misma razón, y como hemos indicado el apartado 3.2 (*número de respuestas*) cuando utilizamos preguntas distintas para medir rasgos distintos (una pregunta para cada rasgo, no escalas de actitudes) detectaremos mejor relaciones entre preguntas si cada pregunta tiene *varias respuestas graduadas* (por ejemplo de *mucho* a *nada*) que si solamente tiene dos respuestas (como *sí* o *no*).

4.2. *Fases* del proceso y *estructura* de todo el cuestionario

De los diversos tipos de escalas o tests nos limitamos aquí a las denominadas escalas *tipo-Likert*. Son las más conocidas y se denominan así por el autor que sistematizó el proceso de construcción (en 1932)[42]. En conjunto es el sistema más sencillo y de características no inferiores a los otros tipos de escalas (o son incluso mejores) por lo que es posiblemente el procedimiento más utilizado. Lo que hizo Likert fue extender a la medición de las actitudes lo que ya era común en la medición de los rasgos de personalidad: la suma de una serie de respuestas a ítems *supuestamente homogéneos* (que expresan el mismo rasgo) sitúa al sujeto en la variable medida

Es conveniente tener *desde el principio* una clara visión de conjunto de todos los pasos que integran el proceso de construcción de una *escala de actitudes* o, en general, de cualquier tipo de test. Aunque aquí tratamos de manera más explícita de las escalas de actitudes que de los tests de personalidad o de otro tipo; el proceso de construcción es básicamente el mismo.

Para mayor claridad vamos a distinguir:

1) Las *fases* o pasos sucesivos en la construcción de una escala.
2) La *estructura* que debe tener el instrumento final o *cuestionario completo* que responderán los sujetos.

Una escala de actitudes o un test (el *conjunto de ítems* con el que pretendemos medir una determinada actitud) se completa con más preguntas, unas de información personal y otras que serán necesarias para validar la escala y en general para llevar a cabo la investigación en la que se utiliza esa escala.

La *estructura* final del instrumento coincide en parte con las *fases* del proceso de construcción de la escala aunque la coincidencia no es exacta; el instrumento es el cuestionario que responderán los sujetos pero el proceso incluye una *fase previa* de preparación y *continúa* con los análisis que hay que ir haciendo después.

4.2.1. *Fases* en el proceso de construcción de una escala de actitudes

El proceso para construir una escala de actitudes se puede describir de varias maneras, pero básicamente se puede concretar en los pasos enunciados en la figura 21.

Vemos en esta figura dos columnas; la *fase de preparación* se refiere a la preparación del instrumento y a la posterior recogida de datos. Una vez que ya tenemos la información que necesitamos (los cuestionarios respondidos) pasamos a la *fase de análisis*.

[42] En Morales, Urosa y Blanco (2003, capítulo 1 se exponen de manera sucinta los diversos tipos *clásicos* de escalas (Thurstone, Likert, Guttman, Osgood, etc.); también se encuentra fácilmente información en Internet (por ejemplo Wuensch, 2006).

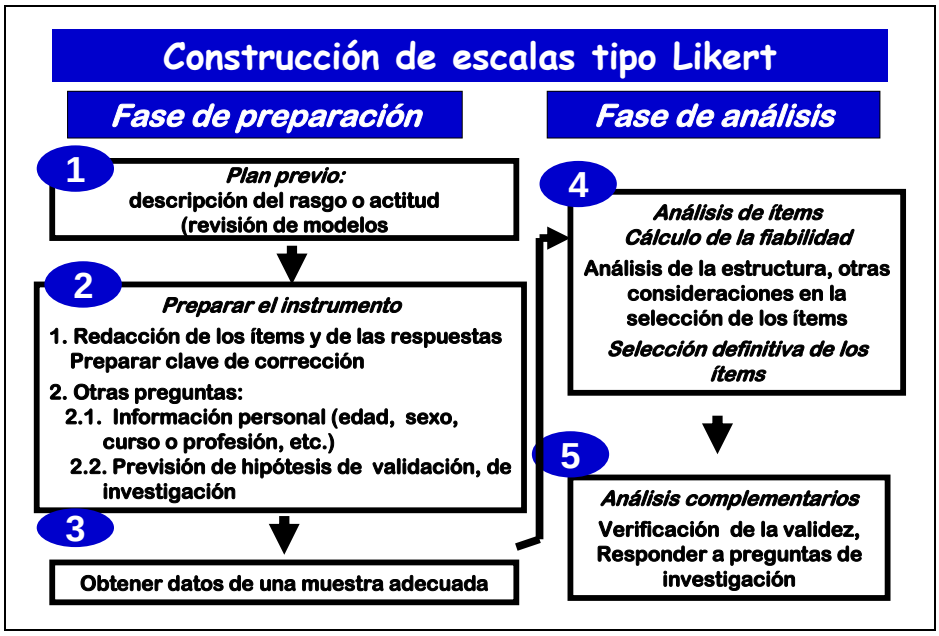


Figura 21

Los análisis indicados en la figura 21 son los típicos en la construcción de escalas de actitudes (análisis de ítems, fiabilidad, validez) pero caben otros muchos análisis. La *validez* (que la escala exprese el rasgo que queremos medir) y la *fiabilidad* (precisión) van a ser dos consideraciones importantes en todo el proceso de construcción de una escala.

4.2.2. Estructura del instrumento

Cuando se construye una *escala de actitudes,* la atención se centra en la redacción de los ítems de la escala, pero esto no es suficiente; además de la escala que se está construyendo se deben preparar *otras preguntas* para obtener información adicional sobre los sujetos.

Estas *otras preguntas* pueden ser otros ítems en sentido literal, pero también puede tratarse de algún otro instrumento (como *otra* escala de actitudes). La estructura del instrumento o *cuestionario completo* que van a responder los sujetos podemos configurarla en tres partes con tres *bloques* de ítems o preguntas, tal como queda esquematizado en la figura 22.

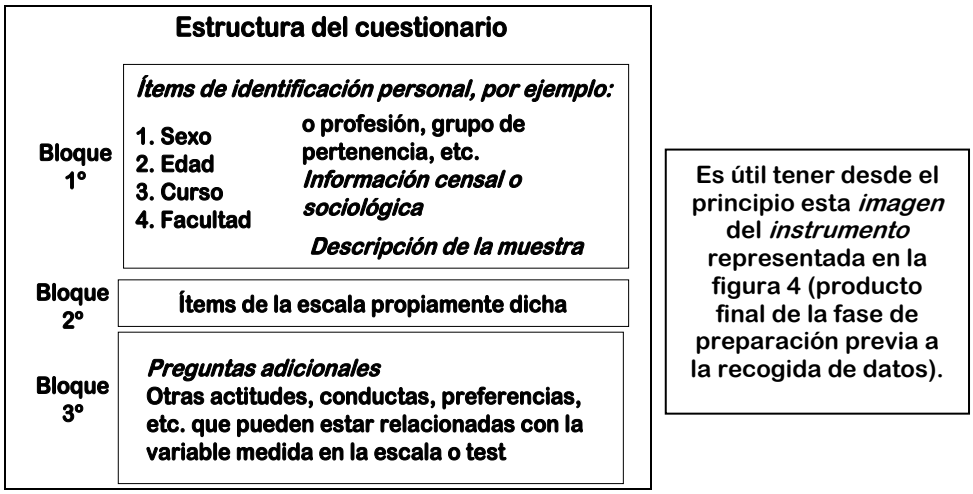


Figura 22

Este modo de ordenar los bloques de preguntas es cómodo y tiene su lógica, pero tampoco es una *norma;* en los cuestionarios convencionales (más que en las escalas) también

se recomienda que la información personal que puede ser más confidencial se deje para el final del cuestionario (Frary, 1996).

Bloque 1 (*información personal*). Las preguntas son *tipo test,* con dos o más respuestas de las que se debe escoger solamente una[43].

Bloque 2 (*escala*). Corresponde a la escala o test de construcción propia (o *ya hecha* y tomada de otras fuentes) que utilizamos para medir la *variable dependiente principal* (el objeto de nuestra investigación).

Bloque 3 (*preguntas adicionales*). Pueden ser otras preguntas o también otras escalas que completan el cuestionario[44].

El *orden* de todas las preguntas del cuestionario es importante porque en este orden se introducen después las respuestas en una hoja de cálculo (como EXCEL) y un orden claro facilita detectar errores y los análisis posteriores.

Todas estas preguntas (ítems) se pueden presentar o con numeración correlativa o en tres (o más) apartados diferenciados y numerados de manera independiente.

5. Puntos de partida en la construcción de una escala

Puede parecer extraño hablar de *puntos de partida*, en plural, en vez de punto de partida, en singular. Realmente el punto de partida es tener una idea más o menos clara de la actitud o rasgo que queremos medir (y así lo tenemos en la figura 21), pero a partir de ahí caben diversas estrategias que se pueden combinar entre sí y que conviene examinar para no encerrarnos desde el principio en esquemas muy rígidos que pueden *bloquear* nuestra creatividad y cerrarnos a otras posibilidades. De una manera u otra el primer paso es siempre tener claro *qué* se desea medir, pero *después* tenemos varios caminos para concretar y dar forma a nuestro instrumento.

Los posibles *puntos de partida* que vamos a comentar son estos cuatro que no se excluyen mutuamente:

1. Definición y *retrato robot*
2. Revisión de instrumentos
3. Traducción de otro idioma
4. Estudio cualitativo previo

5.1. Definición y *retrato robot*

Si se va a construir un instrumento para medir *actitud hacia el estudio,* o *asertividad,* o *inteligencia emocional,* etc., lo primero que suele hacerse es aportar la *definición* del término tal como podemos encontrarla en un diccionario, en un texto o en un autor importante. De muchos términos encontraremos más de una definición y podemos quedarnos con la que nos parezca más adecuada o podemos reformular nuestra propia definición. Esta práctica es correcta y conviene seguirla, pero una mera definición puede ser insuficiente o poco inspiradora para empezar a redactar los ítems porque un mismo término (motivación, autoestima, etc.) se puede concretar de maneras muy distintas; el mismo término *actitud* es muy genérico. Además cabe el que *conceptualicemos* un rasgo sin definición conocida o sin un término acuñado para denominarlo.

No hay que olvidar tampoco que las actitudes, como cualquier otro rasgo que queramos medir, se pueden concebir en *diversos grados de abstracción,* y pueden ser muy genéricos (como *actitud general hacia el estudio)* o muy específicos (como *nivel de aspiraciones, gusto por el estudio, organización del tiempo,* etc., o actitud hacia el estudio de un *determinado*

---

[43] Son las preguntas ya comentadas en el apartado 1.2 sobre *preguntas de identificación personal*.

[44] Información más específica en los apartados 9 *preparar preguntas o instrumentos adicionales* y 13.2 *sugerencias para datos sobre otras variables relacionadas con lo que medimos*.

tipo de materias, etc.). De manera análoga podríamos contar por separado cuántas *naranjas* y cuántas *manzanas* tenemos, o podemos contar cuántas *frutas* tenemos, juntando manzanas y naranjas

El nivel de generalidad lo determina el investigador. A veces se comienza a un nivel muy general pero después, a la vista de lo que *va saliendo*, del análisis de ítems, de las decisiones que se van tomando, hay que *restringir el significado* de lo que estamos midiendo, explicarlo en el lugar oportuno e incluso *cambiar el término para designarlo* y su definición. Podemos comenzar pensando en una escala para medir *actitud hacia el estudio* y al final decidamos que lo que realmente estamos midiendo es *autorregulación en el estudio* o *enfoques en el aprendizaje* o *autoconfianza académica* que pueden ser conceptualizados como constructos distintos, y puede haber otros que también expresan, de una manera u otra, actitud hacia el estudio.

Una buena manera de comenzar a hacer un instrumento como un test o una escala, es hacer en primer lugar una *descripción* de la persona que supuestamente tenga de manera clara la actitud que se desea medir. Esta descripción puede estar basada en una *teoría* ya establecida sobre ese rasgo, o la puede establecer el propio investigador, o se puede apoyar en su propia observación. La descripción puede ser tanto del que tenga una actitud positiva como negativa. Para llamar la atención sobre esta descripción, tenemos puesto en el subtítulo de este apartado *retrato robot*; este retrato robot no sustituye a la definición, pero la aclara y la hace operativa.

Si, por ejemplo, se desea construir una escala para medir *actitud hacia el estudio*, uno puede preguntarse qué caracteriza al que tiene una actitud muy favorable o muy desfavorable hacia el estudio tal como conceptualizamos nosotros esa actitud:

- *Le gusta estudiar,*
- *Tiene un nivel alto de aspiraciones,*
- *No confía en la suerte,*
- *Es organizado y planifica el tiempo de estudio*, etc.

A partir de esta descripción se puede empezar a formular ítems que expresen esas ideas o las contrarias:

- *La preparación de los exámenes la dejo para el final*
- *A mí me basta con aprobar el curso,*
- *No estudio lo que sospecho que no van a preguntar en el examen*, etc.

Este *retrato robot* ayudará a redactar los ítems en torno a un plan coherente. Interesa desde el principio asegurar la *validez conceptual* (que *los ítems midan lo que pretendemos medir*) y poder justificarla. Este *retrato robot* se puede pensar en función de opiniones o conductas con las que previsiblemente estaría de acuerdo el que tuviera una actitud muy positiva (o muy negativa), de actividades que le gustaría hacer, etc.

5.2. Revisión de instrumentos

Puede ayudar también el *revisar otros instrumentos* ya hechos para medir el mismo o parecido rasgo o actitud. En cualquier caso, en un trabajo de investigación en el que se utiliza un instrumento (escala) de nueva construcción, es informativo mencionar otros autores y otros instrumentos hechos para medir básicamente el mismo rasgo aunque sea con otros matices.

Si se utiliza una escala ya hecha (o varias) como *fuente de inspiración*, o se toman determinadas ideas o ítems de uno o más autores (práctica habitual), se debe hacer constar en el texto y en las referencias bibliográficas. Esta revisión de instrumentos puede llevarnos a *traducir* o *adaptar* uno de otro idioma, que suele ser el inglés.

Una *revisión de instrumentos* es útil también en otras situaciones y con otras finalidades.

1) *Concretar el objetivo o tema de la investigación*.

Cuando uno está interesado en crear un instrumento como una escala de actitudes (frecuentemente en el contexto de concretar el tema de una investigación o tesis) pero se encuentra un tanto indeciso y sin acabar de ver qué le podría interesar o le gustaría medir, una revisión de instrumentos *puede ayudar a descubrir o especificar mejor qué es lo que podría hacer*.

2) *Localizar información para el marco teórico*

Normalmente estos instrumentos se localizan en artículos de revistas académicas (de educación, psicología, etc.) en los que también suele encontrarse un *marco teórico adecuado* y bibliografía complementaria; también hay publicaciones con *colecciones de instrumentos* (ver bibliografía).

5.3. Traducción de otro idioma

Hemos mencionado la *revisión de instrumentos* y la *adaptación* de algún instrumento ya hecho como un *punto de partida*. Otro punto de partida puede ser *traducir* directamente un instrumento hecho originalmente en otro idioma (que suele ser el inglés) y que también podemos encontrar ya traducido. En este caso (muy frecuente) ya no hablamos de la *construcción* de una escala (o test) pues ya está hecha. *Traducir* y *adaptar* pueden en la práctica ser *casi* sinónimos; a veces se denomina adaptación a una traducción con ligeras modificaciones; en cada caso hay que especificar de qué se trata: tomar ideas, adaptar o reformula algunos ítems, etc., o literalmente *traducir*.

Traducir o utilizar traducciones ya hechas (tanto de cuestionarios como de escalas o tests) es una práctica muy frecuente por diversas razones[45].

1. Nos puede dar *seguridad* y con mayor razón si se trata de un autor conocido y de prestigio, la han utilizado otros que podemos citar y hay datos sobre su fiabilidad y validez.
2. Nos libera de la tarea de tener que hacer un nuevo instrumento (tarea que podemos juzgar más complicada de lo que realmente es).
3. Además es posible que una traducción sea mejor aceptada que un cuestionario o escala de confección propia por quien tiene que aprobar nuestro instrumento (cuando es ése el caso).

Todas estas razones pueden ser válidas pero hay que tener presentes algunas cautelas porque utilizar instrumentos traducidos no siempre es la mejor práctica. En esta línea hacemos algunas observaciones.

a) Con frecuencia se traduce un instrumento de otro idioma (o se utiliza una traducción ya hecha) porque el *término* para designar el instrumento concuerda con lo que queremos medir nosotros (*motivación, funcionamiento familiar, actitud hacia el estudio, satisfacción laboral,* etc.). Sin embargo con un mismo término se pueden designar instrumentos muy distintos.

[45] El *traducir* un instrumento con una finalidad académica o de investigación (es decir, con una finalidad *no comercial*) es una práctica común y no suele constituir mayor problema (citando las fuentes).

Además del *nombre* hay que *examinar el contenido de los ítems* para verificar si su formulación expresa *lo que realmente queremos medir* y si es apropiado *para nuestros sujetos*: unos ítems nos pueden parecer apropiados, otros muy inapropiados y también podemos echar otros de menos.

b) No es siempre la mejor idea utilizar un test o escala *pensados* para una cultura distinta, y a veces *muy* distinta. Los *ítems indicadores* de la actitud pueden ser relevantes en unas culturas y ambientes pero irrelevantes, y a veces incomprensibles, en culturas y ambientes (o edades) distintos. Ya no se trata solamente de la formulación literal de los ítems, sino de sus *connotaciones,* que pueden remitir a realidades muy diferentes. Los sujetos responden *desde* su situación; las normas grupales y la percepción del entorno pueden diferir mucho entre culturas por lo que las mismas respuestas pueden tener significados distintos; uno se considera *alto* o *bajo* según con quién se compare, de la misma manera que los indicadores de una *buena salud* no son los mismos a los 30 años que a los 80[46].

c) La información aducida para confirmar la *validez* de un instrumento utilizado en otro idioma y en otra cultura se puede asumir en principio, aunque es una información que hay que saber valorar y, a poder ser, replicar. Lo que no se puede *traducir* es la fiabilidad, que hay que calcularla en cada nueva muestra.

d) Una traducción *bien hecha* (hay procedimientos establecidos para hacer una traducción correcta)[47] es especialmente útil para hacer *comparaciones interculturales* en las que interesa que sujetos de distinta cultura y distinta lengua materna respondan al mismo instrumento; éste no es el caso más frecuente pero puede ser de interés e incluso ser el objetivo de una investigación.

Una obvia conclusión de estas consideraciones es que con frecuencia puede ser preferible *crear un nuevo instrumento* (que puede estar *inspirado* en otros de los que tomamos ítems e ideas y que citamos oportunamente) más adaptado a nuestra población y a nuestras intenciones que utilizar una traducción simplemente porque está disponible o aparentemente resulta más sencillo.

Puede haber cierta resistencia a crear instrumentos nuevos cuando realmente puede ser una tarea relativamente sencilla y más creativa y adaptada a nuestros sujetos y a nuestra situación.

5.4. Estudio cualitativo previo

Por estudio cualitativo se entiende aquí algo tan sencillo como *consultar con otros*, individualmente o en grupo. Por ejemplo la *primera redacción* de los ítems se puede hacer con la ayuda de un grupo utilizando una sencilla *tormenta de ideas* (*brainstorming)*. Si vamos a construir una escala de *satisfacción laboral* se pregunta al grupo ¿Con qué afirmaciones estaría de acuerdo una persona *muy satisfecha* o *muy insatisfecha*? Con este procedimiento se puede generar un número grande de ítems y a continuación se pueden rechazar o reformular los ítems ambiguos y simultáneamente se establece la clave de corrección. El análisis de ítems posterior (apartado 12.1) nos dirá qué ítems se pueden retener.

---

46 Esta analogía es de Heine, Lehman, Darring, Peng y Greenholtz (2002) que analizan estos problemas interculturales en una escala de *independencia-interdependencia.*

47 Una traducción deficiente puede cambiar el significado de un ítem. Para traducir un instrumento de un idioma a otro hay procedimientos correctos que se deberían al menos consultar (pueden verse en Morales, Urosa y Blanco, 2003, 72-76); un artículo clásico sobre este tema es el de Hambleton y Patsula (1999) para quienes es un *mito* la creencia de que siempre es preferible *traducir* o *adaptar* de un idioma a otro en vez de *crear* un test o escala nuevos.

Este enfoque puede ser interesante sobre todo para que el instrumento sea *claro* y *relevante* precisamente para el tipo de *población* y de *situación* al que va a ser aplicado. Aunque este enfoque no sea siempre el más adecuado *para comenzar*, sí es muy conveniente (en muchos casos al menos) que el instrumento *provisional* sea revisado por una muestra de sujetos semejante a la que va a utilizar después en el estudio definitivo; formulaciones claras para un adulto pueden no ser tan claras para una muestra de niños.

6. Características de los ítems de las escalas de actitudes

La primera redacción de los ítems de una escala de actitudes suele ser muy espontánea, y es bueno dejar que funcione la propia intuición sin preocuparse mucho de una redacción muy cuidada que puede frenar la inspiración, pero en un segundo momento habrá que depurar las ideas y formulaciones iniciales y darles forma teniendo en cuenta las características que deben tener los ítems.

6.1. En forma de opiniones

En general, y sobre todo tratándose de escalas de actitudes, los ítems suelen formularse en forma de *opiniones* con las que se puede *estar más o menos de acuerdo*. y a través de las opiniones con las que un sujeto está de acuerdo podemos inferir la actitud subyacente. No deben por lo tanto formularse en forma de hechos o datos que se puedan *saber o no saber*, pues no se trata de *medir ciencia.*

Podemos definir las o*piniones* como *actitudes verbalizadas*.

No siempre se formulan los ítems exactamente en forma de *opiniones*, también pueden presentarse como *preguntas* en sentido propio (entre signos de interrogación pero con las mismas respuestas habituales). Ya hemos presentado otros estilos y ejemplos de redacción (apartado 2) que en definitiva también equivalen a *opiniones* (se pide al sujeto que exprese lo que *opina, cree, piensa*…)[48].

Las normas o recomendaciones que damos a continuación deben mantenerse cualquiera que sea la modalidad que escojamos para redactar los ítems.

6.2. Relevancia y claridad

Los ítems deben ser *relevantes* y expresar *claramente* la actitud que se desea medir. La claridad es importante; deben redactarse de manera que todos, en la medida de lo posible, los entiendan de la misma manera.

Para que las preguntas (*ítems*) sean claras hay que tener en cuenta al menos estas tres precauciones relacionadas con expresiones 1) *negativas*, 2) *universales* y 3) que incluyan *más de una idea*. Estas cautelas son válidas para cualquier tipo de cuestionario.

1º Hay que tener cuidado con *expresiones negativas* (como *no*); se pueden incluir pero hay que prestar atención a la posible confusión que pueden crear en el que responde, y con más razón si se trata de *dobles negaciones*.

Las opiniones negativas también se pueden expresar en forma positiva, como *me aburre* ver televisión en vez de *no me gusta* ver televisión.

Estos adverbios negativos (*no*), si parece oportuno incluirlos (en principio es preferible evitarlos), pueden ir subrayados o en negrita para que el que responde entienda bien lo que se le pregunta.

[48] El diccionario de la Real Academia Española define una opinión como *dictamen o juicio que se forma de algo cuestionable* o *fama o concepto en que se tiene de alguien o algo.*

2º Hay que evitar expresiones *universales* como *nunca* o *siempre* porque no suelen ser discriminantes (fácilmente las aceptan o rechazan todos los sujetos), lo mismo que adverbios como *solamente*, que además pueden introducir ambigüedad.

3º Deben contener *una única idea,* pues cuando hay más de una idea se puede estar de acuerdo con una y no con la otra (como *las matemáticas son 'muy importantes' pero también 'muy difíciles'*, en una escala de actitudes hacia esta asignatura)

Los errores y ambigüedades en la formulación de los ítems suelen manifestarse en el análisis de ítems que veremos más adelante; la irrelevancia o ambigüedad puede sospecharse cuando demasiados sujetos escogen la respuesta central o cuando los ítems no discriminan (todos los sujetos responden casi de la misma manera). En cualquier caso los posibles errores indicados hay que evitarlos desde el principio.

Es aconsejable que la primera redacción provisional la revise más de una persona para detectar fallos, sugerir nuevos ítems o mejorar la redacción de los ya

6.3. Discriminación

Los ítems de una escala de actitudes deben ser *discriminantes*, es decir, se deben redactar de tal manera que *previsiblemente* unos sujetos estarán de acuerdo y otros no, o no tanto. El que los ítems discriminen es importante en tests y escalas; es distinto si se trata de un cuestionario con el que no se pretende diferenciar a los sujetos; en este caso preguntamos *lo que nos interesa conocer*, sin buscar diferencias en las respuestas.

Para *medir* necesitamos encontrar *diferencias*; los ítems que todos o casi todos acepten o rechacen no van a contribuir a la fiabilidad de la escala y serán eliminados en el análisis de ítems, o simplemente son inútiles (porque *medir es diferenciar*). Además tenemos más garantía de que los ítems *miden lo mismo* (expresan el mismo rasgo) *si simultáneamente diferencian a los mismos sujetos*. Naturalmente verificaremos después (en el *análisis de ítems*) si los ítems de hecho discriminan o no discriminan, pero ya al redactarlos debemos procurar que sean discriminantes.

El que los ítems discriminen es importante en tests y escalas; es distinto si se trata de un cuestionario con el que no se pretende diferenciar a los sujetos; en este caso preguntamos *lo que nos interesa conocer*, sin buscar diferencias en las respuestas.

La *no discriminación* que puede mostrar el análisis de ítems de una escala de actitudes puede indicar que un ítem no mide lo mismo que los demás o que los sujetos lo entienden de hecho de otra manera. Si un ítem discrimina en sentido contrario (puntúan más alto en *ese* ítem los que en el conjunto de la escala puntúan más bajo) puede haber un error en la clave de corrección.

6.4. Equilibrio entre ítems positivos y negativos

Es preferible redactar los ítems en las dos direcciones, unos en la dirección *positiva* y otros en la dirección *negativa,* es decir, que el estar de acuerdo con un ítem unas veces manifieste una actitud *favorable* y otras veces manifieste una actitud *desfavorable* pero sin utilizar adverbios negativos como *no*; por ejemplo:

En una escala de actitud hacia el estudio:

Opinión *favorable* hacia el estudio: *Estudiar es divertido*
Opinión *desfavorable* hacia el estudio: *Estudiar es aburrido*

En una escala de actitudes hacia las formas democráticas de gobierno[49]:

[49] La escala completa de *actitudes hacia la democracia* en Morales (2006, Anexo XI; 2011 *cuestionarios y escalas*).

Opinión *favorable* hacia la democracia:	*En un sistema democrático los ciudadanos pueden realmente cambiar la sociedad a través de sus votos*
Opinión *desfavorable* hacia la democracia:	*Las elecciones realmente libres son un mito: siempre son manipuladas por unos pocos*

El redactar los ítems *en ambas direcciones* tiene ventajas específicas aunque *no es una norma de obligado cumplimiento;* son muchas las escalas cuyos ítems se redactan en una sola dirección, por lo general en la dirección positiva (favorable al objeto de la actitud). La redacción en ambas direcciones, y mejor si la proporción de ítems en ambas direcciones es la misma, tiene en principio estas ventajas[50].

1. Obliga a una *definición previa más matizada* del rasgo o *constructo*.
2. Requiere una *atención mayor* por parte del que responde y *se controla mejor la aquiescencia* o la tendencia a mostrar acuerdo con todo independientemente del contenido del ítem. Esta es la razón más importante que suele aducirse para formular los ítems en las dos direcciones,
3. Permite comprobar la *coherencia* de las respuestas verificando si los dos tipos de ítems correlacionan positivamente.

Esta coherencia se verifica con facilidad cuando hay un *número aproximado* de ítems positivos y negativos (mejor si el número de ítem es el mismo en las dos direcciones); en este caso podemos hacer lo siguiente:

1º Sumamos a cada sujeto sus respuestas a los dos tipos de ítems (cada sujeto tendrá dos totales),

2º Calculamos *la correlación entre los dos totales parciales* como si se tratara de dos subescalas o dos tests.

Si esta correlación alcanza un valor en torno .50 o más, estará indicando una suficiente *coherencia global* en las respuestas[51]. Comprobar la coherencia es prácticamente lo mismo que comprobar que no está operando la *aquiescencia* o tendencia a mostrar acuerdo casi con cualquier afirmación, incluso con afirmaciones que se contradicen (cuando se da esta *aquiescencia* suele deberse a ambigüedad en la redacción, falta de claridad; etc., se da más en niveles educacionales bajos)[52].

6.5. Ítems negativos y discriminación

Los ítems que expresan una actitud negativa (la *máxima puntuación* corresponderá al *máximo desacuerdo*) merecen un comentario especial porque con frecuencia los ítems más discriminantes, los que mejor diferencian a unos sujetos de otros son precisamente los que tienen una formulación negativa, es decir, *cuando la puntuación más alta corresponde al mayor desacuerdo con la opinión expresada en el ítem.* Esto sucede sobre todo cuando se

---

[50] Las ventajas de una formulación de los ítems en ambas direcciones (favorable y desfavorable) y sin utilizar expresiones negativas (como *no*) está muy investigada, por ejemplo Barnette (2000) que recomienda este formato, sobre todo para evitar la tendencia a dar *respuestas aquiescentes*.

[51] Esta magnitud aproximada (.50) del coeficiente de correlación entre las subescalas formadas por los ítems positivos y negativos es una orientación de Cronbach (1960:446).

[52] En la medida de constructos psicológicos *positivos* (*resiliencia* en este caso) Friborg, Martinussen y Rosenvinge (2006) proponen el uso del Semántico Diferencial para evitar la *aquiescencia*, con buenos resultados psicométricos en una muestra de 334 estudiantes universitarios; en los ítems del Semántico Diferencial se incluyen siempre las dos formulaciones, *positiva* y *negativa* (adjetivos de significados opuestos). Un tratamiento amplio de la aquiescencia y otros problemas metodológicos puede verse en Morales (2006, cap. 5).

trata de medir actitudes *socialmente aceptables;* en este caso la respuesta puede estar muy condicionada por la *aceptabilidad social* de la actitud.

Por ejemplo, en una escala de *actitudes hacia la conservación de la naturaleza* (un constructo o actitud *popular*) previsiblemente casi todos los sujetos estarían de acuerdo con ítems de este estilo:

-Es importante conservar la diversidad biológica,
-Es deseable hacer estudios de impacto ecológico antes de construir una carretera, etc.

Entre los que responden puede haber diferencias reales en su actitud hacia la conservación de la naturaleza que no se detectan con este tipo de ítems, sin embargo podemos encontrar diferencias en ítems como estos:

-Los lobos están bien en lugares acotados, donde no pueden causar daños al ganado
-No se debe detener el progreso de una comarca con la excusa de proteger a unos pájaros;
-El conservar la fauna y otros recursos de valor estético, está bien con tal de que no se perjudique a nadie.

En estos ítems la máxima puntuación (actitud *más favorable* a la conservación de la naturaleza) correspondería al *máximo desacuerdo*. Lo que caracteriza a estas formulaciones es que *se incluye una razón o excusa para mostrar desacuerdo* aunque el objeto de la actitud tenga una valoración social muy positiva.[53]

Este otro ejemplo está tomado de una escala de *actitudes hacia el trabajo cooperativo* pensada para los profesores de un colegio (Martínez, 2008). La eficacia del trabajo en pequeños grupos o equipos de alumnos está muy comprobada experimentalmente y sobre todo en algunos ámbitos académicos es muy popular, se imparten seminarios o talleres para formar a los profesores en estas técnicas y puede ser difícil para algunos profesores manifestar (aunque sea anónimamente) opiniones que van en contra (o no tan a favor) del sentir común.

En este caso (*actitudes hacia el trabajo cooperativo*) entre los mejores ítems, los que más diferencian a unos profesores de otros (todos tienen trabajos en grupo con sus alumnos), tienen en su mayoría una formulación negativa, que incluye alguna *excusa* para justificar una actitud opuesta (o simplemente *no muy favorable*) al trabajo en grupo, por ejemplo:

-Creo que diseñar actividades cooperativas lleva tanto tiempo que no compensa.
-Los estudios y trabajos deben ser individuales, ya que los que se realizan en grupo sólo sirven para enfadarse con algún compañero.
-Los trabajos en grupo no son un buen método porque unos trabajan más y otros menos.
-Los estudios y trabajos deben ser individuales, ya que los que se realizan con compañeros sólo sirven para perder el tiempo.

Todas las *excusas* incluidas en los ítems pueden ser razonables y ajustarse a la realidad, pero no en la misma medida para todos y es ahí donde se manifiestan las diferencias en *actitudes* personales.

En este ejemplo específico estos ítems son también los que tienen mayores correlaciones con *conductas didácticas* que favorecen el trabajo en grupo. En este caso las

---

[53] Estos dos ítems están tomados de una escala para medir *actitudes hacia la conservación de la naturaleza* que consta solamente de 10 ítems, todos en la misma dirección desfavorable, y que discrimina muy bien (la versión completa de esta escala en Morales (2006, Anexo XIII) y también en Morales (2011, *Cuestionarios y escalas*). Otro ejemplo de test con todos los ítems negativos (en la dirección opuesta) para medir *depresión* en niños y adolescentes (13 ítems muy sencillos, con tres respuestas, 2 = *verdadero*, 1 = *verdadero a veces* y 1 = *falso*) en Messer, Angold, Costello, Loeber, Van Kammen, y Stouthamer-Loeber, (1995).

relaciones son negativas: a *mayor desacuerdo* con estas opiniones, *mejor disposición* a organizar trabajos en grupo.

No hay que redactar todos los ítems de esta manera (la recomendación más habitual es incluir opiniones favorables y desfavorables) pero estas formulaciones en sentido *desfavorable* suelen ser eficaces cuando se trata de medir actitudes sobre cuya bondad hay un gran *consenso social* y es por lo tanto probable que las respuestas tengan más que ver (o bastante al menos) con lo que es *socialmente aceptable* que con lo que realmente sienten los sujetos que responden.

6.6. Formulación de los ítems en función de los componentes de las actitudes

A veces se recomienda formular los ítems de una escala de manera que distintos ítems reflejen los tres componentes que suelen distinguirse en las actitudes, *conocimientos, sentimientos* y *conductas.*

Esta estrategia requiere un cierto análisis crítico porque, aunque puede ser útil tener a la vista los componentes clásicos de las actitudes, no se puede proponer como el procedimiento *necesariamente* más adecuado para construir escalas de actitudes. Además la experiencia da que cuando se hace un plan previo según estos componentes y se redactan ítems que en un número aproximadamente igual reflejan cada componente, el análisis de ítems posterior suele desequilibrar el plan inicial. Aun así, y si el constructor de la escala se lo propone, sí es posible construir escalas compuesta por ítems que reflejan de manera equilibrada *conocimientos* (es más apropiado hablar de *creencias*), *sentimientos* y *conductas*[54].

Estos tres componentes de las actitudes merecen un comentario diferenciado en relación a la construcción de escalas.

a) *Conocimientos*

Por *conocimientos* (en cuanto componente de una actitud) no se entiende lo que un sujeto *sabe*, sino lo que *cree* que es cierto aunque objetivamente no lo sea. Esta distinción es importante porque con una escala de actitudes en ningún caso se trata de verificar que se conocen hechos o datos objetivamente ciertos; un sujeto puede tener una actitud muy positiva hacia el deporte (*me gusta hacer deporte, o ver deporte en televisión*) y saber muy poquito de deportes (por ejemplo quién ganó un campeonato importante el año pasado); en cualquier caso no se pretende verificar conocimientos en sentido propio, sino actitudes.

Lo que sí verdad es que una actitud positiva hacia un objeto de la actitud, por ejemplo una actitud positiva hacia el estudio de las ciencias naturales, suele ir acompañada de unos *conocimientos* sobre ciencias naturales mayores de lo que puede considerarse normal, pero los *meros conocimientos* no reflejan una actitud (uno puede saber mucho de algo porque lo ha estudiado para obtener un buena nota en una asignatura).

Escalas y tests que miden actitudes hacia otros grupos (*estereotipos, prejuicios*) suelen tener ítems sobre *cómo son los otros*, por ejemplo con una lista de adjetivos, pero en estos casos las respuestas reflejan *creencias comunes* más que conocimientos, que pueden ser parciales, falsos o estar deformados.

Las actitudes actúan como *filtros* de la información y del conocimiento; tendemos a *ver* la información que confirma nuestras actitudes dejando *fuera* aspectos de la realidad que no concuerdan con nuestras actitudes *previas.*

54 Tenemos un ejemplo en Loureiro y Lima (2009) que construyen una escala de *actitudes altruista*s de 12 ítems en la que cada componente está representado por cuatro ítems; con N = 213, los coeficientes de fiabilidad son .65 (*componente cognitivo*), .81 (*componente afectivo*) .70 (*componente conductual*) y la fiabilidad de la escala completa es de .79

Como se puede esperar que una actitud positiva sí esté relacionada con saber más sobre el objeto de la actitud, preguntas sobre lo que uno *cree que sabe,* o directamente de conocimientos (con respuestas que pueden ser objetivamente correctas o incorrectas) pueden ser un buen complemento a (*complemento a* pero no *parte de*) una escala de actitudes para verificar la hipótesis de que los que tienen una actitud positiva hacia *algo* saben más sobre ese *algo*. También podrían utilizarse (y se utilizan de hecho) *notas escolares* con esta finalidad; los que muestran una actitud positiva hacia las matemáticas probablemente tienen mejores notas en matemáticas[55]. La relación entre actitud y conocimientos (en cuanto *ciencia*) puede ser un dato a favor de la *validez* de la escala.

b) *Sentimientos*

Los *sentimientos* entendidos en un sentido amplio (*agrado, me gusta, estoy a favor de,* etc.) son posiblemente el *componente formal* de las actitudes (*predisposición a reaccionar a favor o en contra*) y se prestan a formular buenos ítems que sí reflejan la actitud del que responde. Una simple lista de adjetivos con *connotaciones valorativas positivas o negativas* (controlados en la clave de corrección, como *bueno, útil, aburrido, complicado,* etc.), que expresan *sentimientos espontáneos,* pueden constituir tanto una sencilla escala de actitudes en sentido propio como *otra manera* de verificar la actitud medida con una escala más compleja.

c) *Conductas*

En el apartado 2 sobre distintas formas de redactar los ítems están incluidas las conductas (conductas propias y valoración de conductas ajenas). Las *conductas habituales* expresan actitudes y también otras variables de interés aunque no las conceptualicemos como actitudes (como los *enfoques de aprendizaje*). En muchos tests y escalas (en tests de personalidad, escalas de actitudes hacia el estudio, motivación, etc.) es habitual formular ítems en términos de *hábitos* y *conductas habituales*.

El componente conductual de las actitudes requiere, sin embargo, alguna matización; hay que tener especial cuidado cuando en una escala de actitudes se formulan ítems en términos de lo que el sujeto *hace*. Una actitud tiende a manifestarse en conductas; por ejemplo el que valora muy positivamente (*actitud*) la conservación de la naturaleza es posible que se afilie a una organización ecologista o compre libros en esa línea (*conductas*) pero también puede hacerlo por otros motivos.

Sobre las conductas como componente de las actitudes hay que hacer dos observaciones.

1) Una conducta puede reflejar sumisión, presión social (*hago lo que hacen todos*), etc., no actitudes internas. Uno puede ir a una *manifestación* (conducta) que en principio sí manifestaría una actitud (política, social, etc.) más por presión grupal (*no quedar mal*) que por convicción.

2) Una actitud puede no reflejarse en conductas porque simplemente no se da la oportunidad (me gusta mucho la ópera, pero *no voy* a la ópera porque es muy caro, o en mi pueblo nunca hay ópera). Otras conductas sí pueden ser más claras (*suelo ver ópera en televisión*, etc.).

Realmente el *componente conductual* de las actitudes es con más propiedad un *componente conativo*[56], es decir, se refiere más a la *intención* y al *deseo* o a *lo que uno haría,* o a la *aprobación* de determinadas conductas, que a la conducta misma y en esta línea se

---

55 En los apartados 1.2.3 y 13.2.1 damos sugerencias para obtener datos de rendimiento académico cuando los cuestionarios son anónimos.

56 En latín *conor* (infinitivo *conari*) significa *intentar*.

pueden formular ítems en una escala de actitudes (*si tuviera la oportunidad no dudaría en afiliarme a una organización ecologista*).

Con todo esto no queremos decir que en una escala de actitudes no se puedan formular ítems en términos de conductas; simplemente que hay que tener *cierta cautela* al formular ítems que expresan conductas.

7. Preparar la clave de corrección

Las respuestas se codifican siempre con números íntegros sucesivos. Si por ejemplo las respuestas son cuatro, se pueden codificar de 0 a 3 o de 1 a 4. En principio es preferible evitar el 0 y comenzar a partir de 1 (cuando sólo hay dos respuestas suelen codificarse como 0 ó 1).

Si en el cuestionario que responden los sujetos las respuestas están identificadas con números, estos van siempre *en el mismo orden* cualquiera que sea el sentido del ítem (*favorable* o *desfavorable*), pero después hay que recodificarlos de manera que a la respuesta más favorable a la actitud le corresponda el número mayor[57], tal como puede verse en el ejemplo de la figura 23 (sobre *percepción de la propia competencia*)[58].

	Respuestas y clave de corrección			
ítems	Totalmente de acuerdo	De acuerdo	Más bien en desacuerdo	Totalmente en desacuerdo
Me manejo bien con las tareas de clase	4	3	2	1
Me resulta difícil hacer las tareas de case	1	2	3	4

Figura 23

En este ejemplo (figura 23) en el cuestionario que *ven y responden* los sujetos no haría falta identificar las respuestas con números porque ya están claras las respuestas verbales, pero si se hubieran identificado con números, en ambos casos irían de 4 (*totalmente de acuerdo*) a 1 (*totalmente en desacuerdo*) aunque *después* hay que invertir el orden de los números cuando *totalmente de acuerdo* es la respuesta *más desfavorable* a lo que se está midiendo.

8. Número de ítems

Por lo respecta al número de ítems en las escalas de actitudes suelen hacerse dos preguntas, cuántos se deben formular como *punto de partida* y *más o menos* cuántos ítems debe tener la escala en su *versión final*. Ninguna de las dos preguntas tiene una respuesta categórica pero sí se pueden dar algunas orientaciones y quizás sobre todo aclarar malentendidos frecuentes.

8.1. Número inicial de ítems

Sobre el *número inicial de ítems* que deben redactarse: no hay un *número óptimo*, pero a mayor número inicial de ítems tendremos una mayor probabilidad de encontrar en el análisis un conjunto de ítems definitivos con una fiabilidad suficientemente alta. Nunnally (1978:605)

---

[57] Las respuestas se pueden introducir en EXCEL por el orden en que vienen en el cuestionario (primera respuesta = 1, segunda respuesta = 2, etc.; lo que sea más cómodo pero sin tener en cuenta la clave); después si se dispone del SPSS se recodifican los números en los ítems que corresponda. Las opciones en el SPSS son *Transformar→Recodificar→En las mismas variables→ Valores antiguos y nuevos* (Morales, Urosa y Blanco, 2003:68).

[58] Estos dos ítems están tomados de Seifert y O'Keefe (2001); con un cuestionario de 15 ítems miden cinco variables relacionadas con el estudio (*atribución externa del éxito, percepción de significado,* etc.), cada variable está expresada por tres ítems (que no van juntos tal como se presentan a los sujetos); en una muestra de 512 alumnos de secundaria los coeficientes de fiabilidad de estas cinco breves escalas están entre .75 y .85 (traducido al español en Morales 2011, *cuestionarios y escalas*).

sugiere un número *máximo* de 40 ítems como punto de partida, pero pueden ser bastantes menos. En la primera versión de la escala una misma idea puede expresarse en ítems distintos con distinta formulación; el análisis de ítems posterior nos dirá cuáles podemos retener si no queremos que sean muy repetitivos.

A mayor número inicial de ítems podremos hacer una mejor selección final, pero no hay que olvidar que *además* de los ítems de la escala (de los que escogeremos los mejores en la versión final, después del análisis de ítems) en el *cuestionario completo* que responderán nuestros sujetos habrá más preguntas y con frecuencia en una misma investigación se utiliza más de una escala o test y no conviene que el instrumento sea excesivamente largo.

8.2. Número de ítems y fiabilidad

Aunque la fiabilidad suele ser mayor al aumentar el número de ítems, no conviene asociar automáticamente el número de ítems con la fiabilidad por dos razones:

1) En definitiva la fiabilidad depende de las diferencias entre los sujetos y los sujetos tienden a diferenciarse con más nitidez cuando aumenta el número de ítems y *también* cuando aumenta el número de respuestas en los ítems (tratado en el apartado 3.2 sobre *número de respuestas*).

2) De hecho se puede conseguir una fiabilidad aceptable e incluso muy alta con muy pocos ítems; tenemos múltiples ejemplos.

Como es muy común pensar que una escala de actitudes debe tener *muchos ítems* para que tenga una fiabilidad aceptable, no sobra aducir algunos ejemplos (se podrían localizar muchos más) de escalas con muy pocos ítems (entre *dos* y *seis* ítems) y que tienen una fiabilidad alta o al menos suficiente. Los ejemplos mencionados aquí son de instrumentos fácilmente localizables y potencialmente útiles en otras investigaciones.

- Seifert y O'Keefe (2001) tienen cinco escalas de *tres ítems* (relacionadas con el estudio) cada una con coeficientes que oscilan entre .75 y .85 (con una muestra de 512 sujetos).
- Meana (2003) utiliza nueve escalas para medir otros tantos valores; cada escala está compuesta por *tres ítems*; tres coeficientes no llegan a .70 (.53, .55, .64) y los otros seis oscilan entre .70 y .83 (en una muestra en torno a 650 sujetos)[59].
- Kember y Leung (2005) presentan una serie de escalas de *dos ítems* para medir la percepción de diversos aspectos de la vida académica; de los 26 coeficientes de fiabilidad 20 son superiores a .75 y los dos más bajos son de .67 y .68 (también con muestras grandes, más de 1000 sujetos).
- Gismero (1996) mide varios rasgos de *personalidad* utilizando adjetivos autodescriptivos o frases muy cortas; con *dos* adjetivos obtiene una fiabilidad de .744 y con *tres* la fiabilidad es de .606 en un caso y .812 en otro.
- Trechera (1997) mide también rasgos de *personalidad* utilizando adjetivos y obtiene coeficientes entre .53 y .62 con *tres* adjetivos y de .73 con *cuatro* adjetivos.[60].
- Burns, Vance, Szadokierski y Stockwell (2006) miden cinco necesidades básicas de los alumnos (*pertenencia, poder, libertad, supervivencia* y *diversión*) cada una con *cinco ítems* y coeficientes de fiabilidad entre .69 y .75[61]

---

59 Se trata del *Work Values Inventory* de Super, ya mencionado (ver bibliografía).

60 La muestra de Gismero es de N = 404; utiliza estos breves tests de personalidad como instrumento complementario para verificar relaciones entre personalidad y *asertividad;* la muestra de Trechera es de N = 1025 y utiliza estos tests de personalidad en relación con una medida de *narcisismo*; en ambos casos los adjetivos que miden el mismo rasgo se han localizado mediante el análisis factorial.

- Campo-Arias, Oviedo y Cogollo (2009) obtienen una fiabilidad de .87 en una escala de actitudes hacia el *cristianismo*[62] con *cinco* ítems;
- Corbiere, Fraccaroli, Mbekou, y Perron (2006) tienen una escala de *autoconcepto académico* de *seis* ítems (con cinco respuestas) referida a las asignaturas de Lengua y Matemáticas (son dos escalas) los coeficientes de fiabilidad obtenidos están entre .73 y .89[63]
- Díaz, Rodríguez-Carvajal, Blanco, Moreno-Jiménez, Gallardo, Valle, y van Dierendonck (2006) analizan seis escalas de *bienestar psicológico* de entre *cuatro* y *seis* ítems con coeficientes de fiabilidad entre .70 y .84[64].
- Lancellotti y Sunil (2009) utilizan cuatro escalas de *tres* o *cuatro* ítems referidas a la misma asignatura (*motivación, autoeficacia, actitud y comprensión de los descriptores del curso*) con coeficientes de fiabilidad entre .78 y .91[65].

8.3. Características de las escalas *muy breves*

Las escalas con muy pocos ítems tienen dos características que favorecen el que la fiabilidad sea alta: *definición simple del rasgo* y *muestras grandes.*

a) Con escalas muy breves suelen medirse actitudes o rasgos concebidos de manera *muy simple,* con ítems muy parecidos unos a otros que son indicadores muy claros del rasgo que se desea medir, de manera que el grado de acuerdo que expresen los sujetos sea el mismo o muy parecido en todos los ítems de la misma escala. Rasgos o actitudes concebidos de manera más compleja no se expresan bien con muy pocos ítems (pueden quedar *fuera* indicadores de la actitud que pueden ser importantes).

b) Como vemos en los ejemplos mencionados, con pocos ítems es más fácil obtener coeficientes altos de fiabilidad cuando *las muestras son grandes* porque en muestras grandes es más probable encontrar sujetos muy distintos en lo que estamos midiendo y consiguientemente sube la fiabilidad.

Conviene tener en cuenta estas dos características porque hacer *varias escalas* y además *muy breves* puede ser útil en investigaciones en las que el interés no está centrado fundamentalmente en estudiar un solo rasgo principal sino en un *abanico más amplio,* como podrían ser *diversas variables* relacionadas con el estudio, con la satisfacción, etc., o actitudes hacia objetos de la actitud del mismo ámbito (diversas asignaturas, tareas, profesiones, etc.), o *una serie de valores* o de rasgos de personalidad; esto es lo que sucede en casi todos los ejemplos citados. No hay que sostener *por principio* que los ítems de una escala deben ser muchos. En estos casos unos pocos ítems nos pueden dar *información fiable* sobre lo que queremos medir y no alargan demasiado el cuestionario que tienen que responder los sujetos.

Hacemos tres indicaciones más sobre estas *escalas breves:*

1. Aun en estos casos, cuando se piensa seleccionar muy pocos ítems como indicadores de un determinado rasgo, conviene comenzar redactando más ítems de los que se piensa retener

---

61 Reproducen la escala, construida con una muestra de 432 alumnos de secundaria

62 Con N = 405 adolescentes (en Cartagena, Colombia); es una adaptación muy reducida de la *Francis Scale of Attitude toward Christianity* (24 ítems) traducida a diversos idiomas; la escala original (en inglés) en Flerea, Klanjsek, Francis and Robbins (2008).

63 En dos muestras de 315 y 993 sujetos (versión en inglés, francés e italiano); escalas traducidas en la figura 12.

64 Con N = 467 sujetos entre 18 y 72 años.

65 Asignatura de *marketing;* en dos muestras de N = 219 y 118

2. Conviene que tengan *al menos* cinco o seis respuestas para asegurar que los sujetos se diferencien de manera más clara y aumente la fiabilidad.

3. Cabe formular *ítems repetitivos*, la misma idea se puede expresar de diversas maneras; frecuentemente una manera de decir las cosas resulta de hecho más discriminante que otra. Luego (después de los análisis) podemos quedarnos con la formulación que más nos convenza si no queremos que haya ítems excesivamente semejantes o que todas las breves escalas tengan un idéntico número de ítems.

9. Preparar preguntas o instrumentos adicionales

Una vez que tenemos los ítems de la escala hay que pensar en el resto de la información que debemos recoger tal como hemos visto en las *fases* del proceso y en la *estructura* del instrumento.

Necesitamos dos tipos de datos además de los que vamos a obtener con la escala:

1. Datos de identificación personal, como se hace en cualquier cuestionario, tal como está visto en el apartado 1.2 sobre *preguntas de identificación personal.*

2. Además es importante recoger *información adicional* sobre otras variables (rasgos, actitudes, valores, etc.) que pueden estar relacionadas con lo que queremos medir con nuestro instrumento.

Cuando vamos leyendo sobre el rasgo o actitud que queremos medir, o revisamos otros estudios y vamos preparando el *marco teórico*, las relaciones con *otras variables* que vemos en otras investigaciones nos pueden sugerir ideas válidas sobre *qué más podríamos preguntar* e incluso como criterio en la misma selección de los ítems y en la búsqueda de determinadas muestras.

Es útil tener desde el principio una idea clara no solamente del *rasgo* que queremos medir, sino también *con qué otros rasgos puede estar relacionado*, a qué grupos puede diferenciar, etc. Más que pensar en *un* rasgo, conviene pensar *desde el comienzo* en toda una *teoría* o *red de relaciones,* aunque sea muy modesta, en torno a ese rasgo (con qué otros rasgos o características de la persona puede estar relacionado).

Si por ejemplo vamos a construir una escala de *autoeficacia* en relación con un tipo de tarea o profesión, podemos preguntarnos *¿Con qué puede estar relacionada la autoeficacia?*

Por ejemplo:

Con experiencias de éxito en esa tarea
Con buenas calificaciones en una materia relacionada con esa tarea
Con satisfacción por la tarea
Con haber tenido buenos modelos

También podemos preguntarnos *¿Qué grupos pueden ser distintos en autoeficacia tal como la concebimos e intentamos medir?* Y podemos pensar en grupos que *han recibido o no han recibido* un entrenamiento especial para esa tarea, o en grupos que por su profesión pueden ser o no ser tan capaces, etc.[66]

Esta información se puede obtener con *simples preguntas,* o con otras escalas o instrumentos que pueden estar ya hechos y que los sujetos responderán al mismo tiempo pues todo está incluido en el mismo cuestionario.

66 Ampliado en el apartado 13.2.1

Los objetivos de estos datos adicionales se pueden desglosar en dos finalidades que pueden coincidir; facilitar la comprobación de a validez y responder a preguntas de investigación.

9.1. Comprobar la validez de la escala.

Cuando se va a construir un instrumento, como una escala de actitudes, conviene repasar los temas referidos a la *validez* y a su comprobación. Comprobar la validez de un instrumento tiene que ver con la comprobación o confirmación del *significado* de lo que medimos, y también con su *utilidad*[67].

9.2. Responder a preguntas de investigación

Independientemente de la utilidad de esta información adicional para confirmar la validez del instrumento, nuestras preguntas de investigación pueden requerir comparar distintos grupos en la actitud medida, ver con qué otras variables se relaciona, etc. Puede suceder que consideremos que la validez esté ya suficientemente establecida, sobre todo si:

1) El instrumento no es de nueva construcción y contamos con resultados de otras investigaciones (con el mismo o parecido instrumento),

2) Cuando en principio nos basta con la *validez conceptual* del instrumento (como en los cuestionarios convencionales, menos claro en las escalas de actitudes).

En cualquier caso el uso del instrumento en muestras nuevas y los análisis con otros datos aporta información sobre su validez (*validar es investigar*).

El recoger más o menos información adicional dependerá de la amplitud de nuestro estudio, pero *alguna información adicional* habrá que recoger que nos permitirá hacer otros análisis sin limitarnos a la mera construcción del instrumento. *Sólo* con nuestra escala o test no podemos ir muy lejos en nuestra investigación.

Estas *preguntas adicionales* se suelen preparar después de haber redactado los ítems de la escala pero se pueden ir pensando desde el comienzo del proceso. En el cuestionario que se presenta a los sujetos estas preguntas adicionales suelen ponerse al final del cuestionario (bloque 3º de la figura 22).

El recoger esta información adicional *al mismo tiempo* que se prueba el instrumento en una primera muestra, que es ya la *muestra definitiva,* supone un *considerable ahorro de tiempo y esfuerzo,* en vez de construir *primero* el instrumento, y *luego,* con la versión definitiva, volver a buscar otros datos en otros sujetos.

En otro apartado presentamos sugerencias y ejemplos sobre cómo hacer preguntas que recojan esta información y que conviene tener presente desde el comienzo[68].

10. Obtener datos de una muestra

Sobre la muestra con la que se construye y prueba una escala hay al menos dos temas de interés, el *tipo de muestra* y el *número de sujetos,* con una especial consideración sobre las muestras *muy pequeñas* (a veces las únicas disponibles) y las llamadas *pruebas piloto*.

10.1. Tipo de muestra

Una vez preparada la *versión inicial* del instrumento, se recogen las respuestas de una muestra para poder hacer los análisis correspondientes, que son fundamentalmente el *análisis*

---

[67] Como puede verse en el índice, sobre la validez tratamos más adelante en otros apartados, incluyendo ejemplos sobre estas *preguntas adicionales.*

[68] En el apartado 13.2, *sugerencias para obtener datos adicionales que faciliten la validación de la escala*.

de ítems y el cálculo de la *fiabilidad.* Estos análisis nos van a permitir dar forma al *instrumento definitivo.* La primera redacción de los ítems tiene un carácter en principio hipotético; suponemos que todos los ítems redactados en primer lugar describen bien un determinado rasgo o actitud, pero esta hipótesis hay que verificarla analizando las respuestas de los sujetos.

El *tipo de muestra* elegido debe ser semejante al tipo de *población* con el que se piensa utilizar después (niños, universitarios, población adulta general, etc.). A mayor *heterogeneidad* en la muestra (pero perteneciente a la *población* seleccionada) obtendremos una fiabilidad alta con mayor facilidad, pero no es legítimo *forzar* la heterogeneidad de la muestra. La heterogeneidad normal que se da en cualquier población la garantizamos mejor en muestras grandes.

10.2. Número de sujetos

Sobre el *número* de sujetos necesario para que los análisis de una escala de actitudes tengan suficiente consistencia no hay un criterio único; damos al menos las orientaciones de dos autores relevantes.

Nunnally (1978:279, 605) sugiere entre un *mínimo tolerable* de 5 sujetos por ítem y un *máximo suficiente* de 10 sujetos por ítem y Kline (1994:74, 79) estima suficiente dos o tres sujetos por ítem con tal de que la muestra total no baje de unos 100 sujetos.

Este tipo de normas sobre el número de sujetos hay que tomarlas como *orientaciones*, lo que está claro es que son preferibles muestras grandes.

Si se piensa hacer después un *análisis factorial* debe de haber unos 10 sujetos por ítem y en cualquier caso los sujetos no deben ser menos de 200. Aunque en un principio no se piense en hacer un análisis factorial, conviene dejar abierta esta posibilidad y disponer de un número adecuado de sujetos[69].

Si no se trata de construir una escala de actitudes sino de un *cuestionario convencional*, por lo general el tamaño de la muestra es el tamaño real de la muestra objeto de nuestra investigación (*nuestros alumnos, una clase, los que han participado de una experiencia*). Si se trata de una muestra a partir de la cual queremos *extrapolar* los resultados a una población determinada (*todos* los alumnos de un curso, de un colegio, de una facultad, etc.) al menos se pueden hacer estas dos observaciones:

a) La muestra debe ser *representativa* de la población a la cual queremos extrapolar los resultados.
b) A mayor número de sujetos en la muestra, habrá un menor margen de error al extrapolar los resultados a toda la población[70].

10.3. Cuando la muestra es muy pequeña

Los instrumento hechos con muestras pequeñas (como pueden ser los alumnos de una clase o simplemente los sujetos disponibles), lo mismo que los análisis de ítems, fiabilidad, etc., que hagamos, pueden también ser informativos y útiles pero en principio referidos solamente a la muestra que nos ha servido para construir el instrumento. El problema está en

---

[69] El *análisis factorial* es un análisis de la *estructura* del instrumento, identifica agrupaciones de ítems (*factores*) que miden aspectos distintos del constructo más general expresado por todos los ítems; no es un análisis *necesario* en la construcción de una escala y no lo tratamos aquí pero es interesante y conviene tenerlo en cuenta como *tema de ampliación.* Ejemplos de *análisis factoriales* de diversas escalas e instrumentos en Morales (2010, un documento sobre el *análisis factorial* en la confección de escalas); también en Morales, Urosa y Blanco (2003, pp.140ss)

[70] Ampliado en Morales (2011, *Tamaño necesario de la muestra: ¿Cuántos sujetos necesitamos?)*

utilizar después este instrumento en otras muestras; los ítems que discriminan en una muestra pueden no discriminar en otras, la fiabilidad puede variar apreciablemente, etc.

Una escala de actitudes se puede construir con una muestra pequeña (30 ó 40 sujetos, e incluso menos) no para *dejar hecha* una escala de actitudes sino con otras finalidades muy específicas en las que también conviene pensar, como pueden ser estas dos:

a) En *actividades didácticas*; por ejemplo, para *enseñar* a construir escalas o simplemente para hacer prácticas de cálculo e interpretaciones estadísticas a partir de datos de los mismos sujetos (siempre será *más eficaz* para ver la *utilidad* de estos análisis y despertar el *interés* por la asignatura y por la investigación).

b) En actividades orientadas sobre todo a la *reflexión* sobre una determinada actitud a partir de datos reales obtenidos en la misma muestra[71].

En estos casos, y aunque la finalidad primera no sea la de construir una escala de actitudes:

1) La escala construida con un pequeño grupo y con las finalidades indicadas puede convertirse en un buen *estudio piloto* y habría que presentarlo como tal, incluso en un trabajo de investigación, pero sin darle un carácter definitivo y señalando sus limitaciones (sobre todo por lo que respecta al número de sujetos); puede ser un *punto de partida* para hacer posteriormente un instrumento con más rigor metodológico y en una muestra mayor.

2) Si estas situaciones se repiten (clases, seminarios, actividades de formación que tenemos periódicamente), se pueden *acumular datos y análisis* de muestras pequeñas (y semejantes) hasta llegar a un número de sujetos apropiado y poder obtener unos resultados más definitivos y extrapolables. Si tenemos un *plan previo*, muchas de nuestras actividades se prestan a ir recogiendo datos en grupos pequeños haciendo así más rentables nuestras tareas habituales.

A veces deseamos construir un instrumento (una escala) que vamos a utilizar en una investigación con una muestra necesariamente muy pequeña (por ejemplo para evaluar una terapia, una experiencia hecha con pocos sujetos, verificar un cambio en una clase, en un grupo pequeño, etc.). En estos casos podemos construir el instrumento con una muestra grande (por ejemplo con 100 o 200 sujetos) de la *misma población*, es decir, de características similares a la que pertenecen los sujetos experimentales a los que se aplicará después ese instrumento.

El construir con una muestra grande un instrumento que luego se va a utilizar experimentalmente en una o varias muestras pequeñas tiene además la ventaja que se da *más cuerpo*, más amplitud y complejidad, a una investigación que si se limita a un grupo muy pequeño puede quedar (o *parecer*) muy limitada (por ejemplo en una tesis).

---

[71] Esta reflexión se facilita presentando datos como el número de sujetos que escogen cada respuesta en cada ítem, e incluso haciendo durante la misma actividad un análisis de ítems simplificado, como el indicado en el apartado 12.1.2 sobre *contraste de medias en cada ítem de los dos grupos con puntuaciones mayores y menores en el total de la escala*; la tabla que acompaña a este apartado es muy ilustrativa.

10.4. Las *pruebas piloto* y la validación d*e expertos*

Al construir una escala de actitudes a veces se aconseja probarla antes en una *muestra piloto* para detectar deficiencias, corregir ítems que no han funcionado bien, etc. Sobre las *pruebas piloto* hay que hacer varias observaciones.

a) Si por *prueba piloto* se entiende que un *grupo reducido* responda a la escala para *analizarla y depurarla* antes de que responda la *muestra definitiva*, como criterio general las pruebas piloto, así entendidas, *pueden ser* una pérdida de tiempo y de recursos.

Dicho así puede parecer muy radical; habrá que matizarlo y valorarlo racionalmente, pero en cualquier caso para analizar una escala hace falta una muestra *suficientemente grande* que puede ser ya la muestra definitiva. Si ya se ha puesto suficiente cuidado en la confección de la escala, analizamos los ítems en una única muestra, prescindimos de los ítems que no funcionan bien y *la misma muestra* es a la vez la muestra piloto y también la muestra definitiva con la que hacemos todos los demás análisis. Evitamos *pasar* la escala *dos veces* con un importante ahorro de tiempo y esfuerzo[72].

b) Si se trata de *tests de conocimientos* (o de otro tipo) que van a tener un *uso repetido* sí puede ser importante hacer un estudio piloto con una *muestra numerosa* (en torno a 200 sujetos) y con bastantes más ítems de los que se piensa retener (Nunnally y Bernstein, 1994:300). Si se construye un test que se va a publicar y está orientado a un uso generalizado, sí es necesaria una prueba piloto en una muestra amplia. Estas orientaciones hay que valorarlas teniendo en cuenta el uso del instrumento y la trascendencia que puede tener en los sujetos que responden.

c) Ya hemos indicado que si construimos una escala con un grupo pequeño, porque es la única muestra disponible o es suficiente para una determinada actividad, de cara a investigaciones futuras se puede considerar como una *prueba piloto* y no definitiva; posteriormente se puede mejorar y probar de nuevo con otras muestras.

d) Una cuestión distinta es que una *versión provisional* de la escala o de todo el instrumento (que incluye más preguntas, no sólo la escala) la revise o responda (o discutan en grupo) un número de sujetos *relativamente pequeño* para detectar y corregir fallos en la redacción de los ítems, verificar que todo se entiende bien, etc. Estas *pruebas* sí se pueden recomendar y podrían denominarse *prueba piloto* (explicando lo que se ha hecho) pero sin entrar en los análisis que solamente se deben hacer en muestras suficientemente grandes y que son las que permiten depurar el instrumento de manera definitiva.

La escala provisional la pueden responder (o *evaluar*) dos tipos de personas según se crea conveniente o sea posible.

1) Una pequeña muestra de sujetos *representativa* de la misma población a la que será aplicada la versión definitiva (por ejemplo un grupo de alumnos). Pueden responder al cuestionario para poder examinar después sus respuestas o pueden directamente comentarlo y hacer sugerencias (dependerá de la edad o capacidad de los sujetos).

2) Un grupo de sujetos con frecuencia denominados *expertos* (que puede ser muy pequeño) conocedores de la situación y de los sujetos que responderán después a la escala definitiva (como podrían ser otros profesores si la población es de alumnos); este grupo puede *evaluar* el instrumento y ofrecer sugerencias para mejorarlo. Por ejemplo se puede presentar

---

[72] Nos estamos refiriendo a *escalas de actitudes* y a otros tipos de tests construidos por lo general para llevar a cabo alguna investigación o tesis.

el listado de ítems para que los evalúen con uno o dos criterios (claro, *sí* o *no*, relevante, *sí* o *no*)[73].

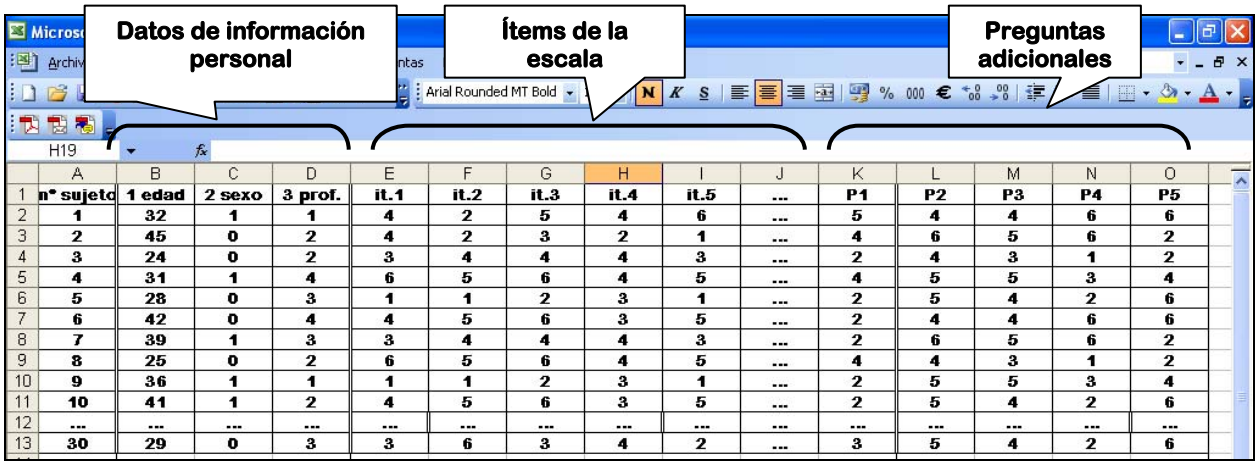
A esta revisión de los ítems por parte de un pequeño grupo se le llama a veces *validar la escala*, pero esta expresión es equívoca, se puede decir pero matizando que por *validar* se entiende esta *primera revisión*, porque una *escala* o un *test* no es válido simplemente porque lo ha revisado un grupo o unos expertos (esto sí se podría decir de un *cuestionario convencional*, no de un test o escala). La validez de los tests y escalas se confirma con estudios experimentales; aunque sí se puede hablar de una *validación previa de tipo conceptual* o *cualitativo*, como parte del proceso de validación o de asegurar que *en principio* los ítems son claros y el instrumento mide lo que se pretende. *El primer control de la validez está en la misma redacción de los ítems.*

11. Introducir los datos en un programa informático

Una vez que tenemos los datos habrá que procesarlos en un programa adecuado; normalmente utilizaremos el SPSS o EXCEL (los datos se pueden importar de un programa a otro)[74]. Suponemos que siempre está disponible al menos una hoja de cálculo como EXCEL por eso hacemos indicaciones referidas a esta hoja de cálculo. Los datos se pueden introducir en EXCEL y después completar los análisis en el SPSS. Es frecuente necesitar ayuda de alguien más informado pero conviene tener alguna idea previa[75].

11.1. Los datos en EXCEL

En la figura 24 tenemos un ejemplo orientador (ejemplo ficticio; hemos visto otro ejemplo sencillo en la figura 14).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	nº sujeto	1 edad	2 sexo	3 prof.	it.1	it.2	it.3	it.4	it.5	...	P1	P2	P3	P4	P5
2	1	32	1	1	4	2	5	4	6	...	5	4	4	6	6
3	2	45	0	2	4	2	3	2	1	...	4	6	5	6	2
4	3	24	0	2	3	4	4	4	3	...	2	4	3	1	2
5	4	31	1	4	6	5	6	4	5	...	4	5	5	3	4
6	5	28	0	3	1	1	2	3	1	...	2	5	4	2	6
7	6	42	0	4	4	5	6	3	5	...	2	4	4	6	6
8	7	39	1	3	3	4	4	4	3	...	2	6	5	6	2
9	8	25	0	2	6	5	6	4	5	...	4	4	3	1	2
10	9	36	1	1	1	1	2	3	1	...	2	5	5	3	4
11	10	41	1	2	4	5	6	3	5	...	2	5	4	2	6
12	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13	30	29	0	3	3	6	3	4	2	...	3	5	4	2	6

Figura 24

No hay una norma definida para denominar las *columnas;* se pueden identificar con la misma numeración del cuestionario, pero puede ser más cómodo poner otro tipo de identificación; este ejemplo puede servir de sugerencia[76]. Toda la información de cada sujeto está en la misma *fila*.

---

[73] Sobre la *validación de expertos* puede verse también lo dicho en el apartado 1.4 sobre *la validez de los cuestionarios*.

[74] SPSS son las siglas de *Statistical Package for the Social Sciences*. En esta *guía para construir cuestionarios y escalas* no se incluyen instrucciones sobre cómo utilizar EXCEL o el SPSS salvo alguna indicación ocasional (en Morales, Urosa y Blanco, 2003, se incluye cómo utilizar el programa SPSS en la construcción de escalas). Conviene disponer de algún manual sobre el manejo de estos (y otros) programas para aprovechar bien las posibilidades que ofrecen, o contar con alguien que pueda prestar su ayuda.

[75] Es más rápido y sencillo disponer de hojas de *lectura óptica*, pero no hay que dar por hecho que están siempre disponibles.

[76] Toda la información debe ser numérica incluso en las variables cualitativas.

En la *primera columna* tenemos el *número del sujeto;* no es necesario numerarlos pero puede ser conveniente para controlar errores (naturalmente esta columna no entra en los análisis); se controla mejor todo el proceso si todos los sujetos están numerados (lo mismo que los cuestionarios).

Los encabezados de las *columnas* deben estar bien identificados, con palabras (o abreviaturas), letras o números (se debe evitar poner más de 6 dígitos)[77]; lo importante es que todo esté claro para quien analiza los datos de manera que se pueda identificar con facilidad a qué pregunta del cuestionario corresponde cada columna.

Los datos se introducen por el orden en el que vienen en el cuestionario. En este ejemplo ficticio vemos que los datos (respuestas) de los sujetos están agrupados en los tres grandes bloques que corresponden a los indicados en la *estructura del instrumento* (figura 22).

En este ejemplo tenemos en las primeras columnas la *información personal* (sexo, edad. profesión, etc.). Esta información se puede indicar con abreviaturas o palabras completas si son cortas (como en la figura 22, *sexo, edad, curso, prof(esión), fac(ultad),* etc.).

En segundo lugar tenemos las respuestas a los ítems de la escala (it. 1, it.2, it. 3, etc.). No es necesario añadir una columna con la suma de los ítems de la escala porque se puede hacer con el mismo EXCEL. En el último bloque están las *preguntas adicionales* (P1, P2, etc.).

Si no se tiene alguna práctica o conocimientos previos, una *práctica prudente* puede ser introducir solamente los datos de dos o tres sujetos y buscar después a alguien con cierta experiencia para que verifique si se está haciendo del modo correcto; introducir en una hoja de cálculo o en un programa informático todas las respuestas de todos los sujetos es una tarea muy laboriosa y conviene corregir posibles errores cuanto antes.

11.2. Cuando algunos sujetos omiten la respuesta a algunos ítems

Un problema que se da con frecuencia es cuando hay sujetos que omiten su respuesta a algunos ítems. Si se van a hacer análisis con EXCEL no puede haber casillas en blanco; es necesario que *todos respondan a todo*; el SPSS sí admite respuestas omitidas.

Podemos proponer tres soluciones.

Primera solución. Si son pocos los sujetos que omiten algún ítem, lo más cómodo suele ser *prescindir* de estos sujetos, sobre todo si trabajamos con EXCEL (y esto vale para todos los análisis).

Si interesa que no baje el tamaño de la muestra hay varios procedimientos para sustituir estos valores que faltan y que más o menos dan resultados parecidos. Exponemos las dos soluciones más habituales a la omisión de respuestas (programadas en el SPSS)[78].

Segunda solución. El procedimiento que parece más sencillo y recomendable consiste en sustituir los valores que faltan por el valor de la *respuesta media* del sujeto (no dejando más de dos decimales); algunos utilizan la respuesta *más frecuente* para sustituir las respuestas omitidas[79].

---

[77] Seis dígitos es el máximo que admite el SPSS y se debe dejar abierta la posibilidad de utilizarlo.

[78] El cómo hacer esta substitución en el SPSS está explicado en Morales, Urosa y Blanco (2003:69-72).

[79] Qué hacer cuando algunos sujetos no responden a algunos ítems puede verse tratado e investigado (precisamente en escalas de actitudes tipo Likert) en Dodeen (2003) que recomienda poner la *respuesta media* del sujeto (su total dividido por el número de ítems que ha respondido) en lugar de las respuestas omitidas.

Tercera solución. Otra solución propuesta es utilizar como puntuación total de *todos los sujetos* no la *suma* de sus respuestas a todos los ítems (lo habitual), sino la *media*, dividiendo la suma de las respuestas de cada sujeto por el número de ítems que ha respondido. Es decir, no se utiliza la media de los ítems respondidos para sustituir las omisiones, sino que esta media calculada para cada sujeto es el total individual (en vez de la suma) que se utiliza después en el resto de los análisis (para calcular medias, desviaciones, análisis de ítems, correlaciones, etc.) (Bortz y Döring 2006, p.224; Wuensch, 2006).

Por ejemplo: si un sujeto en una escala de 4 ítems responde solamente a tres ítems (por ejemplo responde 3, 3 y 4 a tres ítems y omite la respuesta a un cuarto ítem) la media de los tres ítems respondidos sería 3+3+4/3 = 3.33. Esta media de los ítems respondidos (3.33) se puede poner como respuesta al ítem omitido (solución segunda) o se puede utilizar esta media en vez de la suma como total de ese sujeto (y así con todos los sujetos, solución tercera).

Como criterio general parece preferible la segunda solución (poner en los ítems omitidos la media individual de los ítems respondidos para sustituir las omisiones) porque se mantiene la práctica más habitual, que es sumar a cada sujeto todas sus respuestas y así se facilita la comparación con las medias de otros grupos que son calculadas habitualmente a partir de los totales de todos los sujetos.

Hay que advertir que cuando se utiliza esta solución (*media de los ítems respondidos* para sustituir a los omitidos), la fiabilidad de todo el instrumento (y las correlaciones entre los ítems) tiende a aumentar artificialmente. Los dos procedimientos (*segunda y tercera solución*) dan resultados muy semejantes cuando tanto el número de los sujetos que omiten ítems como el número de ítems omitidos es del orden del 20% o inferior (Downey y King, 1998)[80].

Si hemos optado por utilizar como total de cada sujeto su *media* a los ítems respondidos pero queremos utilizar como dato en los análisis la *suma* de todas las respuestas como si todos hubieran respondido a todos los ítems, nos basta multiplicar para cada sujeto la media de los ítems respondidos por el número de ítems (redondeando decimales).

En cualquier caso aunque no haya ítems sin responder, el calcular en un grupo *la media por ítem* (media total dividida por el número de ítems) puede ser útil para hacer *gráficos* ilustrativos que permiten comparar grupos intuitivamente o un mismo grupo en varias variables independientemente del número de ítems de cada variable o instrumento.

Los ítems no respondidos por algunos sujetos pueden ser más problemáticos cuando no podemos suponer que la omisión de respuestas es *aleatoria*, por ejemplo cuando bastantes sujetos de un determinado tipo no responden a determinadas preguntas. En este sentido las respuestas omitidas pueden ser mayor problema en las preguntas sobre *características personales* (que también suelen incluirse en tests y escalas aunque no como ítems de la escala) como podría suceder cuando no se responde a preguntas de *identificación étnica* o *pertenencia a determinados grupos, nivel de ingresos económicos* o *número de artículos publicados en una muestra de profesores universitarios*, etc. Difícilmente se puede suponer que el omitir la respuesta a este tipo de preguntas sea algo *aleatorio* (simple olvido o distracción). Es muy posible que no se responda *porque no se quiere responder*. Siempre cabe explorar si los sujetos que omiten la respuesta a determinados ítems tienen alguna característica común.

Estas omisiones (es decir, menos datos en algunas de estas variables) pueden afectar a:

---

[80] Este estudio (Downey y King, 1998) está hecho con una muestra de 975 adultos, utilizando dos escalas de 15 y 20 ítems.

a) La descripción de la muestra
b) Las correlaciones de la escala con estas variables
c) El contraste de medias entre grupos formados en función de las respuestas a alguna de estas preguntas.

12. Proceso de análisis de una escala de actitudes: finalidad del análisis de ítems e interpretación del coeficiente de fiabilidad.

Una vez que tenemos todos los datos procedemos a los análisis. En la figura 25 tenemos una visión de conjunto del proceso que vamos a seguir para construir una escala de actitudes o un test.

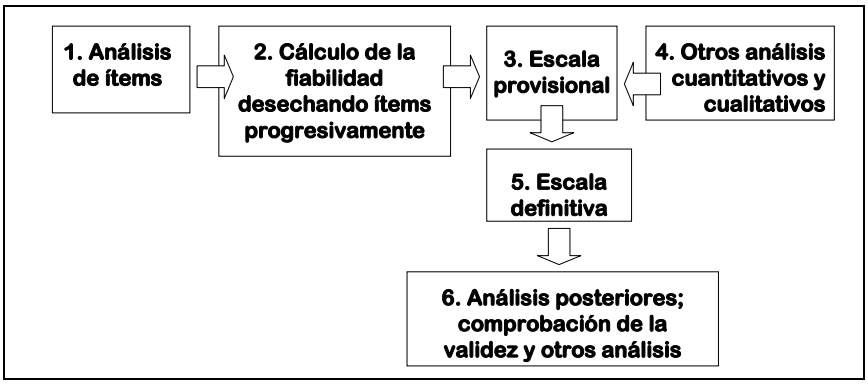


Figura 25

Este proceso es aproximado; la escala indicada como provisional (paso 3) puede ser ya la definitiva aunque antes de hacer la selección definitiva de los ítems caben otros análisis complementarios (de los que podríamos prescindir); unos análisis pueden ser cuantitativos (como el *análisis factorial*)[81] y otros análisis pueden tener un carácter más cualitativo (comentados en otro apartado)[82].

Antes de explicar con más detalle y en apartados diferenciados el *análisis de ítems* (por qué y cómo hacerlo) y la *fiabilidad*, conviene tener una idea inicial clara de lo que estamos haciendo. El análisis de ítems se hace *para seleccionar* los ítems que nos van a dar una *fiabilidad óptima*.

Con el *análisis de ítems* respondemos a esta pregunta hecha para cada ítem:	➔	*¿En qué grado este ítem mide o manifiesta 'lo mismo' que miden todos los demás ítems y que es precisamente lo que queremos medir con toda la escala?*

Hablando con propiedad más que verificar si el ítem *mide* lo mismo que los demás, lo que verificamos *de hecho* es en qué grado se relaciona cada ítem con la suma de todos los demás ítems, y de esta relación *deduciremos* que mide lo mismo que el resto de los ítems[83]. El coeficiente de fiabilidad nos dirá *en qué grado* lo hemos logrado con el conjunto de ítems finalmente seleccionados; este coeficiente de alguna manera *nos resume* lo conseguido con el análisis de ítems.

---

81 El *análisis factorial* (del que no estamos tratando) cabe incorporarlo en el proceso de construcción de la escala o se puede hacer ya con la escala definitiva.

82 Apartado 12.3.2, *otros criterios en torno a la elección de los ítems definitivos*.

83 Con los análisis estadísticos analizamos *números*, no *conceptos*; verificamos *relaciones numéricas* para confirmar *relaciones conceptuales.*

Antes de exponer cómo llevar a cabo el análisis de ítems es importante entender el *significado e interpretación del coeficiente de fiabilidad* que es en definitiva el punto final del análisis de ítems.

Sin entrar en profundidad en el tema de la fiabilidad sí que hay que tener claro cómo se interpretan estos coeficientes, también denominados coeficientes de *consistencia interna*[84]. Estos coeficientes van de 0 a 1; el coeficiente utilizado normalmente es el coeficiente *alpha* (α) de Cronbach. El objetivo del análisis de ítems es en principio seleccionar los ítems que mejor contribuyen a la fiabilidad de toda la escala.

Qué quiere decir un coeficiente de fiabilidad alto.

Al menos podemos indicar tres interpretaciones relacionadas entre sí.

1) Una fiabilidad alta quiere decir que el test o escala *diferencia bien a los sujetos en el rasgo medido* (en lo que tienen *en común* los ítems).

El coeficiente de fiabilidad viene a indicar la *capacidad diferenciadora* o discriminatoria de la escala o test, por eso se encuentra una mayor fiabilidad en muestras más heterogéneas, y también en muestras grandes en las que hay una mayor probabilidad de que haya sujetos más distintos en lo que estamos midiendo.

No se puede ordenar o diferenciar bien a los muy semejantes. Sin diferencias entre los sujetos no hay una fiabilidad alta. También, y por la misma razón, el mismo instrumento aplicado a la misma muestra *después* de un proceso puede tener una fiabilidad menor porque la muestra se ha hecho más homogénea en función de ese proceso.

2) Una fiabilidad alta *apoya* la interpretación de que todos los ítems miden o expresan el mismo rasgo.

Decimos que un coeficiente de fiabilidad alto *apoya pero no prueba que todos los ítems miden lo mismo* o *expresen bien el mismo rasgo* porque una fiabilidad alta lo que expresa literalmente es que los ítems están *relacionados entre sí;* puntuaciones altas o bajas en cada ítem se corresponden a puntuaciones altas o bajas en todos los demás ítems; para poder afirmar que todos los ítems miden *lo mismo* hacen falta también *controles conceptuales.*

La necesidad de *controles conceptuales* podemos verla con un ejemplo hipotético, absurdo pero claro. Si a un grupo formado por niñas de 10 años que están recibiendo clases de *ballet* y de niños de 14 años que son miembros de un equipo de fútbol, y les preguntamos la edad, sexo, peso, altura y si practican el fútbol y el *ballet*, y sumamos sus respuestas a todas estas preguntas como si se tratara de un test, tendremos una fiabilidad muy alta (todos los ítems diferencian a los mismos sujetos; todos los niños responden de una manera y todas las niñas de otra) y *no estamos midiendo nada interpretable* a pesar de esa fiabilidad tan alta. En ningún caso la estadística substituye al sentido común y al análisis lógico de lo que estamos haciendo.

3) El coeficiente de fiabilidad se puede interpretar como la *correlación estimada* con otra escala semejante (de ítems parecidos)

Los sujetos hubieran quedado *ordenados de manera parecida* en tests o escalas semejantes en tipo y número de ítems.

---

84 Un tratamiento más amplio de la fiabilidad de tests y escalas en Morales (2008, capítulo 6) y con más extensión en Morales (2006, capítulos 9 y 10). También es útil repasar todo lo referido a los *coeficientes de correlación*, que son centrales en todos estos análisis (Morales, 2006, cap. 5, o cualquier otro texto de estadística).

Dos observaciones importantes sobre la fiabilidad.

1) La fiabilidad en sentido propio *no es una propiedad del instrumento* sino de los datos recogidos en una muestra. La magnitud del coeficiente de fiabilidad puede variar de muestra a muestra (será mayor en muestras con mayores diferencias interindividuales) y por lo tanto *debe calcularse en cada nueva muestra* (norma de la APA)[85].

2) Lo que no podemos concluir de una fiabilidad alta es que prueba o supone la *validez* del instrumento, es decir, que mide realmente lo que pretendemos o decimos que medimos; la validez requiere un tratamiento diferenciado.

12.1. Análisis de ítems

Los ítems, tal como los hemos redactado, constituyen una *definición operativa*, provisional e hipotética, del rasgo que vamos a medir. Ahora tenemos que comprobar si cada ítem, supuestos los controles lógicos iniciales (los hemos redactado para que midan el mismo rasgo) *mide lo mismo que los demás*, y por lo tanto es *sumable* en una puntuación total que es la que después interpretamos y utilizamos. Esta comprobación la hacemos en cada ítem mediante el análisis denominado *análisis de ítems.*

Lo que queremos comprobar es si las respuestas tienden a *covariar*, es decir, si los sujetos tienden a responder de *manera coherente*, de manera que podamos deducir que todos los ítems *son indicadores del mismo rasgo*. En definitiva comprobamos si los ítems tienden a *diferenciar* a los sujetos, si *discriminan* adecuadamente.

Los procedimientos que podemos utilizar para analizar los ítems son dos, la *correlación ítem-total* (el más habitual) y el *contraste de medias de los grupos extremos*. Con ambos métodos llegaremos a resultados muy parecidos.

12.1.1. Correlación ítem-total

Propiamente se trata de la *correlación de cada ítem con la suma de todos los demás* (o *correlación de cada ítem con el total menos el ítem*) que suele denominarse *correlación ítem-total corregida* (corregida porque en este total no se incluye el ítem que estamos analizando).

Lo que comprobamos es en qué medida el puntuar alto en un ítem supone de hecho obtener un total alto en todos los demás ítems; y viceversa, si puntuar bajo en un ítem se corresponde con un total menor en la suma de todos los ítems.

Los ítems con una mayor correlación con el total son los que en principio *tienen más en común* con los demás y por lo tanto *podemos suponer* que *miden lo mismo que los demás* con más claridad. Los ítems con correlaciones más bajas con el total los eliminaremos de nuestra escala porque no miden claramente lo mismo que los demás (el puntuar alto en esos ítems no supone puntuar alto en los demás).

Calcular la correlación de cada ítem con el total (y no con el total *menos el ítem analizado*), es sencillo con una hoja de cálculo como EXCEL, en ese caso hay una fórmula que convierte esta correlación ítem-total en la correlación ítem-total *menos el ítem,* que es la que nos interesa, pero este procedimiento es laborioso[86].

---

85 *American Psychological Association;* sus normas son una referencia autorizada. El aportar los coeficientes de fiabilidad obtenidos en otras muestras es informativo pero no es suficiente.

86 Esta fórmula la tenemos en Morales, 2008, cap. 5, apartado 4.3. Lo que es sencillo en EXCEL es calcular la *correlación ítem-total sin restar al total el ítem analizado,* por lo que estas correlaciones serán algo mayores, sobre todo si los ítems son pocos. Si en la matriz de datos (*filas* sujetos, y *columnas* ítems) añadimos una última columna con la suma de todos los ítems; en la matriz de correlaciones tendremos en la última fila la correlación de cada ítem con el total. Aun así con

Suponemos que habitualmente utilizamos el programa SPSS, que nos da para cada ítem la *correlación ítem-total* (menos el ítem) y la *fiabilidad de todo el test o escala si suprimimos ese ítem* y se puede apreciar rápidamente qué ítems se pueden eliminar para que suba la fiabilidad. Esta información es útil y acelera el proceso, pero conviene tener presentes otras consideraciones que veremos en otro apartado (nº 12.3.2).

En la tabla 1 tenemos un ejemplo de la información que nos da el SPSS.[87]. En este caso se trata del análisis de una breve escala de *autoeficacia materna* (Zurdo-Garay, 2011).

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento
VAR00001	**26,5488**	**11,362**	**,471**	**,623**
VAR00002	**26,5610**	**11,583**	**,380**	**,644**
VAR00003	**26,5122**	**11,241**	**,425**	**,633**
VAR00004	**26,7073**	**11,815**	**,348**	**,651**
VAR00005	**26,0244**	**12,715**	**,359**	**,651**
VAR00006	**26,0244**	**13,777**	**,132**	**,686**
VAR00007	**26,1341**	**13,105**	**,209**	**,677**
VAR00008	**26,4390**	**11,607**	**,445**	**,630**
VAR00009	**26,1220**	**12,133**	**,368**	**,647**

Tabla 1

La información que nos da el SPSS y que vemos en esta tabla 1 es la siguiente:

- La *fiabilidad* de toda la escala con todos los ítems iniciales; en este caso la fiabilidad de la escala formada por estos 9 ítems es (con 82 sujetos) de α = .677 (esta información no aparece en la tabla).
- La media y la varianza de *toda* la escala si suprimimos cada ítem (dos primeras columnas).
- La correlación de cada ítem con la suma de todos los demás (eso significa *correlación elemento-total corregida*).
- La fiabilidad de la escala si suprimimos el ítem.

Además el programa nos da la información descriptiva de cada ítem (si la pedimos en el cuadro de diálogo; media y desviación). Realmente la información de interés es la que tenemos en las *dos últimas columnas.*

En la presentación de un trabajo de investigación (o de una tesis o de un artículo) queda toda la información más clara y se interpreta con mayor facilidad si en la tabla 1 (*output* del SPSS):

1) Copiamos la formulación de los ítems (se puede poner abreviada).
2) Nos limitamos a poner la información de las dos últimas columnas que son las que realmente nos interesan (si lo deseamos, la información completa puede ir en un anexo).

---

este procedimiento (sugerido por Trochim, 2006) todavía nos faltaría calcular el coeficiente de fiabilidad. En conjunto y para construir escalas de actitudes y tests en general, es preferible utilizar el programa SPSS.

87 Una explicación detallada, paso a paso, sobre cómo utilizar el SPSS en la construcción de escalas puede verse en Morales, Urosa y Blanco (2003). Las opciones en el menú del SPSS para el análisis de ítems y el cálculo de la fiabilidad son *Analizar→Escalas→Análisis de Fiabilidad*

La tabla 1 quedaría en este caso como la tabla 2 en la que hemos copiado la formulación de los ítems y hemos suprimido las dos primeras columnas de datos de la tabla 1.

Ítems de la escala de *autoeficacia materna* Respuestas: (4) *Sí, siempre o casi siempre,* (3) *Muchas veces*, (2) *Pocas veces*, (1) *No, nunca o casi nunca*	**Correlación elemento-total corregida**	**Alfa de Cronbach si se elimina el elemento**
1. Soy capaz de saber lo que le sucede cuando él/ella está molesto.	**,471**	**,623**
2. En general me siento muy capaz como mamá	**,380**	**,644**
3. ¿Se siente capacitada para ayudarle en sus tareas escolares / en su aprendizaje?	**,425**	**,633**
4. ¿Se siente capaz de hacer que él/ella obedezca?	**,348**	**,651**
5. ¿Se siente en la capacidad de brindarle los medios necesarios para que tenga una buena educación?	**,359**	**,651**
6. ¿Se siente capaz de atenderle por si misma en lo que necesita cada día (ropa, alimentos, etc.)?	**,132**	**,686**
7. Si se enferma, ¿Se siente en la capacidad de decidir qué hacer o qué cuidados necesita?	**,209**	**,677**
8. ¿Se siente capaz de hablar o comunicarse con él/ella cuando tiene problemas o dificultades?	**,445**	**,630**
9. ¿Se ve a sí misma capaz de satisfacer las necesidades básicas que el niño/a tiene (alimentación, vestido, salud)?	**,368**	**,647**

Tabla 2

Podemos observar que:

Si suprimimos el ítem 6 (es el que tiene una menor correlación con el total) la fiabilidad sube de .677 (la obtenida con todos los ítems) a .686

Si suprimimos el ítem 7, la fiabilidad no cambia, sigue en .677

En principio podemos suprimir estos dos ítems que son también los que menos relación tienen con la suma de los demás; la fiabilidad sube de hecho a .689 (podemos redondear a .69) si dejamos los 7 mejores ítems.

Las *correlaciones bajas con el total* de algunos ítems que en principio nos parece que *expresan conceptualmente bien* lo que queremos medir, pueden tener algunas de estas explicaciones que puede ser útil examinar y explicar:

1) Las medias son más bien altas o más bien bajas y las desviaciones típicas son pequeñas (al menos en términos relativos): la mayoría tiende a estar de acuerdo o en desacuerdo y por lo tanto no diferencian a los sujetos.

2) En estos ítems con correlaciones bajas con el total también puede suceder que haya diferencias claras entre los sujetos (conviene fijarse en la *magnitud relativa* de la desviación típica), pero lo que sucede es que miden *algo distinto* a lo expresado en los demás ítems, *al menos tal como lo entienden los sujetos;* se puede puntuar *alto* en estos ítems (*siempre, mucho, de acuerdo,* etc.) y a la vez *o alto o bajo* en los demás ítems (eso significa *no relación* o relación baja). En el ejemplo de la tabla 2 unos sujetos podrían responder *siempre* al ítem 6 y *nunca* o *pocas veces* a casi todos los demás y otros sujetos podrían responder también *siempre* al mismo ítem 6 y *siempre* o *muchas veces* a otros ítems.

En ocasiones podemos advertir que un ítem tiene una correlación *negativa* con el total; en estos casos podemos comprobar si no hay un error en la clave de corrección; si no hay error ciertamente hay que eliminarlo.

A la vista de la tabla con las correlaciones ítem-total podemos hacernos estas dos preguntas que pueden prestarse a buenos *comentarios cualitativos* sobre los ítems con mayores y menores correlaciones con el total:

→*¿Qué ideas expresan o qué tienen en común los ítems que discriminan más, que diferencian mejor a los que tienen totales más altos o más bajos*? (ahí está el *núcleo* de lo que estamos midiendo, son los ítems que mejor diferencian a los sujetos con totales más altos de los sujetos con totales más bajos*).*

→*¿Qué ideas expresan o qué tienen en común los ítems que discriminan menos?*

Un programa como el SPSS facilita notablemente el proceso, pero conviene hacer algunas observaciones que son también aplicables si utilizamos el contraste de los grupos extremos que veremos a continuación.

a) Es cuestionable (al menos no es lo mejor necesariamente) seguir *cuasi mecánicamente* procedimientos automáticos; el constructor del instrumento puede intervenir con *sus propios criterios* sobre lo que quiere medir y sobre las características del instrumento (por ejemplo, *incluir un número idéntico de ítems positivos y negativos*[88]). Ampliamos este punto en el apartado 12.3.2, *otros criterios en torno a la elección de los ítems definitivos*. Es el autor del instrumento quien hace y da forma a su instrumento, no un programa de ordenador.

b) Por otra parte estos programas informáticos nos dan la fiabilidad si suprimimos los ítems *de uno en uno,* pero no si suprimimos más de uno a la vez, y puede no interesar ir eliminando ítems *uno a uno* sino en bloques escogidos con algún criterio (como el tener una casi idéntica o muy parecida correlación ítem-total, o buscando que haya un número idéntico de ítems positivos y negativos).

c) Además es muy normal que con varios subconjuntos de ítems obtengamos una fiabilidad idéntica o similar, por lo que habrá que acudir a otros criterios (al menos se puede pensar en esta posibilidad) en la selección definitiva de los ítems (como consideraciones conceptuales y resultados del *análisis factorial*).

12.1.2. Contraste de medias en cada ítem entre los dos grupos con puntuaciones mayores y menores en el total de la escala.

El procedimiento anterior es el que se hace habitualmente con el SPSS. Si no disponemos del programa adecuado hay un procedimiento que aporta una información semejante. Aunque suponemos que los análisis los haremos habitualmente con el SPSS, no sobra indicar cómo llevar a cabo el análisis de ítems con el otro procedimiento porque además ayuda a comprender mejor lo que estamos haciendo (se entiende con más facilidad lo que es una diferencia entre dos medias que un coeficiente de correlación). Consiste en comparar en cada ítem el 25% con puntuación total *más alta* con el 25% con puntuación total *más baja*[89]. Para llevar a cabo este análisis:

---

88 Ya hemos indicado (en el apartado 6.4 *ítems positivos y negativos*) que es útil calcular la correlación entre los dos subtotales (sumando por separado las respuestas a los ítems positivos y negativos); si la correlación está en torno a .50 podemos excluir el influjo de la *aquiescencia* en las respuestas.

89 Este análisis se puede hacer fácilmente con una hoja de cálculo tipo EXCEL; en Morales, Urosa y Blanco (2003) también se explica cómo hacer este contraste de medias con el SPSS; si se dispone del SPSS es preferible el método anterior (*correlación ítem-total*).

1º Ordenamos a los sujetos de más a menos, según el total obtenido en toda la escala, y seleccionamos dos subgrupos:

El *grupo superior*, el 25% con puntuación total más alta
El *grupo inferior*, el 25% con puntuación total más baja;

El 50% central no entra en este análisis (en la correlación ítem-total sí entran todos los sujetos).

2º Calculamos la media y la desviación típica en cada ítem de cada uno de los dos grupos, superior e inferior;

3º Contrastamos las medias de estos dos grupos mediante la *t de Student*.

Lo que esperamos es que el 25% con una puntuación total superior tenga una media significativamente más alta en cada ítem que el 25% inferior. Podremos en este caso concluir que los ítems que simultáneamente diferencian a los mismos sujetos están *midiendo lo mismo*. En principio prescindiremos de los ítems que no discriminan (valores de la t de Student no significativos), y si son muchos o demasiados los que discriminan (y esto sucede con frecuencia), podemos quedarnos con los más discriminantes; siempre hay ítems mejores que otros en términos relativos. Ya hemos indicado (a propósito del análisis anterior, *correlación ítem-total*) que en la elección definitiva de los ítems pueden intervenir además otros criterios. En cualquier caso retendremos solamente ítems que discriminan, que diferencian bien a los sujetos (es lo que pretendemos con un instrumento de medida), y son estos los ítems de los que podemos suponer que miden básicamente el mismo rasgo.

Para hacer estos análisis conviene disponer los datos de manera clara y tener a la vista algún modelo como el de la tabla 3 (no hay un modo único de presentar los datos). En este ejemplo de un total de 40 sujetos (número muy bajo para construir una escala de actitudes si se tratara de un caso real) comparamos las respuestas de los 10 sujetos (25%) con totales más altos con las respuestas de los 10 sujetos (25%) con totales más bajos.

Análisis de ítems: contraste de medias en cada ítem entre el 25 % con total más alto y el 25 % con total más bajo

ítem		5	4	3	2	1	M	$M_s - M_i$
Nº 1	25% Sup.	II (2)	IIIII (5)	III (3)			3.9	2.50
	25 % Inf.			I (1)	II (2)	IIIIIII (7)	1.4	
Nº 2	25% Sup.	IIII (4)	IIIIII (6)				4.4	.30
	25 % Inf.	III (3)	IIIII (5)	II (2)			4.1	
Nº 3	25% Sup.		II (2)	IIIII (5)	II (2)	I (1)	2.8	-1.1
	25 % Inf.	IIII (4)	II (2)	III (3)	I (1)		3.9	

Tabla 3

En este ejemplo ficticio (tabla 3) tenemos:

1) La distribución de frecuencias de los dos grupos en cada ítem,
2) Las medias (M) de cada grupo
3) La diferencia entre las medias (M_s-M_i).

También es útil calcular la media y la desviación típica de cada ítem en *toda* la muestra; las desviaciones típicas de los ítems en *toda* la muestra nos harán falta después para calcular la fiabilidad (si no utilizamos el SPSS).

En un análisis más completo, en la tabla 3 habría que incluir en cada ítem *las desviaciones típicas* y el valor de la *t de Student* al contrastar las medias; la información de la tabla 3 es suficiente como ejemplo ilustrativo[90].

En esta tabla 3 podemos observar lo siguiente:

→El ítem nº 1 discrimina muy bien, los sujetos con total más alto reparten sus respuestas entre 5 y 3 y los sujetos de total más bajo entre 3 y 1.

→El ítem nº 2 discrimina poco, casi todos responden de la misma manera, la diferencia es muy pequeña y habrá que eliminarlo en la escala definitiva.

→El ítem nº 3 discrimina *negativamente,* los del grupo inferior superan a los del grupo superior; claramente este ítem *no es sumable con los demás*, no mide lo mismo y hay que rechazarlo. En este caso también puede suceder que esté mal la clave de corrección y conviene revisarla.

Con cualquiera de los dos procedimientos (correlación ítem-total y contraste de medias entre los dos grupos extremos) obtenemos un dato (r ó t) sobre la *calidad* del ítem; ambos tipos de información nos dicen si podemos considerar que el ítem discrimina adecuadamente y consecuentemente si podemos considerar que mide lo mismo que los demás.

¿Cuál de los dos análisis es preferible?

Los dos análisis, *correlación ítem-total* y *contraste de medias entre grupos extremos*, aportan información semejante; prácticamente con los dos procedimientos se llega a la misma selección de ítems, sobre todo si nos vamos a quedar con los mejores (con los más discriminantes). En la práctica el escoger un método u otro es cuestión de conveniencia y lo habitual será utilizar el SPSS (con la correlación ítem-total y el coeficiente de fiabilidad ya programados); es el procedimiento más cómodo y el que hoy día suele seguirse rutinariamente. También se puede hacer con EXCEL pero al no estar programado directamente resulta más laborioso.

En *procesos de aprendizaje* sobre cómo construir escalas y utilizando un ejemplo real en el que los mismos participantes han generado los ítems, quizás el contraste de medias entre los grupos extremos es intuitivamente más claro: se entiende fácilmente que los ítems que simultáneamente están diferenciando a los mismos sujetos están expresando el mismo rasgo.[91]

12.2. Cálculo de la fiabilidad

El coeficiente de fiabilidad utilizado habitualmente es el coeficiente α de Cronbach. La fórmula es la siguiente:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_t^2}\right)$$

En esta fórmula k es el número de ítems, σ_i es la desviación típica de cada ítem (hay que sumar las varianzas o desviaciones típicas de los ítems elevadas previamente al cuadrado) y σ_t es la desviación típica de los totales.

---

90 En otro apartado (10.3, *cuando la muestra es muy pequeña*) hemos indicado que este sencillo análisis (cómo se reparten las respuestas de los grupos con totales más altos y más bajos en cada ítem) puede ser un buen estímulo para reflexionar sobre el contenido de los ítems y de la actitud que supuestamente reflejan.

91 El autor del procedimiento (Likert) recomienda y utiliza el contaste de medias, pero en su época no se disponía de los programas de ordenador con los que contamos hoy día.

La fórmula es muy laboriosa (aunque el cálculo puede quedar facilitado utilizando EXCEL) y más todavía si tenemos que calcular la fiabilidad con distintas combinaciones de ítems, pero estamos suponiendo que para construir escalas utilizamos el programa SPSS que nos da directamente el valor de este coeficiente.

Si no disponemos del SPSS una alternativa sencilla al cálculo del coeficiente α (que en principio es el preferible) es utilizar alguna de las fórmulas basadas en la partición del test o escala en *dos mitades.* Para calcular estos coeficientes haremos lo siguiente:

1º Al *corregir* la escala a cada sujeto se le calculan dos puntuaciones totales, una en los ítems pares y otra en los ítems impares (la suma de los dos *subtotales* será el total de cada sujeto).

2º Después se calcula la *correlación entre las dos mitades* pues esta correlación entra en estas fórmulas de la fiabilidad basadas en la partición del test en dos mitades[92].

Conviene repasar en otro lugar todo lo referente a estas fórmulas; en principio es preferible utilizar las fórmulas del coeficiente α de Cronbach o Kuder-Richardson 20[93].

12.2.1. Cómo *estimar* la fiabilidad en una nueva muestra a partir de la fiabilidad conocida en otra muestra y de las desviaciones de las dos muestras.

Entre las muchas fórmulas en torno a la fiabilidad hay una que no es especialmente laboriosa y que puede ser ocasionalmente muy útil cuando utilizamos un test o una escala *ya hecha (no necesariamente de confección propia) y utilizada en otros estudios.* Cuando utilizamos una escala ajena y ya probada en otras muestras (normalmente localizada en alguna tesis o estudio publicado) solemos encontrar el coeficiente de fiabilidad calculado en *esa otra muestra*, además de otros datos descriptivos, como la media y la desviación típica de los totales de la escala.

Como la magnitud de la fiabilidad depende de la heterogeneidad de la muestra (aunque no se trata de una relación *sistemática*) conociendo 1º la *fiabilidad* y la *desviación típica* encontradas *en otra muestra* y 2º la *desviación típica* encontrada en *nuestra muestra* podemos *estimar la fiabilidad aproximada* en *nuestra muestra* mediante esta fórmula (Guilford y Fruchter, 1973:420; Morales, 2008:231):

$$r_{nn} = 1 - \frac{\sigma_o^2 (1 - r_{oo})}{\sigma_n^2}$$

r_{nn} = *fiabilidad estimada* en la *nueva muestra*
σ_o y r_{oo} = desviación típica y fiabilidad ya calculadas (observadas) en otra muestra,
σ_n = desviación típica en la *nueva muestra* (en la que deseamos *estimar* la fiabilidad).

Por ejemplo, si en una escala de actitudes hemos obtenido en una muestra una desviación típica de 6.86 y una fiabilidad de α = .78 (o hemos visto publicada en otro lugar esta información) ¿qué fiabilidad podemos esperar en otra muestra cuya desviación típica vemos que es 7.28?

Aplicando la fórmula anterior de la *fiabilidad estimada* en nuestra muestra, tendríamos:

---

92 Hay varias fórmulas de la fiabilidad basadas en la partición del test en dos mitades (pueden verse en Morales, 2008, cap. 6) y conviene revisarlas antes de escoger una, pero en cualquier caso siempre es preferible calcular el coeficiente α de Cronbach ya programado en el SPSS.

93 Las dos fórmulas son idénticas aunque varíen los símbolos; la de Cronbach se utiliza con ítems en los que hay varias respuestas graduadas, y la de Kuder-Richardson cuando las respuestas son 1 ó 0; el programa del SPSS para analizar los ítems y calcular la fiabilidad es el mismo.

$$\alpha = 1 - \frac{6.68^2 (1 - .78)}{7.28^2} = .8147$$

Bien entendido que hay que presentar este coeficiente como una *estimación* de la fiabilidad en la nueva muestra[94].

12.2.2. Cuándo un coeficiente de fiabilidad es suficientemente alto

Aunque no hay un valor óptimo de referencia se pueden dar algunas orientaciones:

1) Un valor en torno a .70 se considera aceptable; es un valor muy habitual (Schmitt, 1996) y es también el valor mínimo recomendado por Nunnally (1978:245-346). Valores en torno a .60 son también aceptables.

2) Con valores muy inferiores (hasta .50) podemos utilizar el instrumento en trabajos de investigación (Schmitt, 1996; Guilford, 1954:388-389)[95].

En definitiva no hay un *valor mínimo sagrado* para aceptar un coeficiente de fiabilidad como adecuado y medidas con una fiabilidad relativamente baja pueden ser muy útiles en trabajos de investigación.

3) Para dar *información fiable a cada sujeto* estos coeficientes de fiabilidad deben ser bastante más altos (.80 o mayores) porque al subir la fiabilidad, baja el error típico o margen de oscilación entre ocasiones o medidas semejantes. Un coeficiente de fiabilidad relativamente bajo puede ser problemático para hacer diagnósticos individuales, pero normalmente las escalas de actitudes no se utilizan con esta finalidad.

Cuando la fiabilidad obtenida nos parece baja (y siempre que lo estimemos oportuno) podemos obtener sin mucho esfuerzo otro tipo de información complementaria derivada del coeficiente de fiabilidad.

a) Las *correlaciones entre variables* se ven afectadas por la baja fiabilidad de los instrumentos pero buscando la fórmula apropiada podemos *estimar* cuál sería el valor de la correlación si la fiabilidad de nuestro instrumento fuera óptima[96].

b) Cuando baja la fiabilidad, sube el *error típico o margen de oscilación* probable de las puntuaciones individuales. Este error típico lo podemos calcular a partir del coeficiente de fiabilidad y de la desviación típica. De cada sujeto la información más justa y razonable (sobre todo para tomar decisiones) no es la puntuación obtenida de hecho, sino los límites probables máximo y mínimo (*intervalos de confianza*) entre los que podemos estimar que se encuentra su verdadera puntuación[97]. En situaciones de diagnóstico y orientación individual una baja fiabilidad podemos obviarla calculando esos márgenes de error y tenerlos en cuenta; la información es más imprecisa pero también *más segura*.

---

94 Y en este caso indicando la fórmula y la fuente de donde se tomó (no se trata de una fórmula muy conocida).

95 Gómez Fernández (1981) cita coeficientes *inferiores* a .50 en la versión española de un test de Cattell; los tests de personalidad de Cattell suelen medir rasgos concebidos a un nivel muy complejo. Cattell no considera una fiabilidad muy alta como deseable porque implica una simplicidad que juzga excesiva en la concepción del rasgo (*ítems muy repetitivos*) aunque naturalmente este nivel de complejidad o simplicidad depende de lo que el autor del instrumento quiere medir y hacer.

96 Las fórmulas adecuadas (denominadas *corregidas por atenuación*) pueden verse en Morales, 2008, cap. 5; apartado 4.1.; no deben utilizarse con muestras inferiores a unos 300 sujetos. Son fórmulas muy sencillas; *basta dividir el coeficiente de correlación por la raíz cuadrada del coeficiente de fiabilidad conocido*, o *por el producto de las raíces cuadradas de los dos coeficientes de fiabilidad si conocemos los dos*.

97 Las fórmulas del *error típico de la medida* (de las puntuaciones individuales), los *intervalos de confianza* y su interpretación en Morales (2008, cap. 6).

12.3. Selección de los ítems definitivos

Aunque llevemos a cabo todo el proceso con un programa de ordenador (SPSS), debemos tener muy claro qué es lo que estamos haciendo. Lo que vamos a hacer es *calcular la fiabilidad con distintos subconjuntos de ítems* para quedarnos finalmente con la selección de ítems que más nos convenza como versión definitiva de nuestro instrumento.

La norma de retener el subconjunto de ítems que nos de la máxima fiabilidad es válida en principio, aunque siguiendo este criterio de manera muy literal no obtendremos necesariamente el mejor instrumento posible, por eso veremos en otro apartado (12.3.2) otros criterios complementarios que pueden ser útiles para elegir los ítems definitivos.

12.3.1. Según el análisis de ítems

El proceso, como vamos viendo, es éste:

1º Calculamos el coeficiente de fiabilidad con todos los ítems iniciales;

2º Vamos eliminando los peores ítems (los que tienen una menor correlación con el total) y volvemos a calcular la fiabilidad y así sucesivamente hasta que nos quedamos con el conjunto de ítems que nos da la mayor fiabilidad. El SPSS ya nos va indicando qué ítems hay que suprimir para que suba la fiabilidad.

Si hemos analizado los ítems mediante el contraste de medias entre los grupos extremos, el criterio será la diferencia entre estos dos grupos (a mayor diferencia, mejor ítem, más discriminante).

3º Cuando al eliminar ítems vemos que baja la fiabilidad, *en principio* damos por terminado el trabajo de construcción de la escala; nos quedamos con el subconjunto de ítems que forme una escala con una fiabilidad óptima. Decimos *en principio* porque caben otras consideraciones en la selección definitiva de los ítems.

Los ítems los vamos suprimiendo de uno en uno, o en pequeños bloques. No se trata de un proceso totalmente *mecánico*, pues como comentaremos después, pueden entrar otras consideraciones en la elección de los ítems, pero los ítems que vamos reteniendo deben ser ítems que correlacionan bien con el total (o que diferencian bien a los sujetos en los grupos extremos).

Como vamos a calcular la fiabilidad con distintas combinaciones de ítems, se puede ir dejando constancia del proceso (*su historia*) tal como aparece en la tabla 4; es una información útil que además se puede presentar en un trabajo de investigación o en una tesis.

En vez de poner en la primera columna los ítems que vamos *eliminando*, podemos poner los que *retenemos* en cada versión.

ítems en las versiones sucesivas	*número de ítems*	*media*	*desviación típica*	*fiabilidad* α
Todos los ítems				
Eliminamos ítems nº				
Eliminamos ítems nº				
Eliminamos ítems nº				

Tabla 4

También cabe seguir el *procedimiento inverso*, recomendado por algunos autores y que puede ser preferible. En vez de ir eliminando progresivamente los ítems que menos discriminan, podemos comenzar reteniendo los que más discriminan.

1º Calculamos la fiabilidad con el subconjunto de ítems que más discriminan (mayor correlación con el total o mayor diferencia entre los grupos extremos).

2º Añadimos unos pocos ítems, los más discriminantes de los que nos quedan, y volvemos a calcular la fiabilidad.

3º Damos la tarea por terminada cuando la fiabilidad empieza a bajar o simplemente no sube de manera apreciable.

Con este procedimiento nos quedará normalmente una escala *más breve*, sobre todo si partimos de muchos ítems. Los ítems se pueden ir añadiendo de uno o en uno, o en pequeños bloques de ítems de discriminación parecida. Como antes, se pueden tener también criterios más conceptuales, para que nos quede un instrumento equilibrado y a nuestro gusto.

Estos procesos, seguidos de manera *automática*, nos llevan a construir instrumentos (*escalas de actitudes* en nuestro caso) de una calidad adecuada:

1º La validez, al menos conceptual, la hemos ya procurado al redactar los ítems,

2º El análisis de ítems nos permite a desechar los peores ítems y conseguir una fiabilidad aceptable (al menos la mayor posible en nuestro caso).

El proceso de construcción de una escala de actitudes no tiene por qué ser *automático.* En la elección definitiva del conjunto de ítems que van a formar la escala definitiva pueden entrar también otros criterios *más conceptuales* que *modulen* esta selección; con frecuencia tendremos versiones con distinto número de ítems que apenas difieren en fiabilidad.

Como estamos tratando del análisis de ítems y de la fiabilidad como criterios (no únicos) de calidad, no debemos olvidar que aunque es verdad que *en general* a mayor número de ítems tendremos una mayor fiabilidad, también es verdad que:

1) *Con más respuestas en los ítems* también aumenta la fiabilidad, y en muchos casos puede ser preferible aumentar el número de respuestas en vez de aumentar el número de ítems.

2) Ya hemos visto en el apartado 8.2 sobre *número de ítems y fiabilidad* ejemplos de escalas con muy pocos ítems (entre dos y cinco) y con una fiabilidad aceptable e incluso alta, aunque en estos casos se trata de ítems con formulaciones muy parecidas (se miden actitudes o rasgos concebidos de manera muy simple) y de muestras grandes (en las que es más probable que haya sujetos muy distintos).

12.3.2. Otros criterios en torno a la elección de los ítems definitivos

Siguiendo el proceso indicado (con el SPSS) podemos construir fácilmente una escala o test de buena calidad, al menos tomando la correlación ítem-total y la fiabilidad como criterios. Aun así podemos incorporar otros criterios en la selección definitiva de los ítems si tenemos en cuenta que:

a) Con diversos conjuntos de ítems podemos llegar a coeficientes de fiabilidad muy parecidos.

b) Diferencias pequeñas en fiabilidad no tienen mayor importancia, además estos coeficientes no serán los mismos en muestras distintas. Puede ser preferible otra combinación de ítems que por alguna razón nos gusta más *aunque baje algo la fiabilidad* (bien entendido que se mantiene dentro de unos límites aceptables). Proponemos algunos de estos criterios.

1º *Equilibrio entre ítems positivos y negativos*

Ya hemos visto en otro apartado (6.4, *ítems positivos y negativos*) que tiene sus ventajas (que podemos repasar) el que haya ítems en las dos direcciones (*favorables* y *desfavorables* a

la actitud medida), incluso podemos procurar que ambos tipos de ítems entren en la misma o parecida proporción. En la escala definitiva podemos incorporar los mejores ítems de cada dirección, controlando con la clave de corrección el que todas las respuestas se puedan sumar (el *máximo acuerdo* en unos ítems será equivalente al *máximo desacuerdo* en otros ítems).

También un examen más pormenorizado de los ítems (de su correlación ítem-total) nos puede llevar a la conclusión de que todos deberían estar en la misma dirección (favorable o desfavorable).

2º Cuidar más la *representatividad* del contenido de los ítems

En le selección definitiva de los ítems que van a conformar nuestra escala, podemos buscar el que todos expresen el mismo rasgo de manera más nítida para facilitar la *interpretabilidad* de los datos en función de un rasgo o una actitud definida en términos más precisos. A veces el prescindir de uno o dos ítems (baja correlación ítem-total) nos puede llevar a prescindir de otros de significado semejante, que funcionan bien empíricamente (buena correlación ítem-total) pero que de alguna manera desequilibran el significado de lo que medimos de hecho.

Un ejemplo posible. En una escala de *motivación de logro*, y según nuestro marco teórico y plan inicial, podemos incluir ítems que expresan *nivel alto de aspiraciones, constancia en el trabajo, aceptación de riesgos, objetivos a largo plazo,* etc. Si vemos que dos o tres de estos ítems sobre *aceptación de riesgos* (o sobre cualquier otro de los aspectos que hemos especificado) no discriminan en el conjunto de la escala (aunque algún otro ítem semejante sí discrimine bien), puede ser preferible no incluir en la escala el componente de *actitud hacia el riesgo* (aunque baje algo la fiabilidad) y nos quedará un concepto de *motivación de logro* más equilibrado aunque quizás con un significado más restringido. Si nos interesa también medir la *actitud hacia el riesgo*, podemos hacerlo con otro instrumento o con unas preguntas adicionales. Nos quedará un concepto de *actitud hacia el estudio* con un significado más restringido pero más claro que el que buscábamos en primer lugar.

También puede suceder que eliminando un ítem *apenas* baje la fiabilidad pero si lo eliminamos la escala puede quedar menos coherente con otros criterios (una idea o aspecto del constructo puede quedar mal representada o de manera muy incompleta); en ese caso puede merecer la pena retenerlo.

3º Incluir de manera equilibrada aspectos distintos del mismo rasgo general (subescalas)

A veces nos interesa medir un rasgo o actitud concebidos a un nivel muy general pero de manera que el instrumento se pueda descomponer en *subescalas* que miden aspectos distintos. *Con esta finalidad* al construir y analizar la escala podemos seleccionar los ítems de manera que todos los aspectos estén suficientemente representados. En este caso tanto la escala formada por todos los ítems como cada subescala (*si* también se van a utilizar como escalas independientes) deben tener una fiabilidad aceptable.

De este tipo de escalas tenemos muchos ejemplos; uno es una escala de *autoeficacia docente* (Prieto, 2007)[98] de la que se deriva una *puntuación total* en autoeficacia y cuatro puntuaciones parciales en *autoeficacia* 1) *para planificar,* 2) *para implicar activamente a los alumnos,* 3) *para interactuar positivamente en el aula* y 4) *para evaluar*.

Sobre estos instrumentos *subdivididles* en subescalas:

a) Es preferible (aunque no necesario) que en estos instrumentos cada subescala esté compuesta por un mismo número de ítems; de esta manera se facilita la comparación de las

98 Instrumento reproducido en Morales (2011, *cuestionarios y escalas*).

medias de las diversas subescalas, aunque dividiendo la media de cada subescala por el número de ítems que la compone también tenemos valores directamente comparables.

b) Además de *toda* la escala podemos utilizar las *subescalas* en el resto de los análisis (correlaciones con otras variables, comparar grupos).

c) Las subescalas son útiles con una *finalidad diagnóstica*; en el ejemplo mencionado (Prieto, 2007) podríamos preguntarnos *¿En qué aspecto de la autoeficacia convendría tener alguna actividad formativa con los profesores?*

El *análisis factorial* es un buen procedimiento (es el habitual) para identificar los ítems de cada subescala. También se puede seguir con *cada subescala* el proceso seguido para construir *toda* la escala: 1º se identifican los ítems de cada subescala con criterios conceptuales y 2º se examina después la fiabilidad de cada subescala.

4º Incorporación de nuevos ítems

A veces nos encontramos con esta situación: unos pocos ítems nos convencen porque expresan bien lo que realmente deseamos medir, pero con estos ítems no llegamos a una fiabilidad que juzgamos adecuada. En este caso podemos acudir a las fórmulas que nos dicen cuántos ítems del mismo estilo (de formulaciones parecidas) deberíamos añadir para alcanzar una fiabilidad determinada[99].

También puede suceder que encontremos ítems que haya que eliminar por su baja relación con el total pero que conceptualmente nos parecen interesantes; es posible que estén formulados de manera que discriminan poco y podemos intentar una redacción nueva.

El incorporar nuevos ítems supone que pensamos mejorar el instrumento en ediciones o investigaciones futuras.

5º Preparación de dos versiones, corta y larga, de la misma escala

Es muy normal que con un número muy reducido de ítems consigamos una fiabilidad alta (a veces muy alta). Podemos verificar qué fiabilidad obtenemos seleccionando solamente los mejores ítems, que pueden ser muy pocos (por ejemplo entre tres y cinco). Nos puede interesar tener disponibles dos versiones de la misma escala; una la versión normal o *larga,* y otra *breve,* que puede ser muy útil para otros usos de estos instrumentos (como instrumento *complementario* en otras investigaciones, para dar un rápido *feedback* a un grupo que puede responder en muy poco tiempo a unas pocas preguntas y sin mucho esfuerzo por nuestra parte para analizarlas, etc.).

12.3.3. Explicación o *redefinición* del rasgo medido por nuestro instrumento

En esta observación no nos referimos a la selección definitiva de los ítems, sino a las *consecuencias* de esta selección. A veces, y a la vista de los ítems que han sido retenidos en la escala definitiva, habrá que *redefinir* lo que pretendemos medir o al menos explicarlo adecuadamente aunque se mantenga el nombre del instrumento.

Los términos para designar los rasgos suelen ser muy genéricos, y de hecho instrumentos con el mismo nombre (como *actitud hacia el estudio*, *autoestima,* etc.) pueden no coincidir en lo que de hecho miden, que puede ser definido con unos límites o más amplios o más ajustados. Podemos comenzar, por ejemplo, construyendo una escala de *actitud hacia el estudio,* pero al eliminar una serie de ítems y fijarnos en los que nos quedan, puede ser preferible hablar de *nivel de aspiraciones,* o de *autorregulación en el estudio,* etc. (lo que de

[99] Estas fórmulas que relacionan la longitud de un test y su fiabilidad están en Morales, Urosa y Blanco (2003) y también suelen encontrarse en textos que se tratan de manera más específica sobre la fiabilidad, la construcción de tests o la psicometría en general.

hecho vemos que estamos midiendo con los ítems seleccionados). Al menos debe quedar explicado de alguna manera.

13. Comprobación de la validez de la escala y otros análisis posteriores

Una vez que tenemos ya la versión definitiva de la escala, se hacen los demás análisis según los datos de que dispongamos.

a) Se pueden calcular *datos descriptivos* (medias y desviaciones) de las diversas submuestras si las hay,

b) Podemos construir *baremos* o *normas* de interpretación de los resultados individuales (como los *percentiles, estaninos,* u otro tipo de puntuaciones)[100].

c) Sobre todo podemos comprobar de manera más específica y planificada la *validez* del *nuevo instrumento* con los datos que hemos obtenido simultáneamente[101] y continuar con nuestra investigación.

13.1. Conceptos básicos sobre la validez de tests y escalas

No tratamos aquí de manera específica sobre la validez, pero es útil recordar ahora algunas ideas básicas sobre la validez de tests y escalas y cómo confirmarla. No hay una prueba de validez en sentido estricto, pero sí podemos tener datos que apoyen una determinada *interpretación* o avalen la *utilidad* del instrumento.[102]

La confirmación de la validez más que un cálculo es un *proceso*; los llamados *coeficientes de validez* son simples correlaciones con un determinado criterio que no confirman necesariamente la validez de un instrumento, sino una interpretación específica de los datos obtenidos con ese instrumento; (hablar de *coeficientes de validez* es un tanto equívoco; no hay *un* coeficiente de validez análogo a los coeficientes de fiabilidad).

La validez de un instrumento no se prueba de manera categórica (*sí* o *no*), pero sí se pueden acumular datos que van clarificando y ampliando el significado de lo que medimos al ver con qué otras variables se relaciona; en expresión Cronbach (1971) *validar es investigar*. Cuando se construye un instrumento (una escala) para hacer una determinada investigación (como una tesis), la misma investigación ya suele aportar información sobre la validez del instrumento.

Aunque hablamos de la validez *de* tests y escalas, como si fueran *propiedades* de estos instrumentos, hablando con propiedad la validez no es una propiedad del instrumento, sino de las *inferencias e interpretaciones* que hagamos con los datos obtenidos.

Una visión de conjunto de lo que entendemos por *validez* y de los modos de comprobarla está resumida la figura 26[103]. Es importante tener una cierta claridad sobre el concepto de validez porque de hecho se utiliza con *adjetivos* y *significados* distintos. La validez se refiere unas veces (las *filas* de la figura 26) a la *interpretación* de lo que medimos (validez de *constructo* y validez *predictiva*) y otras veces (las *columnas* de la figura 26) al

---

100 Cómo construir estas *normas de interpretación* en Morales (2008, cap. 4, *Tipos de puntuaciones individuales*).

101 Ya indicamos al comienzo, al explicar el proceso de construcción de una escala (figura 21) que *además* de redactar los ítems conviene pensar en *preguntas adicionales*, precisamente para confirmar la validez.

102 Sobre la validez las normas de la *American Educational Research Association* dicen que *la validez se refiere al grado en el que la evidencia y la teoría apoyan (support) las interpretaciones de los tests de acuerdo con el uso que se va a hacer de estos tests.*

103 Adaptada y modificada de Morales, Urosa y Blanco (2003) y Morales (2006, donde se trata con amplitud el tema de la validez, capítulos 12, *aspectos conceptuales,* y 13, *aspectos metodológicos*).

modo de *comprobación* (*analizando el contenido* o *con estudios experimentales*)[104]. El concepto de validez más importante e integrador es el de *validez de constructo.*

	Cómo comprobamos el significado y la utilidad	
Las interpretaciones se pueden reducir a dos grandes tipos	**Analizando el contenido** **VALIDEZ LÓGICA, APARENTE o CONCEPTUAL**	**Con métodos experimentales** ***Verificando hipótesis coherentes*** **VALIDEZ EXPERIMENTAL**
a) Sobre el SIGNIFICADO **¿Medimos lo que decimos que medimos?** **¿Con qué otras variables se relaciona? (ampliamos y matizamos el significado)** **VALIDEZ DE CONSTRUCTO**	**Necesario pero no siempre suficiente**	**¿Por qué?** **A pesar de la *validez aparente:* Podemos medir en parte *algo distinto* (como capacidad lectora)** **En las respuestas *pueden influir otras variables* (presentar una buena imagen, aquiescencia, cansancio, no entender lo que se pregunta o entenderlo de otra manera, etc.)**
b) Sobre la UTILIDAD del instrumento **VALIDEZ PREDICTIVA** **(en sentido amplio)**	**No es una estrategia válida aunque el análisis del contenido ayuda formular hipótesis predictivas**	**Necesario siempre,** **métodos correlacionales** **Posibles problemas: escoger criterios adecuados, validez y fiabilidad del criterio**

Figura 26

Con los estudios de *validación* pretendemos, sobre todo, dos finalidades que se apoyan mutuamente (*significado* y *utilidad*, a y b en la figura 26).

a) *Confirmar el significado* previsto de la variable (de la actitud o rasgo) que pretendemos medir (*validez de constructo*). Es el significado más habitual de *validez* aplicado a tests y escalas.

Se trata de verificar que la interpretación que hacemos es correcta. Si, por ejemplo, decimos que estamos midiendo *actitud hacia el estudio*, verificamos que es *eso*, y no otra cosa lo que medimos con nuestro instrumento. Como se indica en la figura 26, en las respuestas pueden influir otras variables, como pueden ser el deseo más o menos consciente de *presentar una buena imagen de uno mismo, capacidad lectora*, etc.; por eso necesitamos o es conveniente una verificación experimental, supuesta siempre la *validez conceptual.* Este tipo de validez suele denominarse *validez de constructo* (constructo = rasgo).

Confirmamos el significado *comprobando hipótesis* basadas en el mismo significado de lo que pretendemos medir con nuestro instrumento. Podemos utilizar dos tipos de estrategias que se complementan:

1. Validez *convergente:* por ejemplo comprobando relaciones *esperadas y plausibles* (positivas o negativas) con otras medidas:

   1.1. Unas medidas pueden ser otros instrumentos que pretendidamente miden *lo mismo* (si hacemos una escala de autoconcepto esperaremos una correlación significativa con otras escalas de autoconcepto o de constructos semejantes).

   1.2. Otros instrumentos pueden medir *otras cosas* pero con las que esperamos que haya relación (como entre actitud hacia el estudio y calificaciones escolares).

Podemos también comprobar si la escala *diferencia grupos* que *ya sabemos* (al menos lo suponemos como hipótesis) que son diferentes en esa variable.

---

[104] No hay que olvidar *la validez ética* del uso pretendido (y sus consecuencias) de estos instrumentos.

2. Validez *divergente:* comprobando que el rasgo *no tiene relación* con otros rasgos con los que no esperamos que la tenga o *sí* la tiene pero es menor que con otros rasgos con los que es más claro que debe tenerla mayor.

Hay otros métodos para confirmar, matizar, describir mejor o explorar el significado de lo que medimos, como es el *análisis factorial* (un análisis de la *estructura* del instrumento).

b) *Comprobar la utilidad* práctica del instrumento

En este caso verificamos si existen correlaciones apreciables con determinados criterios (como *rendimiento académico, éxito en una tarea,* etc.); se trata de *validez predictiva* en un sentido amplio pues no siempre pretendemos *predecir*.

Esta comprobación de la *utilidad* también aporta datos a la comprobación del *significado (validez de constructo)*; si comprobamos que un test de inteligencia tiene una correlación significativa y apreciable con rendimiento académico estamos comprobando la *utilidad* del test y a la vez esta correlación nos confirma que el test de inteligencia mide precisamente inteligencia.

13.2. Sugerencias para obtener datos adicionales que faciliten la validación de la escala

Si la escala o test no es de nueva construcción, o se trata de una adaptación de otro instrumento, la validez confirmada experimentalmente en otros estudios puede ser suficiente para juzgar que el test o escala mide lo que se pretende, sobre todo si la *validez conceptual* es muy clara. Nosotros podemos confirmar o ampliar el significado y la utilidad de la escala o test con análisis semejantes a los que vamos a exponer para validar un nuevo instrumento; casi cualquier investigación hecha con uno de estos instrumentos aporta información sobre su validez.

Las sugerencias puestas a continuación están estructuradas pensando en los posibles análisis sobre la *validez,* pero en cualquier caso siempre es útil e informativo hacer algún estudio de tipo correlacional o de comparación de grupos independientemente de la intención de validar el instrumento (puede haber ya otras muchas investigaciones sobre la validez del mismo instrumento o de otros parecidos).

Cuando construimos o utilizamos un instrumento como una escala de actitudes *no nos quedamos ahí,* solemos tener un planteamiento de investigación (como en una tesis) y por lo general los análisis propios de la investigación ya aportan información sobre la validez del instrumento. Si lo que nos planteamos es solamente construir el instrumento, además habrá que validarlo de alguna manera, harán falta otro tipo de datos, y eso ya es una investigación.

Estos datos adicionales ya los hemos mencionado en uno de los apartados anteriores (nº 9, *preparar preguntas o instrumentos adicionales: estructura del instrumento*). Cuando preparamos una escala de actitudes debemos preguntarnos:

- → *¿Con qué otros rasgos, conductas, actitudes, etc., puede estar relacionada la actitud que quiero medir?*
- → *¿Qué grupos pueden ser distintos en esta actitud?*

De aquí nos saldrán ideas e hipótesis para formular preguntas que nos permitirán verificar la validez de la escala o hacer una investigación más completa. Muchas de estas ideas las podemos encontrar en la literatura experimental sobre el constructo que queremos medir, sobre todo en estudios sobre la construcción de instrumentos semejantes.

Estos *otros datos* podemos obtenerlos con otras escalas y otros tests y es frecuente hacerlo así, pero las sugerencias que vamos a dar están orientadas a obtener la información de

interés de manera *más rápida y económica*, de manera que el cuestionario que responden los sujetos no sea muy largo.

Las mismas sugerencias ya hechas para formular los ítems de una escala son válidas para formular otras preguntas que no van a formar parte de la escala pero sí van a formar parte del cuestionario o instrumento que van a responder los sujetos[105].

13.2.1. Confirmación del significado pretendido *(validez de constructo)*

Los análisis pueden tener dos enfoques *básicos* (hay más), como son:

a) Los estudios correlacionales,
b) Las comparaciones entre grupos.

Como ya se ha indicado, con estos análisis podemos responder a otras preguntas de investigación que no siempre tienen por finalidad validar un instrumento.

a) *Análisis correlacionales*

Por lo que respecta a los *análisis correlacionales*, y para tener una visión de conjunto, vamos a pensar en tres tipos de relaciones (o en su caso, de *no* relaciones):

1º Relaciones *positivas* con *el mismo rasgo* (más o menos) medido de otra manera, con otro instrumento; ésta es la confirmación más directa de la validez de constructo.

2º Relaciones *positivas o negativas* con otros *rasgos distintos* que, al menos como hipótesis razonable, pueden estar relacionados positiva o negativamente con el rasgo que medimos con nuestro instrumento.

3º Relaciones muy bajas, no significativas, con otros rasgos con los que esperamos o que *no* haya relación o que ésta sea menor que otras.

Vamos a dar sugerencias metodológicas sobre estas tres estrategias, pero dado lo común que son los análisis correlacionales para validar un instrumento (y en general en cualquier investigación) en la figura 27 tenemos un esquema con los tipos de variables que podemos utilizar.

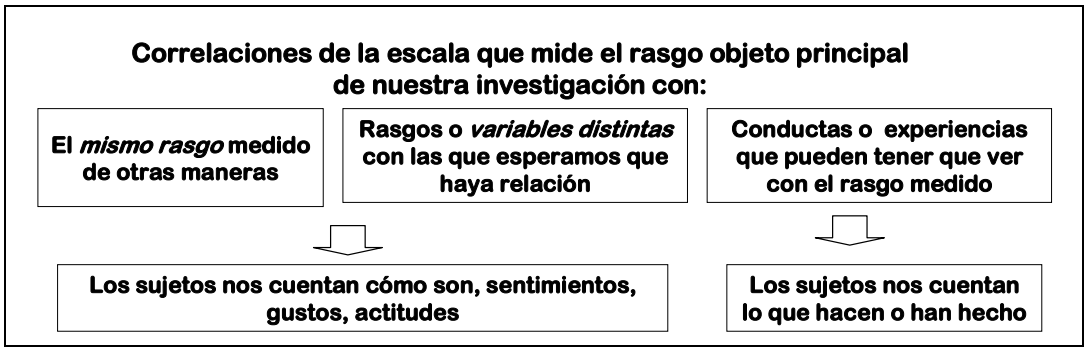


Figura 27

1º *Relación con otros modos de medir el mismo rasgo*

Comprobamos la relación entre lo que mide nuestro instrumento y otros modos de medir *el mismo rasgo.* Otros modos de medir *lo mismo* pueden ser:

a) Un ítem de *autodescripción global* que resume la misma actitud o rasgo que queremos medir con nuestro test o escala.

---

[105] En Morales (2006, en los anexos) pueden verse numerosos ejemplos de escalas junto con *preguntas adicionales* de diverso tipo (preguntas independientes, listas de adjetivos, *Diferencial Semántico,* etc.) para hacer estudios de validación y otros análisis; también hay ejemplos de estas preguntas en Morales, Urosa y Blanco (2003).

Por ejemplo, para confirmar la validez de una escala de *asertividad*, se puede *medir* también la asertividad con esta pregunta en la que se describe la personalidad asertiva (Gismero, 1996):

En qué medida cree Vd. que esta descripción refleja cómo es Vd.:

> *Soy una persona que sabe defender sus derechos y 'plantarse' ante quien sea necesario, sin que eso me cree gran conflicto y a la vez sin hacer daño a otros ni provocar en ellos agresividad.*

Las respuestas pueden ser seis (como en este caso) o más, describiendo solamente los extremos *yo soy así* y *yo no soy así en absoluto* (en este ejemplo la correlación con la escala de *asertividad* es de .557; con N= 404, p <.001).

Son bastantes las investigaciones (muchas fácilmente localizables) que muestran que un solo ítem bien formulado correlaciona bien con instrumentos de más ítems y más elaborados.[106]

b) Otro *mini-test* de pocos ítems que más o menos mida lo mismo, por ejemplo:

1. Desde otra perspectiva (pueden ser *conductas probables* en vez de opiniones),
2. Con un breve instrumento que mida el mismo rasgo pero construido con una técnica distinta (por ejemplo un *Semántico Diferencial*)[107].

c) *Otra escala* o test (o *factor,* o subescala, o selección de ítems, etc., de *otro instrumento)* de otro autor y que supuestamente comprueba lo mismo (o muy semejante).

d) Si es posible, se puede comprobar la relación entre *autoevaluación* (sobre todo si se trata de la medición de un rasgo propio más que la actitud *hacia otra cosa)* y *héteroevaluacion*.

Estos *nuevos instrumentos* tienen un valor complementario, pueden ser muy breves y también pueden limitarse a simples preguntas.

2º Comprobación de relaciones esperadas con otras variables.

Antes nos referíamos a *otros modos* de medir *el mismo rasgo,* ahora se trataría de comprobar también la relación entre lo que mide nuestro instrumento y *otras variables distintas* que no mide nuestro instrumento y con las que esperamos que haya relación *(positiva o negativa).*

Estas *otras variables* pueden ser de muchos tipos, damos algunos ejemplos.

a) *Rasgos o características personales.*

Estos rasgos personales podríamos comprobarlos con tests de personalidad pero también con procedimientos más sencillos, como pueden ser 1) listas de adjetivos o 2) una serie de ítems autodescriptivos en los que cada uno expresa un rasgo distinto y 3) como antes, una *autodescripción global* que resume *otro rasgo* presumiblemente relacionado con el que medimos con nuestra escala.

---

[106] Mencionamos (en nota a pié de página) ejemplos en el apartado 1.3. *Los cuestionarios: ¿Podemos 'medir' actitudes con una sola pregunta?;* entre otros ejemplos posibles Davey, Barrat, Burrow y Deeks (2007) en una muestra de N = 400 encuentran una correlación de .78 entre un solo ítem de *ansiedad* y un test completo de ansiedad (STAI, *State Trait Anxiety Inventory*).

[107] En Morales (2006, pág. 601) se utiliza un diferencial semántico de 12 pares de adjetivos para valorar el sistema democrático de gobierno, la suma de las respuestas al diferencial semántico (analizado como una escala) tiene una fiabilidad de .935 (con N = 323 adolescentes de 13 a 15 años) y una correlación de .84 con una escala de actitudes hacia la democracia.

1) *Lista de adjetivos* que podrían *equivaler* a una serie de tests de personalidad. Una escala de motivación de logro podríamos esperar que tuviera correlaciones significativas con autodescripciones como *ambicioso, constante, organizado,* etc.

En la tabla 5 tenemos las correlaciones de una escala de *actitud hacia el estudio* con una selección de *adjetivos autodescriptivos* (respuestas de 4 = *mucho* a 1 = *nada*) en una muestra de 174 niños y niñas.[108]

Inteligente	r = 0.368	(p < .001)
Perezoso	r = -0.355	(p < .001)
Trabajador	r = 0.439	(p < .001)

Tabla 5

Son relaciones *plausibles:* positivas con autodescribirse como *inteligente* y *trabajador,* negativa con *perezoso* (lo que además avala la *sinceridad* básica de las respuestas[109]).

También se pueden *sumar* adjetivos que reflejen más o menos el mismo rasgo, seleccionados con criterios lógicos o mediante el *análisis factorial* con el que se pueden localizar grupos de adjetivos relacionados entre sí y que reflejan un mismo rasgo subyacente a todos ellos.

2) *Lista de breves autodescripciones*. Ya hemos visto que una lista de adjetivos puede constituir un modo breve y económico de medir rasgos de personalidad (como instrumento complementario, no para hacer un psicodiagnóstico); la misma función la puede cumplir una serie de breves frases autodescriptivas.

En la tabla 6 tenemos los coeficientes de correlación entre un test sobre el *ver sentido a la vida* (basado en la teoría de Frankl) y afirmaciones autodescriptivas con cinco respuestas: *no soy así en absoluto, algo, a medias, bastante, me describe muy bien*[110].

Soy una persona feliz…	.572 p<.001
Me desaliento con facilidad, tengo una cierta tendencia a la depresión…	-.481 p<.001
Acepto con facilidad las cosas que me molestan y me adapto a las situaciones…	.230 p<.05
Soy muy estable emocionalmente…	.373 p<.001
Soy una persona más bien insegura…	-.405 p<.001
Dependo mucho de los demás, me cuesta decir que no si los demás dicen que sí	-.348 p<.01
Soy más bien suspicaz, me siento herido con facilidad…	-.374 p<.001
Soy una persona tranquila, nada nerviosa…	.423 p<.001

Tabla 6

Estas *autodescripciones* (excepto la primera, *soy una persona feliz*) están *inspiradas* en los estudios citados en el manual del test; test completos (por ejemplo de *neuroticismo*) mencionados en el manual por sus correlaciones (positivas o negativas) con la variable medida se han convertido en simples afirmaciones que cumplen bien su función (ver con qué rasgos de personalidad *conecta*, positiva o negativamente, el *ver sentido a la vida*).

3) Una *autodescripción global* de *otro rasgo distinto* previsiblemente relacionado con el que medimos con nuestro instrumento; por ejemplo; para medir *autoestima* o *satisfacción con uno mismo* (Gismero, 1996) y verificar la correlación con una escala de *asertividad*:

---

[108] El estudio completo en Morales (2006, 535-547); la *actitud hacia el estudio* (tal como se mide aquí) también está relacionada con atribuciones internas y externas del éxito y del fracaso.

[109] Es muy complicado mentir tan sistemáticamente.

[110] La muestra es de N = 80; la versión en español de este test y más información y análisis en Morales (2011, *cuestionarios y escalas*).

En qué medida cree Vd. que esta descripción refleja cómo es Vd.:

> *En conjunto estoy satisfecho de ser tal como soy; me encuentro a gusto y no cambiaría demasiadas cosas en mí. Puede decirse que en conjunto soy la persona que me gustaría ser.*

La correlación entre este ítem de *satisfacción con uno mismo* (seis respuestas) y *asertividad* (se trata de validar la escala de *asertividad*) es de .583 ($p<.001$)[111].

b) *Expectativas, preferencias, otras actitudes, 'gustos y disgustos',* etc.

Palabras distintas pueden sugerir posibilidades distintas que, como hipótesis, puedan tener relación con la actitud medida por nuestro instrumento.

Por ejemplo, escalas que miden actitudes o variables relacionadas con el estudio (*actitud general hacia el estudio, autorregulación en el estudio, enfoques en el aprendizaje, autoeficacia académica, motivación,* etc.) pueden estar relacionadas con otras variables del ámbito académico que podemos *medir* con preguntas en torno a:

- *Gusto por la asignatura, carrera, etc.,*
- *Dificultad percibida,*
- *Utilidad percibida,*
- *Nivel de aspiraciones,*
- *Influjo de la suerte en los exámenes.*

Con frecuencia nos interesa tener información sobre el *rendimiento académico* de los alumnos a la que no tenemos acceso porque los cuestionarios son (y deben ser) anónimos, pero siempre cabe hacer preguntas que de manera indirecta y menos exacta nos den una información equivalente o muy probablemente relacionada con calificaciones[112].

c) *Experiencias, conductas y hábitos personales.*

Esta categoría de *posibles preguntas* no nos remite a *sentimientos* y *creencias* sobre uno mismo, sino a lo que los sujetos *hacen* o *han experimentado,* como pueden ser experiencias de trabajo o de otro tipo, contactos con otros tipos de personas, etc. Cuando construimos un instrumento como una escala de actitudes y revisamos estudios experimentales sobre instrumentos semejantes, tenemos que estar *atentos desde el principio a estas otras variables* que se ha comprobado que están relacionadas con la que es objeto de nuestro interés; luego ya veremos la manera de obtener esta información de la manera más conveniente que puede ser muy sencilla.

En la tabla 7 tenemos los coeficientes de correlación de una escala *hacia las personas con discapacidad* (N = 138 estudiantes de psicología) con dos ítems que remiten a *experiencias con discapacitados* (Batz, 2005).

¿Ha tenido contacto personal con personas con discapacidad? *(dos respuestas, Sí = 1, No = 0)*	.178 p = .037
¿Ha tenido alguna experiencia de trabajo con personas con discapacidad? *(dos respuestas, Sí = 1, No = 0)*	.256 p = .002

Tabla 7

[111] Más información bibliográfica y ejemplos sobre el uso de estas preguntas y breves instrumentos en los procesos de validación en Gismero (1996:140) y Morales (2006: 468).

[112] Hemos visto algunas sugerencias en el apartado 1.2.3

Las dos correlaciones son bajas pero coherentes con lo que se pretende medir (mejor actitud hacia los discapacitados los que han respondido *sí* a estas dos preguntas que reflejan experiencias personales, no opiniones) y *estadísticamente significativas* (podemos excluir el azar como explicación). Con ítems dicotómicos (*sí* o *no*) no es fácil encontrar correlaciones grandes, pero son dos preguntas que remiten a experiencias, no se trata de opiniones.

En vez de calcular la correlación de la escala con estas dos preguntas (*sí o no*, 1 ó 0) podríamos haber comparado las medias en la escala de los que responden *sí* o *no* en esas preguntas; las conclusiones hubieran sido las mismas (medias más altas en los que responden *sí*).

3º *Comprobar que no hay relación donde no esperamos que la haya*

Esta *no relación* nos ayuda a distinguir unos rasgos de otros, sobre todo cuando pertenecen al misma ámbito conceptual y es fácil confundirlos (por ejemplo *asertividad* y *agresividad*).

La *no relación* no hay que entenderla necesariamente de manera literal (r = 0); puede tratarse de relaciones incluso estadísticamente significativas y de magnitud moderada, pero *menores* que con otros rasgos y esta menor relación se puede *razonar de manera plausible*.

No es nada infrecuente que *no* encontremos relación donde *sí* esperábamos encontrarla y tendremos que buscar *hipótesis explicativas* (fijándonos en las características de la muestra, de la situación, del instrumento, etc.).

b) *Comparaciones entre grupos*

El otro enfoque mencionado al principio de este apartado consiste en *comparar grupos supuestamente distintos* en aquello que estamos midiendo. Partimos de la hipótesis de que los grupos son distintos (*cualquier* diferencia entre grupos no es una prueba de validez); una escala que midiera gusto por la música *debería diferenciar* a los alumnos de un conservatorio o a músicos profesionales de la población general.

Observaciones sobre las diferencias entre grupos.

a) Todas las comparaciones entre grupos equivalen a análisis correlacionales (como en definitiva todos los análisis estadísticos). Nos da lo mismo, por ejemplo, preguntarnos *si los niños y niñas son distintos en el rasgo o actitud A* (y haremos un contraste de medias), que preguntarnos *si el sexo está relacionado o tiene que ver con la actitud A* (y calcularemos un coeficiente de correlación entre género, 1 ó 0, y la actitud A). En definitiva se trata de comprobar, de una manera u otra, si las diferencias en una variable se corresponden con diferencias en otra variable. Es más, no sólo hay una obvia relación conceptual entre las dos preguntas (*diferencia* entre los sexos en esa actitud o *relación* entre sexo y actitud), sino que disponemos de una fórmula para transformar un valor de la t de Student en un coeficiente de correlación[113].

b) Aun así nuestras preguntas espontáneas las hacemos unas veces en términos de relación, y otras en términos de diferencias; ambos enfoques nos ayudan a formular hipótesis que podemos intentar confirmar, y además aunque los procedimientos de análisis son en principio distintos (contraste de medias o correlación) en última instancia aportan la misma información.

c) Al preparar nuestro instrumento de recogida de datos, debemos pensar qué preguntas podemos hacer que identifiquen a los sujetos según *grupos de pertenencia* (profesiones, sexo,

[113] Morales (2008:284). Hay una serie de fórmulas que transforman unos valores en otros que pueden ser de utilidad.

etc.) o *según características personales* de interés que permitan subdividir la muestra en subgrupos (preferencias por A ó B, preguntas en relación con el estilo de vida, valores, etc.).

d) Si tenemos dos grupos el análisis estadístico obvio será un contraste de medias (o un coeficiente de correlación); si tenemos más de dos grupos el análisis estadístico apropiado es el análisis de varianza.

Resumiendo

Los análisis sugeridos (correlaciones y comparaciones de grupos) no agotan todos los análisis posibles en relación con la validez, pero sí son los más obvios y con frecuencia suficientes. Como ya hemos indicado antes se trata en todos los casos de obtener datos para poder *verificar hipótesis (*y también para *explorar…):*

a) El instrumento mide algo relacionado (positiva o negativamente) con *otras cosas* con las que lógicamente esperamos que haya relación (validación convergente).

Unas veces comprobamos relación *con el mismo rasgo* medido de otra manera (o por otras personas, por ejemplo auto y hétero-evaluacion);

Otras veces comprobamos relación *con rasgos distintos* pero lógicamente relacionados, al menos como hipótesis.

b) El instrumento mide algo que *no está relacionado* con lo que no se espera que lo esté (o tiene una menor relación) (validación divergente). La *no relación* también es útil para ver que no confundimos unas cosas con otras, y porque también son datos informativos.

Otra estrategia para confirmar la validez es verificar diferencias entre grupos que también, según hipótesis razonadas, podemos suponer que son distintos en aquello que estamos midiendo; por ejemplo, en una escala de *actitudes hacia el deporte* tendrá una media mayor un grupo de deportistas profesionales que un grupo equivalente (en edad, sexo, nivel socioeconómico) de la población general, y lo mismo podría decirse de un de *bomberos* con respecto a *actitud hacia el riesgo*.

13.2.2. Confirmación de la *utilidad* del instrumento *(validez predictiva)*

Significado y *utilidad* son conceptos distintos. Para confirmar la utilidad, básicamente se trata de calcular *coeficientes de correlación* entre el instrumento (lo que mide el test o escala) y determinados criterios (como rendimiento académico, determinadas habilidades, etc.).

Un problema frecuente en estos estudios suele ser que se presta mucha atención a la calidad del *instrumento predictor* (fiabilidad, validez) y no tanta (o ninguna) a cómo medimos las variables criterio. A poder ser es preferible disponer de varios criterios; por ejemplo si se trata (como es frecuente) de predecir rendimiento académico, además de notas medias finales podemos tener en cuenta número de notas altas, calificaciones en determinadas asignaturas, etc.

Aunque se trata ahora de comprobar la *utilidad* del instrumento también se pueden confirmar de esta manera *hipótesis plausibles* que apoyan o confirman el significado de lo que estamos midiendo con nuestra escala.

Estos coeficientes de *correlación con un criterio* suelen denominarse *coeficientes de validez* pero hay que caer en la cuenta de que el término es equívoco porque la validez no se concreta en *un* coeficiente específico como sí sucede con el coeficiente de fiabilidad.

También cabe hacer estudios con una finalidad *exploratoria* o que respondan a hipótesis no directamente relacionadas con la validez.

El complemento de la construcción de un instrumento pueden ser además unas *normas* de interpretación *(baremos)*, para que los sujetos que respondan puedan interpretar sus resultados individuales (como son los percentiles, estaninos, etc.)[114]

14. Bibliografía

En esta bibliografía distinguimos tres apartados:

1) Las referencias bibliográficas citadas[115],
2) Publicaciones sobre cómo construir escalas,
3) Publicaciones o documentos en los que se reproducen numerosos instrumentos.

14.1. *Referencias bibliográficas*

AMERICAN PSYCHOLOGICAL ASSOCIATION (2001). *Publication Manual of the American Psychological Association.* Washington D.C.: Author

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION and NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (1999). *Standards for Educational and Psychological Testing.* Washington DC: American Educational Research Association.

BARNETTE, JACKSON J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: if you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement,* 60 (3), 361-370.

BATZ, RUBY (2005). *Actitud del estudiante de psicología clínica de la Universidad Rafael Landívar hacia las personas con discapacidad.* Tesis de licenciatura. Guatemala: Universidad Rafael Landívar.

BIGGS, J., KEMBER, D. and LEUNG, D.Y.P. (2001). The revised two-factor Study Process Questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, 71, 133-149.

BLANCO, A., PRIETO, L., TORRE, J.C. y GARCÍA, M. (2009). Adaptación, validación y evaluación de la invarianza factorial del cuestionario revisado de procesos de estudio (R-SPQ-2F) en distintos contextos culturales: diseño del estudio y primeros resultados. *Actas del IX Congreso Nacional de Modelos de Investigación Educativa sobre "Educación, investigación y desarrollo social"*. Huelva: AIDIPE-Universidad de Huelva, pp. 1535-1543.

BORTZ, JÜRGEN; DÖRING, NICOLA (2006) *Forschungsmethoden und Evaluation*. (4ª ed. revisada). Heidelberg: Springer.

BOURNER, HILL; HUGHES, MARK & BOURNER TOM (2001). First-year Undergraduate Experiences of Group Project Work. *Assessment & Evaluation in Higher Education*, Vol. 26, No. 1, 20-39

BURDEN, PETER (2008). The use of 'Ethos indicators' in tertiary education in Japan. *Assessment & Evaluation in Higher Education***,** Vol. 33 Issue 3, p315-327

BURNS, MATTHEW K.; VANCE, DIANE; SZADOKIERSKI, ISADORA; STOCKWELL, CHUCK (2006). Student Needs Survey; A Psychometrically Sound Measure of the Five Basic Needs. *International Journal of Reality Therapy*. Vol. 25 Issue 2, 4-8

[114] *Tipos de puntuaciones individuales*, en cap. 4 de Morales (2008)

[115] En una tesis o trabajo de investigación bajo el epígrafe de *referencias bibliográficas* se ponen únicamente las publicaciones de hecho consultadas. La norma habitual (que en este caso no se sigue) es poner solamente las iniciales de los nombres propios.

CAMPO-ARIAS, ADALBERTO; OVIEDO, HEIDI CELINA and COGOLLO, ZULEIMA (2009). Internal Consistency of a Five-Item Form of the Francis Scale of Attitude Toward Christianity Among Adolescent Students. *Journal of Social Psychology*, Vol. 149, issue 2, 258-262

CAÑADAS OSINSKI, ISABEL y SÁNCHEZ BRUNO, ALFONSO (1998), Categorías de respuesta en escalas tipo Likert. *Psicothema*, vol. 10, nº 3, 623-631.

CHANG, LEI (1997). Dependability of Anchoring Labels of Likert-Type Scales. *Educational and Psychological Measurement*, 57 (5), 800-807.

CHEANG, KAI I. (2009). Effect of Learner-Centered Teaching on Motivation and Learning Strategies in a Third-Year Pharmacotherapy Course. *American Journal of Pharmaceutical Education* 73 (3) Article 42. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2703280/

CORBIERE, MARC; FRACCAROLI, FRANCO; MBEKOU, VALENTIN & PERRON, JACQUES (2006). Academic self-concept and academic interest measurement. *European Journal of Psychology of Education*, 2006, Vol. XXI, nº 1, 3-15.

CRONBACH, LEE J. (1960). *Essentials of Psychological Testing.* Second edition. New York: Harper and Row.

CRONBACH, LEE J. (1971). Test Validation en THORNDIKE, R.L., (Ed.) (1971) *Educational Measurement*. Washington D.C.: American Council of Education, 2nd edit., 335-355.

CRUMBAUGH, JAMES C. and MAHOLIC, LEONARD T. (1969). *Manual for the Purpose in Life Test*. Brookport, Illinois: Psychometric Affiliates.

DAVEY H., BARRATT A., BUTOW P. and DEEKS J. (2007). A one-item question with a Likert or Visual Analog Scale adequately measured current anxiety. *Journal of Clinical Epidemiology*, 60 (4):356-360.

DAVIES, RANDALL S. (2008). Designing a Response Scale to Improve Average Group Response Reliability *Evaluation & Research in Education*, 2008, Vol. 21 Issue 2, p134-146.

DÍAZ, DARÍO; RODRÍGUEZ-CARVAJAL, RAQUEL; BLANCO, AMALIO; MORENO-JIMÉNEZ, BERNARDO; GALLARDO, ISMAEL; VALLE, CARMEN y VAN DIERENDONCK, DIRK (2006). Adaptación española de las escalas de bienestar psicológico de Ryff. *Psicothema*, 2006, Vol. 18, pág. 572-577 http://www.psicothema.com/tabla.asp?Make=2006&Team=1003

DODEEN, HAMZEH M. (2003). Effectiveness of Valid Mean Substitution in Treating Missing Data in Attitude Assessment. *Assessment & Evaluation in Higher Education*. Vol. 28, nº 5, 505-513.

DOWNEY, RONALD G. AND KING, CRAIG V. (1998). Missing Data in Likert Ratings: A Comparison of Replacement Methods. *Journal of General Psychology*, Vol. 125 Issue 2, p175-191.

FANNING, ELIZABETH (2005). Formatting a Paper-based Survey Questionnaire: Best Practices. *Practical Assessment Research & Evaluation*, 10(12). Available online: http://pareonline.net/getvn.asp?v=10&n=12

FLERE, SERGEJ; KLANJSEK, RUDI; FRANCIS, LESLIE J. and ROBBINS, MANDY (2008). The psychometric properties of the Slovenian translation of the Francis Scale of Attitude toward Christianity: a study among Roman Catholic undergraduate students. *Journal of Beliefs & Values: Studies in Religion & Education* Vol. 29, No. 3, 313–319

FRARY, ROBERT B. (1996). Hints for designing effective questionnaires. *Practical Assessment, Research & Evaluation*, 5 (3) http://pareonline.net/getvn.asp?v=5&n=3

FRIBORG, ODDGEIR; MARTINUSSEN, MONICA and ROSENVINGE, JAN H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality & Individual Differences*, Vol. 40 Issue 5, p873-884.

GARDNER, DONALD G.; CUMMINGS, L.L.; DUNHAM, RANDALL B. and PIERCE, JON L. (1998). Single-item versus multiple item measurement: an empirical comparison. *Educational and Psychological Measurement*, 58 (6), 898-915.

GARVIN, J., BUTCHER, A., STEFANI, A., AND TARIQ, V., LEWIS, N., BLUMSOM, R., GOVIER, R. & HILL, AND J. (1995) Group projects for .first-year university students: an evaluation , *Assessment & Evaluation in Higher Education*, 20, pp. 279–294

GISMERO GONZÁLEZ, ELENA (1996). *Habilidades sociales y anorexia nerviosa.* Madrid: Universidad Pontificia Comillas.

GÓMEZ FERNÁNDEZ, D. (1981). El 'ESP-E', un nuevo cuestionario de personalidad a disposición de la población infantil española. *Revista de Psicología General y Aplicada,* 36, 450-472.

GUILFORD, JOY P. (1954). *Psychometric Methods.* New York: McGraw-Hill.

GUILFORD, JOY P. and FRUCHTER, B. (1973). *Fundamental Statistics in Psychology and Education.* New York: McGraw-Hill (en español, *Estadística aplicada a la psicología y la educación,* 1984, México: McGraw-Hill).

HAMBLETON, RONALD.K. and PATSULA, LIANE (1999). Increasing the validity of AdaptedTests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1* http://data.memberclicks.com/site/atpu/volume%201%20issue%201Increasing%20validity.pdf (consultado 29/12/10)

HEINE, STEVEN J.; LEHMAN, DARRIN R; PENG, KAIPING and GREENHOLTZ, JOE (2002). What's Wrong with Cross-Cultural Comparisons of Subjective Likert Scales? The Reference-Group Effect. *Journal of Personality and Social Psychology*, 82, 6, 903-91. http://blog.sciencenet.cn/upload/blog/file/2009/3/2009318111257156614.pdf (consultado 29/12/10).

HERNÁNDEZ SAMPIERI, ROBERTO; FERNÁNDEZ COLADO, CARLOS y BAPTISTA LUCIO, PILAR (2010). Quinta edición. *Metodología de la investigación*. México: McGraw-Hill

HERNÁNDEZ, ANA; ESPEJO, BEGOÑA and GONZÁLEZ-ROMÁ¸ VICENTE (2006**).** The functioning of central categories middle level and sometimes in graded response scales: does the label matter? *Psicothema,*Vol. 18, nº 2, pp. 300-306 http://www.psicothema.com/psicothema.asp?id=3214

KEMBER, DAVID and LEUNG, DORIS Y.P. (2005). The influence of active learning experiences on the development of graduate capabilities. *Studies in Higher Education,* Vol. 30 Issue 2, p155-170.

KLINE, PAUL (1994). *An Easy Guide to Factor Analysis.* Newbury Park: Sage.

LAIRD, THOMAS F. NELSON; SHOUP, RICK and KUH, GEORGE D. (2005). *Deep Learning and College Outcomes: Do Fields of Study Differ?* Paper presented at the Annual Meeting of the Association for Institutional Research, San Diego, CA.

LANCELLOTTI, MATTHEW and SUNIL, THOMAS (2009). To Take or Not To Take: Effects of Motivation, Selfefficacy, and Class-Related Factors on Course Attitudes. *Marketing Education Review,* Vol. 19 Issue 2, p35-47

LOUREIRO, A. & LIMA, M. L. (2009). Escala de atitudes altruístas: Estudo de validação e fiabilidade. *Laboratório de Psicologia, 7,* 73-83.

MARSHALL, MARY G., (1998). The Texas A & M University System. *Questionnaire Design: Asking Questions with a Purpose*, http://learningstore.uwex.edu/assets/pdfs/g3658-2.pdf

MARTÍNEZ, ROSARIO (2008). *Análisis de la actitud y la aplicación de estrategias de trabajo cooperativo de los profesores/as del Liceo Javier*. Tesis de Licenciatura en Educación y Aprendizaje. Guatemala: Universidad Rafael Landívar.

MEANA, RUFINO (2003). *La experiencia subjetiva de sentido y su relación con variables psicológicas y sociodemográficas*. Tesis doctoral. Madrid: Universidad Pontificia Comillas.

MESSER, S. C, ANGOLD, ADRIAN; COSTELLO, E. JEAN, LOEBER, ROLF, VAN KAMMEN, WELMOET, & STOUTHAMER-LOEBER, MAGDA (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents: Factor composition and structure across development. *International Journal of Methods in Psychiatric Research, 5*, 251-262. http://www.wpic.pitt.edu/research/famhist/PDF_Articles/John%20Wiley/M%2015.pdf (consultado 1, 01, 11)

MILLS, PAUL C.; WOODALL, PETER F. (2004). A comparison of the responses of first and second year veterinary science students to group project work. *Teaching in Higher Education***,** Vol. 9 Issue 4, p477-489.

MORALES VALLEJO, PEDRO (2011). *Tamaño de la muestra: ¿Cuántos sujetos necesitamos?* http://www.upcomillas.es/personal/peter/investigacion/Tama%F1oMuestra.pdf

MORALES VALLEJO, PEDRO (2011). *Análisis factorial en la construcción e interpretación de tests, escalas y cuestionarios.* http://www.upcomillas.es/personal/peter/investigacion/AnalisisFactorial.pdf

MORALES VALLEJO, PEDRO (2011). *Cuestionarios y escalas* http://www.upcomillas.es/personal/peter/otrosdocumentos/CuestionariosyEscalas.doc

MORALES VALLEJO, PEDRO (2010). *Evaluación de los valores: análisis de listas de ordenamiento* http://www.upcomillas.es/personal/peter/otrosdocumentos/ValoresMetodo.pdf

MORALES VALLEJO, PEDRO (2010). *Tamaño de la muestra: ¿Cuántos sujetos necesitamos?* http://www.upcomillas.es/personal/peter/investigacion/Tama%F1oMuestra.pdf

MORALES VALLEJO, PEDRO, (2009). *Análisis de varianza para muestras relacionadas.* http://www.upcomillas.es/personal/peter/analisisdevarianza/MuestrasRelacionadas.pdf

MORALES VALLEJO, PEDRO (2008). *Estadística aplicada a las ciencias sociales.* Madrid: Universidad Pontificia Comillas.

MORALES VALLEJO, PEDRO (2006). *Medición de actitudes en Psicología y Educación, construcción de escalas y problemas metodológicos,* tercera edición revisada. Madrid: Universidad Comillas.

MORALES VALLEJO, PEDRO; UROSA SANZ, BELÉN y BLANCO BLANCO, ÁNGELES (2003). *Construcción de escalas de actitudes tipo Likert. Una guía práctica*. Madrid: La Muralla.

NAGY, Mark S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, Volume 75, Number 1, pp. 77-86

NUNNALLY, JUM C. (1978). *Psychometric Theory.* 2nd edit. New York: McGraw-Hill.

NUNNALLY JUM C. and BERNSTEIN, IRA H. (1994). *Psychometric Theory.* 3rd edit. New York: McGraw-Hill.

OLSEN, D., KUH, GD, SCHILLING, KM, SCHILLING, K., CONNOLLY, M., SIMMONS, A., & VESPER, N. (1998, November). *Great expectations: What students expect from college and what they get*. Paper presented at the annual meeting of the Association for the Study of Higher Education, Miami.

PRIETO NAVARRO, L. (2007). *Autoeficacia del profesor universitario. Eficacia percibida y práctica docente.* Madrid: Narcea.

SCHMITT, NEAL (1996). Uses and abuses of Coefficient Alpha. *Psychological Assessment,* 8 (4), 350-353 (disponible en http://ist-socrates.berkeley.edu/~maccoun/PP279_Schmitt.pdf) (consultado 30, 09, 2008).

SEIFERT, T.L. and O'KEEFE, B.A. (2001). The relationship of work avoidance and learning goals to perceived competence, externality and meaning. *British Journal of Educational Psychology*, 71, 1, 81-92.

SIMONS, JOKE; DEWITTE, SIEGFRIED and LENS, WILLY (2004). The role of different types of instrumentality in motivation, study strategies, and performance: Know why you learn, so you'll know what you learn! *British Journal of Educational Psychology*, Vol. 74 Issue 3, p343-360,

STAPLETON, LAURA M.; CAFARELLI, MICHAEL; ALMARIO, MIGUEL N. and CHING , TOM (2010). Prevalence and characteristics of student attitude surveys used in public elementary schools in the United States. *Practical Assessment, Research & Evaluation* Volume 15, Number 9 http://pareonline.net/pdf/v15n9.pdf

SUPER, DONALD E. (1968). *Work Values Inventory.* New York: Houghton-Mifflin. Disponible en University of Richmond, Career Development Center, *Work Values Inventory*, http://cdc.richmond.edu/common/pdf/valuesworkinventory.pdf . Una adaptación para responder *online* en Saint Anselm College, Career Education Services, *Work Values Inventory*, http://www.anselm.edu/administration/CES/WorkValues.htm

TIAN, XIAOWEN (2007). Do assessment methods matter? A sensitivity test. *Assessment & Evaluation in Higher Education*, Vol. 32 Issue 4, p387-401

TRECHERA, JOSÉ LUIS (1997). *El trastorno narcisista de la personalidad: concepto, medida y cambio.* Córdoba: Publicaciones ETEA.

TROCHIM, WILLIAM M. *The Research Method Knowledge Base*, 2nd Edition, http://www.socialresearchmethods.net/kb/ (version current as of October 20, 2006) (consultado 9, 05, 2009).

WANOUS, JOHN P.; REICHERS, ARNON E and HUDY, MICHAEL, J. (1997). Overall Job Satisfaction: How Good Are Single-Item Measures? *Journal of Applied Psychology* Vol. 82, No. 2, 247-252 http://www.bath.ac.uk/soc-pol/wam-net/Launch-mini-conference/WAMSEM6/WANOUS%20et%20al.%20Overall%20job%20sat.pdf

WENG, LI-JEN (2004). Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability. *Educational and Psychological Measurement*, 64, 6, 956-972.

WILDING, JOHN and ANDREWS, BERNICE (2006). Life goals, approaches to study and performance in an undergraduate cohort. *British Journal of Educational Psychology*, Mar2006, Vol. 76 Issue 1, p171-182

WUENSCH, KARL L. (2006) *Research Design Lessons, scaling,* http://core.ecu.edu/psyc/wuenschk/docs2210/Research-5-Scaling.doc (consultado 3, Dic. 2009).

YORKE, MANTZ (2009). 'Student experience' surveys: some methodological considerations and an empirical investigation. *Assessment & Evaluation in Higher Education*, vol. 34, nº 6, 721-739.

ZURDO GARAY – GORDOVIL, MARÍA MERCEDES (2011) *Determinantes emocionales y cognitivos de la conducta de apoyo materna. Estudio comparativo de madres con hijos que presentan o no problemas de rendimiento escolar.* Tesis doctoral. Madrid: Universidad Pontificia Comillas.

14.2. *Bibliografía sobre construcción de instrumentos*

DEVELLIS, ROBERT (1991). *Scale Development, Theory and Applications*. Newbury Park: Sage.

EDWARDS, A.L., (1957a). *Techniques of Attitude Scale Construction*. New York: Appleton-Century-Crofts.

GABLE, ROBERT K. and WOLF, MARIAN B. (1986). *Instrument Development in the Affective Domain.* Boston/Dordrecht/Lancaster: Kluwer-Nijhoff Publishing.

HENERSON, M.E., MORRIS, L.L. and FIZT-GIBBON, C.T**.** (1978). *How to Measure Attitudes*, Beverly Hills: Sage.

KING, M. AND ZIEGLER, M. (1975). *Research Projects in Social Psychology*. Monterrey: Brooks-Cole.

KLINE, P. (1986). *A Handbook of Test Construction.* New York: Methuen.

LIKERT, R. (1932). A Technique for the Measurement of Attitudes, *Archives of Psychology*, 140, 44-53 [en español en WAINERMAN, C.H. (Ed.), (1976). *Escalas de medición en las ciencias sociales.* Buenos Aires: Nueva Visión, 199-260 y en SUMMERS, GENE F. (Ed.) (1976). *Medición de actitudes*. México: Trillas, 182-193].

MORALES VALLEJO, PEDRO (2011). *Análisis factorial en la construcción e interpretación de tests, escalas y cuestionarios.* http://www.upcomillas.es/personal/peter/investigacion/AnalisisFactorial.pdf

MORALES VALLEJO, PEDRO (2011). *Evaluación de los valores: análisis de listas de ordenamiento* http://www.upcomillas.es/personal/peter/otrosdocumentos/ValoresMetodo.pdf

MORALES VALLEJO, PEDRO (2006). *Medición de actitudes en Psicología y Educación, construcción de escalas y problemas metodológicos,* tercera edición revisada. Madrid: Universidad Pontificia Comillas.

MORALES VALLEJO, PEDRO; UROSA SANZ, BELÉN y BLANCO BLANCO, ÁNGELES (2003). *Construcción de escalas de actitudes tipo Likert. Una guía práctica*. Madrid: La Muralla.

MORRIS, LYNN LYONS, FIZT-GIBBON, CAROL TAYLOR, and LINDHEIM, ELAINE (1987). *How to measure attitudes.* Newbury Park & London: Sage.

NUNNALLY, JUM C. (1978). *Psychometric Theory*. New York: McGraw-Hill.

SPECTOR, PAUL E. (1992). *Summating Ratings Scale Construction: An Introduction.* Newbury Park & London: Sage.

WAINER, HOWARD and MESSICK, SAMUEL (Eds.) (1983). *Principals of Psychological Measurement*. Hillsdale, N. J.: Lawrence Erlbraun.

11.3. *Colecciones de instrumentos*

BEARDEN, WILLIAM O.; NETEMEYER, RICHARD G. and MOBLEY, MARY E. (1993). *Handbook of Marketing Scales.* Newbury Park: Sage.

BORICH, G.D., and MADDEN, S.K. (1977). *Evaluating Classroom Instruction, A Sourcebook of Instruments.* Reading, Mass.: Addison-Wesley.

COHEN, L. (1976). *Educational Research in Classrooms and Schools.* London: Harper & Row.

FISHER, JOEL and CORCORAN, KEVIN J. (1994). *Measures for Clinical Practice: A Sourcebook; Couples, Families and Children.* Portland: Portland State University, Oregon.

HILL, PETER C. and HOOD JR., RALPH W. (1999). *Measures of Religiosity*. Birmingham, Alabama: Religious Education Press.

LAKE, D.G., MILES, M.B. and EARLE JR., R.B. (1973). *Measuring Human Behavior*. New York: Teachers College, Columbia University

MILLER, D.C. (1977). *Handbook of Research Design.* New York: David McKay.

MORALES VALLEJO, PEDRO (2011). *Cuestionarios y escalas*. http://www.upcomillas.es/personal/peter/otrosdocumentos/CuestionariosyEscalas.pdf

NEWMARK, CHARLES S. (Ed.) (1996). *Major Psychological Assessment Instruments.* Second Edition. Boston: Allyn & Bacon.

ROBINSON, JOHN P. and SHAVER, PHILLIP R. (1980). *Measures of Social Psychological Attitudes.* Ann Arbor, Mich.: Institute of Social Research, the University of Michigan.

ROBINSON, JOHN P.; SHAVER, PHILLIP R. and WRIGHTSMAN, LAWRENCE S. (Eds.) (1991). *Measures of Personality and Social Psychological Attitudes.* New York: Academic Press.

ROBINSON, JOHN P.; SHAVER, PHILLIP R. and WRIGHTSMAN, LAWRENCE S. (Eds.) (1999). *Measures of Political Attitudes.* New York: Academic Press.

SCHUESSLER, K.F. (1982). *Measuring Social Life Feelings.* San Francisco: Jossey-Bass

SHAW, M.E. and WRIGHT, J.M. (1967). *Scales for the Measurement of Attitudes.* New York: McGraw-Hill.

STRAUSS, M.A. and BROWN, B.W. (1978). *Family Measurement Techniques, Abstracts of Published Instruments, 1935-1974*. Minneapolis: University of Minnesota Press.