CAPÍTULO 2. ELABORACIÓN DE UNA ESCALA PARA MEDIR LA CALIDAD DE ESTUDIOS PRIMARIOS

CAPÍTULO 2. ELABORACIÓN DE UNA ESCALA PARA MEDIR LA CALIDAD DE ESTUDIOS PRIMARIOS.

1. INTRODUCCIÓN.

Tal como se concluyó en el capítulo anterior puede decirse que, en general, la evaluación de la formación continua no presenta estudios de validez y fiabilidad; generalmente, consiste en la evaluación de los resultados, principalmente al nivel más inicial de los propuestos por Kirkpatrick (1999), referido al estudio de la satisfacción; y suele basarse en modelos poco operativos y metodología con un bajo nivel de estandarización. Estas conclusiones son acordes a las halladas por otros autores que han estudiado la evaluación de la formación en distintos contextos tales como, por ejemplo, Trevisan (2004) en el ámbito de la formación de adultos, Follete y Beitz (2003) en el contexto de formación en psicología clínica y Moresky, Eliades, Bhimani, Bunney y VanRooyen (2001) en la formación para trabajadores de organizaciones no gubernamentales.

Teniendo en cuenta este punto de partida, el **objetivo** en este capítulo fue determinar las variables a considerar para el estudio de la **calidad** de un programa, dando especial relevancia a las cuestiones metodológicas, pero también considerando algunas sustantivas relativas a nuestro ámbito de intervención de acuerdo con la literatura al uso.

Se elaboró una **escala** para medir la calidad de los estudios primarios que actualmente no es considerada como un instrumento cerrado, sino en fase de prueba. La utilidad de esta escala es la posibilidad de ser aplicada en un estudio específico o en un conjunto de estudios de temática común, y la obtención de datos empíricos que ayuden a detectar puntos débiles y a proponer posibles mejoras para aumentar su calidad metodológica. En el **capítulo 3** se aplicará la escala resultante a los programas de formación continua obtenidos tras una búsqueda bibliográfica. Se intenta de este modo obtener una evidencia más para corroborar o rechazar algunos aspectos encontrados en la revisión de modelos como la falta de estandarización y operacionalización en los estudios en este ámbito, intentando detectar otras posibles flaquezas y proponiendo posteriormente posibles mejoras al respecto.

El hecho de que los programas posean un diseño y evaluación de calidad metodológica es de interés pues es un modo de dar credibilidad a los resultados obtenidos ya que, en principio, podría partirse de la hipótesis de que los sesgos de los estudios disminuirán al aumentar la calidad de su diseño.

A este respecto, habría que acabar con el mito de que todo diseño experimental es de calidad y otros diseños carecen de ella porque no todo estudio tiene por objetivo establecer una relación causal y, a su vez, no todo diseño aleatorio tiene necesariamente que ser mejor que los no aleatorios para aportar evidencias empíricas de inferencias causales válidas; así, posiblemente, un diseño no aleatorio con un control de covariante fiable aporte resultados más válidos que un diseño experimental con mortalidad diferencial (Shadish y Myers, 2004; Shadish y Ragsdale, 1996), lo que a su vez lleva implícito que calidad metodológica se suele ligar a efecto causal.

La calidad del diseño de los programas es un concepto complejo y multidisciplinar. En la literatura, existen distintas aproximaciones al estudio de esta calidad; su operacionalización depende del constructo "calidad de intervención" que se tome como referente (validez interna, validez externa, calidad del informe, adecuación de análisis estadísticos realizados, implicaciones éticas, repercusiones para el área de intervención, etc.). Básicamente, existen dos planteamientos para medir la calidad sistemáticamente: estudiar un único componente de la intervención como indicador de calidad (por ejemplo, tipo de asignación de usuarios, tamaño de la muestra, etc.) u obtener un índice global de calidad a partir de las puntuaciones asignadas a una serie de ítems ponderados sobre la calidad de cada intervención particular (Chacón, 2004; Chacón, Sánchez-Meca, Sanduvete y Alarcón, en elaboración).

En este caso, se ha optado por estudiar la calidad mediante la opción del índice global por los siguientes motivos principales:

- Si se mide adecuadamente, permite el estudio de sus **índices métricos** (índices de fiabilidad y validez).

- Es el instrumento de medición **más utilizado**, por lo que resulta fácil encontrar ítems de diversas fuentes a partir de los cuales crear una nueva escala (Jüni, Altman y Egger, 2001).

- Posibilita no sólo el estudio de la calidad a partir de un **índice global cuantitativo** de fácil interpretación sino que, además, aporta información detallada de distintos aspectos a partir de los diferentes ítems. En este trabajo en concreto, interesa más esta última posibilidad, pues se pretende buscar variables (de los aspectos del diseño y la evaluación) que modulan los resultados de los programas, aunque también se calculará el índice global y se estudiará si existe alguna relación entre éste y otras variables de interés.

Una posible duda que puede surgir es que, si las escalas son los instrumentos más utilizados para medir la calidad de los estudios primarios, ¿qué necesidad hay de elaborar una más? Se intentarán aportar las siguientes novedades con respecto a lo ya existente:

- **Sistematizar indicadores útiles**: no sólo se pretende calcular un índice general de calidad; también se pretende obtener información con cada uno de los ítems acerca de cómo mejorar los programas, principalmente desde un punto de vista metodológico.

- Lograr medir con dicha escala estudios primarios con **cualquier tipo de diseño**. De este modo, se intenta cubrir la falta de atención en los diseños no experimentales que comúnmente se da en la literatura.

A continuación, se pasa a describir el proceso de elaboración de la escala. En primer lugar, se recogieron distintos ítems que posteriormente pasarían a formar parte de un cuestionario exploratorio de medición de la calidad en estudios primarios; además, se realizó un estudio exploratorio de la calidad en el diseño de programas psicológicos, sociales y de la educación con los ítems considerados relevantes; en

segundo lugar, se estudió la validez de contenido de dicho cuestionario; a partir de los resultados obtenidos, en tercer lugar se depuraron estos ítems y volvió a hacerse un estudio exploratorio, esta vez con los programas referidos a la formación continua; finalmente, en base al compendio de todos los resultados previos, se obtuvo una versión integradora de la escala, tratando de lograr un instrumento que midiera de manera homóloga la calidad de los estudios primarios independientemente del diseño que presentaran y una mayor concreción y operatividad de sus ítems. Este proceso deductivo-inductivo aún no ha llegado a su fin. El último paso hasta ahora dado ha sido el estudio exploratorio que se presentará en el próximo capítulo.

2. PRIMERA FASE: COMPILACIÓN DE ÍTEMS EXISTENTES EN LA LITERATURA SOBRE CALIDAD EN ESTUDIOS PRIMARIOS Y APLICACIÓN EXPLORATORIA.

La primera fase de elaboración de la escala para medir la calidad de los estudios primarios obtuvo como resultado la recopilación exhaustiva de todos los ítems utilizados para tal fin en la literatura. Posteriormente, se realizó una aplicación exploratoria a estudios de ámbitos social, educacional y psicológico con los ítems considerados más relevantes para ir analizando el funcionamiento de éstos y para obtener una aproximación de las características que generalmente presentaban las intervenciones en estos contextos de actuación. A continuación, se detallan el método utilizado y los resultados obtenidos en esta primera fase.

2.1. Compilación de ítems sobre calidad en estudios primarios.

2.1.1. Método.

2.1.1.1. Muestra.

Se recogieron los **artículos** referidos a la medición de la calidad de los estudios publicados, disponibles en las 11 bases de datos informatizadas a las que se tuvo acceso a través de la Universidad de Sevilla y que resultaron de interés, con la finalidad de detectar aquellos ítems que usualmente se consideraban relevantes. Concretamente, las bases de datos utilizadas fueron EBSCO Online, Medline, Serfila, CABHealth, CINAHL, Econlit, MathSci, Current Contents, Humanities Index, ERIC y PsycINFO. Otros artículos relacionados con la temática fueron obtenidos gracias a la colaboración de la Unidad de Meta-análisis de la Universidad de Murcia.

Concretamente, los **artículos** estudiados fueron los siguientes **27**: Sánchez-Meca y Ato (1989); O'Rourke y Detsky (1989); Weisz, Hawley, Pilkonis, Woody y Follette (2000); Tritchler (1999); Jüni, Altman y Egger (2001); Sánchez-Meca (1997); Sutton, Abrams, Jones, Sheldon y Sing (2000); Moher, Jones y Lepage (2001); Moher, Schulz y Altman (2001); Begg, Cho, Eastwood, Horton, Moher, Olkin, Pitkin, Rennie, Schulz, Simel y Stroup (1996); Moher, Jadad, Nichol, Penman, Tugwell y Walsh (1995); Moher, Pham, Jones, Cook, Jadad, Moher, Tugwell y Klassen (1998); Des Jarlais, Lyles, Crepaz y "the TREND group" (2004); Bossuyt, Reitsma, Bruns, Gatsonis, Glasziou, Irwig, Lijmer, Moher, Rennie y de Vet (2003); Bossuyt, Reitsma, Bruns,

Gatsonis, Glasziou, Irwig, Moher, Rennie, de Vet, y Lijmer (2003); Olivares, Rosa y Sánchez-Meca (2000); Education Group for Guidelines on Evaluation (1999); Campbell, Elbourne y Altman (2004); Bosch, Guardiola y grupo de trabajo Foundation Workshop 2002 (2003); Altman, Schulz, Moher, Egger, Davidoff, Elbourne, Gotzsche y Lang (2001); Brown (1991); Jüni, Witschi, Bloch y Egger (1999); McGuire, Bates, Dretzke, McGivern, Rembold, Seabold, Turpin y Levin (1985); Emerson, Burdick, Hoaglin, Mosteller y Chalmers (1990); Moher, Jadad y Tugwell (1996); MacPherson, White, Commings, Jobst, Rose y Niemzow (2002); y Greenland (1994).

2.1.1.2. Instrumentos.

Para la recopilación definitiva de dichos artículos, se utilizó el software **Procite-5** con el que, entre otras funciones, se pudo eliminar del listado de resultados encontrados en las distintas bases de datos aquellos que ya estaban incluidos, con lo que se evitaban las repeticiones.

2.1.1.3. Procedimiento.

Los pasos para la compilación de los ítems de medición de la calidad de estudios primarios fueron tres:

1º. Búsqueda de artículos: en primer lugar, se llevó a cabo una revisión bibliográfica y se recopilaron todos los artículos encontrados que hicieran referencia a la medición de la calidad de estudios primarios (el listado de artículos encontrados se presenta en el apartado referido a la muestra). Las **palabras clave** introducidas para la realización de dicha búsqueda fueron "quality" (calidad) y "primary studies" (estudios primarios).

2º. **Recogida de los ítems concretos:** de los artículos previamente recogidos, tomamos todos los ítems encontrados que se habían elaborado con la idea de medir la calidad. No se trató de que fueran mutuamente excluyentes sino que fueran exhaustivos por lo que, aunque algunos pueden parecer redundantes, a partir de los datos se dio la posibilidad de depurarlos a posteriori y así definir los constructos con mayor precisión. En el anexo III (pág. xix) se relaciona la procedencia de los ítems incluidos en el cuestionario exploratorio.

3º. Estructuración en dominios y subdominios: tomando como base la estructuración propuesta por Sánchez-Meca (1997) en relación a las variables moderadoras de un meta-análisis, todos los ítems fueron organizados en diferentes dominios y subdominios.

2.1.2. Resultados.

Se compilaron todos los ítems encontrados en la documentación recopilada que hacían referencia a estudios de calidad del diseño. La mayoría de estos ítems provinieron de las escalas de medición de la calidad de los estudios primarios que más frecuentemente aparecen en la literatura (Chacón, Sánchez-Meca, Sanduvete y Alarcón, en elaboración). Concretamente son: a) CONSORT (Consolidated Standards of Reporting Randomized Trials) (Moher, Schulz, y Altman, 2001); b) TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) (Des Jarlais, Lyles, Crepaz y "the TREND Group", 2004); c) STRICTA (Standards for Reporting

Interventions in Controlled Trials of Acupuncture) (MacPerson, White, Commings, Jobst, Rose y Niemzow, 2002); d) STARD (Standards for Reporting of Diagnostic Accuracy) (Bossuyt, Reitsma, Bruns, Gatsonis, Glasziou, Irwig, Moher, Rennie, de Vet y Lijmer, 2003); y e) Guía para evaluar los artículos referentes a la intervención en educación (Education Group for Guidelines on Evaluation, 1999). En la tabla 2.1 se presentan las principales características de cada una de estas escalas:

	CONSORT	**TREND**	**STRICTA**	**STARD**	**EDUCACIÓN**
Año de publicación	1996 (revisado en 2001)	2004	2002	2003	1999
Número de ítems	22	22	6	25	17
Adecuado para medir…	Estudios aleatorios	Estudios no aleatorios	Acupuntura	Diagnóstico médico	Intervenciones en educación

Tabla 2.1. Características de las escalas de medición de la calidad más comunes.

En definitiva, 43 ítems fueron compilados, todos aquellos que se consideraron útiles para la medida de la calidad del diseño (se eliminaron aquellos que hacían referencia exclusivamente a la calidad del reporte). Se pretendió ser exhaustivos, sin analizar el posible solapamiento parcial entre los ítems encontrados. En el anexo II (pág. vii), se muestra el listado de ítems resultante, tanto en español como en inglés; en el anexo III (pág. xix), se especifica la procedencia de cada uno de estos 43 ítems. Tal y como puede observarse, los ítems incluidos en el cuestionario de validez de contenido sobre la escala para medir la calidad de los estudios primarios fueron divididos en tres dominios (Sánchez-Meca, 1997):

- Características **extrínsecas**, referidas al contexto en el que se escribió cada estudio.

- Características **sustantivas**, referidas al contenido de cada intervención. Se distinguieron tres subdominios:

    - **Muestra**: descripción de las personas participantes.

    - **Contexto**: aquello que queda ajeno al núcleo de intervención pero que podría influir en los resultados obtenidos.

    - **Tratamiento**: la intervención en sí y el modo en que se realizó el estudio.

- Características **metodológicas**, referidas a aspectos tales como el diseño y el análisis de datos.

Con estos ítems, en la segunda fase se realizó el estudio de validez de contenido que se describe en el próximo apartado.

2.2. Aplicación exploratoria.

2.2.1. Método.

2.2.1.1. Muestra.

La búsqueda se realizó en las **bases de datos** informatizadas disponibles en la Universidad de Sevilla, de las que se extrajeron los artículos que interesaban.

Se compilaron **2087** resúmenes de los que se estudiaron finalmente **1899** tras eliminar aquellos trabajos en los que faltaban datos, cuya intervención se llevaba a cabo con sujetos no humanos y/o que eran la replicación de otro estudio previamente incluido.

2.2.1.2. Instrumentos.

El software "**Procite-5**" se utilizó para el tratamiento de la información de los resúmenes; el programa estadístico "**SPSS 12.0**" para la codificación de los datos y su posterior análisis; y, finalmente, un **sistema de categorías** compuesto por 19 ítems, elegidos de entre todos los encontrados por su utilidad a la hora de determinar la calidad de los estudios primarios (Chacón, Sánchez-Meca, Alarcón, & Marín, 2002; Chacón, García, Alarcón y Sanduvete, 2003).

2.2.1.3. Procedimiento.

Se realizó una búsqueda en todas las bases de datos a las que se tenía acceso en la Universidad de Sevilla, que guardaran relación con el tema a estudiar (se excluyeron aquellas específicas sobre algunos temas que no interesaban como, por ejemplo, de economía). Se recogieron artículos publicados hasta julio de 2005.

Las **palabras clave** utilizadas fueron random (aleatorio), non-random (no aleatorio), effect size (tamaño de efecto), quasi-experimental (cuasi-experimental), experimental, meta-analysis (meta-análisis), intervention program (programa de intervención), evaluation (evaluación), assessment (detección), social y education (educación). Se introdujeron en inglés porque las bases de datos utilizadas estaban en este idioma. Las razones por las que se tomaron estas palabras clave y no otras fueron las siguientes:

- Por un lado, se pretendía encontrar intervenciones que presentaran cualquier tipo de diseño:

  - Diseños experimentales, caracterizados porque, además de que el experimentador manipula alguna/s variable/s independiente/s y pretende encontrar efecto en una variable dependiente, los sujetos se asignan aleatoriamente a las distintas condiciones experimentales. Para encontrar este tipo de diseño, se usaron las palabras "aleatorio" y "experimental".

  - Diseños cuasi-experimentales, que se diferencian de los anteriores principalmente en que los sujetos no son asignados aleatoriamente a las distintas condiciones del tratamiento. Es por ello que se tomaron como palabras clave "no aleatorio" y "quasi-experimental".

- Se intentaban detectar intervenciones planificadas, sistematizadas, temporalizadas y con recursos humanos y materiales especificados; es decir, que mostraran cierto control; que siguieran aquellas pautas de carácter científico que permiten la posibilidad de que estos programas sean evaluados de manera fidedigna; en

definitiva, se intentaban detectar intervenciones que presentaran "evaluabilidad". Para encontrar intervenciones con estas características, se utilizaron las palabras "programa de intervención", "evaluación" y "detección".

- Se intentaba encontrar estudios en los que se explicitaran los resultados encontrados con la suficiente precisión como para que, si ello fuera de interés, se pudieran incluir en un estudio meta-analítico. Por ello, se tomaron las palabras clave "tamaño de efecto" y "meta-análisis".

- Finalmente, se pretendía encontrar intervenciones en distintos contextos; en diferentes ámbitos. Es por ello que se incluyeron, entre las palabras clave utilizadas, "social" (con la intención de encontrar programas en el contexto comunitario) y "educacional" (en busca de intervenciones en el ámbito educativo).

De los ítems recopilados respecto a medición de la calidad en estudios primarios, se eligieron aquellos considerados de mayor interés en función de su utilidad y se incluyeron como sistema de categorías para codificar los estudios encontrados.

Tres codificadores independientes categorizaron dichos estudios, obteniéndose una fiabilidad intercodificadores adecuada, con un índice de correlación intraclase de 0.85.

2.2.2. Resultados.

Respecto al **análisis bibliográfico** exploratorio referido a la calidad que suelen presentar los estudios primarios en los ámbitos psicológico, social y educacional, los resultados se muestran en el anexo IV (pág. xxiii) (Chacón, García, Alarcón y Sanduvete, 2003).

A grandes rasgos, puede decirse que los estudios mostraron dos características:

- Un **bajo grado de especificación.** Así, la orientación teórica no fue explicitada en el 67.2% de las ocasiones; la edad no fue especificada en el 82.2% de los casos; y no solió presentarse tamaño de efecto ni datos suficientes para calcularlo.

- Un **grado medio de control y estandarización**: en el 63.4% de las ocasiones, no hubo aleatorización en la asignación de las personas a los grupos, pero sí se dio control de alguna variable extraña; la mayoría de los estudios presentaron un diseño pre-experimental o cuasiexperimental; en la mayoría de las ocasiones, sólo se daba medida posterior; en la mayoría de los casos, sólo algunas variables eran medidas en todos los momentos; al menos una variable dependiente fue semiestándar en el 71.9% de los casos; y se dieron usualmente técnicas de control, pero no el ciego.

Tras este análisis exploratorio, se pasó a tratar de determinar con más precisión y certeza los aspectos que, según los expertos en la materia, influyen en el nivel de calidad metodológica de un estudio.

3. SEGUNDA FASE: ESTUDIO DE VALIDEZ DE CONTENIDO.

Todos los ítems recogidos en la primera fase, una vez estructurados en las distintas dimensiones, sirvieron como cuestionario exploratorio del que se realizó un estudio de validez de contenido para tratar de determinar, según la opinión de expertos en la materia, qué ítems eran viables, útiles para medir calidad y representativos de su dimensión. A continuación, se describe el método utilizado y los resultados obtenidos en dicho estudio de validez de contenido.

3.1. Método.

3.1.1. Muestra.

La muestra para la realización del estudio de validez de contenido de la escala elaborada estuvo compuesta por **30 expertos** en cuestiones de diseño y evaluación de programas y de medición de la calidad (miembros del "Methods Group de la Campbell Collaboration" y miembros de la "Asociación Española de Metodología de las Ciencias del Comportamiento"). Concretamente, fueron 12 mujeres y 18 hombres, de proveniencia europea y norteamericana (el 50% de los participantes fueron españoles por cuestiones de accesibilidad). La media de edad fue de 47 años y 14 la media de años de experiencia en estas cuestiones.

3.1.2. Instrumentos.

Se utilizó un **cuestionario** que contenía los 43 ítems recopilados en la fase anterior (ver anexo II, pág vii). Se utilizó una escala de 3 puntos (siendo -1 la puntuación más baja; 0 la intermedia y 1, la más alta) para medir, en cada uno de los ítems, los tres constructos más frecuentemente utilizados en los estudios de validez de contenido (Chacón, Pérez, Holgado y Lara, 2001a):

1. **Representatividad**, que hacía referencia al grado en que cada ítem representaba el dominio al que había sido asignado.

2. **Utilidad**, referido a la medida en que cada ítem específico era útil para evaluar la calidad de los estudios con respecto al dominio donde fue asignado.

3. **Viabilidad del dato**, referido a si las circunstancias posibilitaban la existencia del dato y su recogida.

Los ítems fueron diferenciados en tres dimensiones; concretamente, 6 formaron parte de las características **extrínsecas**, 14 de las características **sustantivas** (5 respecto a la muestra, 3 en relación al contexto y 6 en el tratamiento) y 23 de las características metodológicas. En cada dominio y subdominio, además, se dejó un espacio por si algún participante quería especificar alguna posibilidad no incluida.

Otro instrumento utilizado fue **internet** para la distribución de los cuestionarios y su posterior recogida.

Por último, se utilizó el software **Microsoft Excel** para codificar y analizar los datos.

3.1.3. Procedimiento.

Se llevó a cabo un estudio de validez de contenido para determinar qué ítems eran considerados los más aptos para medir cada uno de los dominios referidos a la calidad. El procedimiento sistemático puede resumirse en tres pasos:

1. **Selección de la muestra:** Se obtuvo la colaboración de 30 personas expertas en Meta-análisis, evaluación de la calidad y diseño (una descripción más detallada se presentó en el apartado 3.1.1 referido a la muestra).

2. **Distribución del instrumento:** el cuestionario fue repartido a cada experto a través de correo electrónico o en persona, en diversos encuentros de divulgación científica, en tres ocasiones (Ritter y Sue, 2007); concretamente, la primera solicitud se realizó en el *V Annual Campbell Collaboration Colloquium* celebrado en Lisboa en febrero de 2005; la segunda, en el *IX Congreso de Metodología de las Ciencias Sociales y de la Salud*, celebrado en Granada, en septiembre de 2005; y la tercera, a través de correo electrónico a quienes aún no habían respondido.

3. **Análisis de datos:** una vez recogidas las respuestas, se aplicó el índice de Osterlind (1998) con la finalidad de determinar el grado de acuerdo entre expertos respecto a las puntuaciones otorgadas a cada uno de los ítems dentro de su dominio y respecto a los tres constructos: representatividad, utilidad y viabilidad del dato. Concretamente, la fórmula del índice de Osterlind es la siguiente:

$$I_{ik} = \frac{(N-1)\sum_{j=1}^{n} X_{ijk} + N\sum_{j=1}^{n} X_{ijk} - \sum_{j=1}^{n} X_{ijk}}{2(N-1)n}$$

Siendo:

- N = número de dominios.
- X_{ijk} = el valor que cada juez otorga a cada ítem.
- n = número de jueces.

Este índice podía oscilar entre -1 y 1, ya que la escala utilizada fue de tres puntos (-1, 0 y 1), siguiendo las recomendaciones del autor. Obtener el valor extremo negativo en un ítem (-1), se interpretaría como que todos los expertos puntuaron a ese ítem con el valor más negativo; del mismo modo, el valor extremo positivo (1), implicaría que todos los expertos estuvieron de acuerdo en que el ítem descrito merecía en el aspecto estudiado la puntuación máxima. Siguiendo también las recomendaciones del autor, se consideraron valores aceptables aquellos que se encontraron por encima del 0.5 (Osterlind, 1998).

3.2. Resultados.

A continuación, en la tabla 2.2 se presenta cada ítem del cuestionario con sus respectivas puntuaciones obtenidas en función del índice de Osterlind (1998) en los tres aspectos estudiados: **representatividad (columna R), utilidad (U)** y **viabilidad del dato (V)**. En negrita se marcan los valores iguales o superiores a 0.5; con negrita y

cursiva se destacan aquellos ítems que obtuvieron una puntuación mayor a ésta en los tres aspectos.

CARACTERÍSTICAS EXTRÍNSECAS (N=30)	**R**	**U**	**V**
1. Tipo de publicación	0.429	**0.571**	**0.714**
2. Año de publicación	-0.071	0.214	**0.857**
3. Índice de impacto de la revista	-0.143	0.071	0.286
4. Base de datos en que se encontró	-0.214	0.357	0.357
5. Entrenamiento de los investigadores	0.071	**0.5**	0
6. Estructura recomendada por la APA	-0.143	0	0.143
CARACTERÍSTICAS SUSTANTIVAS (N=30)			
De la muestra			
7. Rango de edad	**0.571**	**0.5**	**0.571**
8. Media de edad	**0.786**	**0.786**	**0.714**
9. Desviación típica de la edad	0.429	0.143	0.429
10. Origen cultural	0.143	0.214	0.286
11. Nivel socioeconómico	-0.071	0.071	-0.286
Del contexto			
12. Contexto de intervención	-0.214	0.071	0
13. Campo de intervención	**0.5**	0.357	**0.857**
14. País	0.357	0.429	**0.714**
Del tratamiento			
15. Orientación teórica	0.286	**0.8**	0
16. Evidencia empírica previa	0.143	0.286	0.071
17. Periodo de tratamiento	**0.786**	**0.929**	**0.643**
18. Grado de intensidad del tratamiento	**0.786**	**0.929**	**0.786**
19. Unidades (en grupo o individual)	**1**	**0.929**	**0.929**
20. Los puntos fuertes y débiles son discutidos	0.429	0.143	0
CARACTERÍSTICAS METODOLÓGICAS (N=30)			
21. Criterios de inclusión y exclusión de las unidades de la muestra especificados	**0.643**	**0.929**	**0.5**
22. Asignación aleatoria de las unidades a los grupos	**0.929**	**1**	**0.643**
23. Tipo de metodología/ diseño	**0.857**	**0.929**	**0.643**
24. Tamaño de la muestra	**0.786**	**0.857**	**1**
25. Estadístico para calcular el tamaño de la muestra	0.429	**0.5**	0.286
26. Mortalidad experimental	**0.714**	**0.857**	0.143
27. Sin mortalidad	**0.571**	**0.5**	0.429
28. Mortalidad experimental entre grupos	**0.714**	**0.857**	0.071
29. Exclusiones posteriores a la asignación aleatoria	**0.643**	**0.643**	0.214
30. Periodo de línea base	0.071	0.214	0
31. Periodo de seguimiento	**0.571**	**0.714**	0.286
32. Momentos de medida	**0.929**	**0.929**	**1**
33. Las medidas del pre-test aparecen en el pos-test	**0.786**	**0.857**	0.429
34. Variables dependientes normalizadas	**0.571**	**0.643**	0.357
35. Homogeneidad de la intervención	**0.571**	0.357	-0.143
36. Técnicas de control	**0.714**	**0.857**	0.214
37. Definición del constructo	**0.857**	**0.714**	-0.071
38. Métodos estadísticos para inferir los valores perdidos	**0.643**	**0.571**	0.214
39. Especificación de los intervalos de confianza en los análisis estadísticos	0.143	0.214	**0.5**
40. Tamaño de efecto y valor	**0.714**	**0.786**	**0.571**
41. Otros datos aparte de los objetivos marcados	0.143	0.214	0.357
42. Interpretación de los resultados	0.143	0.071	0.214
43. Interpretación de los sesgos de los resultados	0.429	0.214	0.071

Tabla 2.2. Índices de representatividad, utilidad y viabilidad del dato obtenidos en cada uno de los ítems de la escala de calidad depurada.

Como puede verse en la tabla 2.2, los ítems que obtuvieron valoraciones consideradas positivas (por encima de 0.5) en los tres aspectos medidos son los siguientes:

- Respecto a las **características sustantivas**, concretamente referidas a las de la **muestra** (de las personas participantes en la intervención), dos ítems fueron considerados representativos de la dimensión, útiles y viables:

    - El **rango de edad** (ítem 7, con valores de 0.57, 0.5 y 0.57 en representatividad, utilidad y viabilidad respectivamente.

    - La **media de edad (**ítem 8), con valores aún más positivos; concretamente, 0.79, 0.79 y 0.71 respectivamente.

- Aún haciendo referencia a las **características sustantivas**, pero esta vez centrándonos en el **tratamiento**, los ítems con valoraciones consideradas positivas en los tres aspectos estudiados fueron:

    - El **periodo de tratamiento** (ítem 17), con unas valoraciones de 0.79, 0.93 y 0.64 en representatividad, utilidad y viabilidad del dato respectivamente.

    - El **grado de intensidad del tratamiento** (ítem 18), con valores concretos de 0.79, 0.93 y 0.79.

    - La **unidad** (ítem 19), con valoraciones muy positivas; concretamente 1 en representatividad y 0.93 tanto en utilidad como en viabilidad del dato.

- Los ítems que a partir de ahora se muestran hacen referencia a **características metodológicas**:

    - **Criterios de inclusión y exclusión de las unidades de la muestra especificados** (ítem 21). Sus valores obtenidos fueron 0.64 en representatividad, 0.92 en utilidad y 0.5 en viabilidad del dato. Por tanto, aunque los valores de los tres aspectos estudiados coincidieron o estuvieron por encima del valor considerado crítico (0.5), el que destacó fue el de utilidad.

    - **Asignación aleatoria de las unidades a los grupos** (ítem 22). Este ítem obtuvo valores muy positivos en representatividad de su dimensión y utilidad (0.93 y 1 respectivamente) y un valor no tan positivo, aunque sí por encima de 0.5 en viabilidad del dato (0.64, concretamente).

    - **Tipo de metodología/ diseño** (ítem 23). Siguiendo la línea del ítem anterior, éste obtuvo valoraciones positivas en los tres aspectos recogidos, pero especialmente en los referidos a representatividad y utilidad; concretamente, obtuvo 0.86 en representatividad, 0.93 en utilidad y 0.64 en viabilidad del dato.

- **Tamaño de la muestra** (ítem 24). Obtuvo altas puntuaciones en representatividad y utilidad (0.79 y 0.86 respectivamente). En viabilidad del dato, obtuvo la máxima puntuación (1).

- **Momentos de medida** (ítem 32). Este ítem obtuvo valoraciones muy positivas en los tres aspectos estudiados (0.93 en representatividad y utilidad y 1 en viabilidad del dato).

- **Tamaño de efecto y valor** (ítem 40). Este ítem obtuvo 0.71 en representatividad, 0.79 en utilidad y 0.57 en viabilidad del dato.

- Numerosos ítems obtuvieron valoraciones consideradas positivas en dos de los tres aspectos estudiados. A continuación, se presentan aquéllos que obtuvieron valores por encima de 0.5 en representatividad y utilidad, pero no en viabilidad del dato, todos ellos dentro del dominio de **características metodológicas**:

  - **Mortalidad experimental** (ítem 26). Este ítem obtuvo unas valoraciones muy positivas en representatividad y utilidad (0.71 y 0.86 respectivamente) y, sin embargo, obtuvo una puntuación muy baja en viabilidad del dato (0.14).

  - **Sin mortalidad** (ítem 27). Este ítem obtuvo índices superiores a 0.5, aunque no muy elevados; concretamente, recibió 0.57 en representatividad y 0.5 en utilidad.

  - **Mortalidad experimental entre grupos** (ítem 28). Sus índices, bastante elevados tanto en representatividad como en utilidad, fueron 0.71 y 0.86.

  - **Exclusiones posteriores a la asignación aleatoria** (ítem 29). Los índices obtenidos fueron 0.64 tanto en representatividad como en utilidad.

  - **Periodo de seguimiento** (ítem 31). Este ítem obtuvo índices considerados como positivos tanto en representatividad como en utilidad, siendo el valor en éste último aspecto algo más alto (concretamente, obtuvo 0.57 y 0.71 respectivamente).

  - **Las medidas del pre-test aparecen en el post-test** (ítem 33). Obtuvo índices muy positivos tanto en representatividad como en utilidad (concretamente, 0.79 y 0.86 respectivamente).

  - **Variables dependientes normalizadas** (ítem 34). Los expertos puntuaron este aspecto como representativo de su dimensión y útil, aunque no obtuvo índices muy elevados (0.57 y 0.64 respectivamente).

  - **Técnicas de control** (ítem 36). Este aspecto obtuvo índices bastante positivos tanto en representatividad (0.71) como en utilidad (0.86).

  - **Definición del constructo** (ítem 37). Este ítem fue considerado por los expertos como sumamente representativo (0.86) y muy útil (0.71). Llama la

atención el hecho de que su viabilidad fue considerada tan baja que el índice alcanzó un valor negativo (concretamente, -0.07).

- **Métodos estadísticos para inferir los valores perdidos** (ítem 38). Aun no obteniendo valoraciones muy elevadas, este ítem fue considerado tanto representativo (obteniendo un índice de 0.64) como útil (0.57).

- Un solo ítem, concretamente el número 1 dentro del dominio de **características extrínsecas**, referido al **tipo de publicación**, obtuvo valoraciones por encima del 0.5 en utilidad y viabilidad del dato (aproximadamente 0.57 y 0.71 respectivamente) y por debajo de este valor, aunque muy cercano a él, en representatividad (aproximadamente, 0.43).

- Por último, también un solo ítem, concretamente el referido al **campo de intervención** (ítem 13, dentro del subdominio de las **características sustantivas del contexto**) obtuvo índices con valoraciones iguales o por encima de 0.5 en representatividad (0.5) y viabilidad del dato (0.86), pero no en utilidad (0.36).

La información sobre estos resultados hasta ahora descrita aparece esquematizada de manera general en el anexo V (pág. xxvii), donde se muestran los ítems que obtuvieron índices iguales o mayores a 0.5 combinando de manera distinta los aspectos valorados; de esta forma, la línea A se refiere a aquellos que obtuvieron estas valoraciones en representatividad, utilidad y viabilidad del dato (R+U+V); la B a los que obtuvieron índices con estos valores en representatividad y utilidad, pero no en viabilidad (R+U); la C a los que los obtuvieron en utilidad y viabilidad (U+V); y la D a los que los obtuvieron en representatividad y viabilidad (R+V).

La tabla 2.3 que a continuación se presenta muestra el porcentaje de ítems que obtuvieron índices considerados adecuados respecto a los aspectos valorados, tanto de manera global (columna de "escala total"), como en función de los dominios/dimensiones a los que fueron asignados (características extrínsecas, sustantivas y metodológicas). La primera columna explicita las distintas combinaciones de los aspectos valorados: representatividad (R), utilidad (U) y/o viabilidad del dato (V). Entre paréntesis se muestra la frecuencia (el número de ítems) correspondiente a cada porcentaje.

ASPECTOS	CARACT. EXTRÍNSECAS (6)	CARACT. SUSTANTIVAS (14)	CARACT. METODOLOGICAS (23)	ESCALA TOTAL (43)
R, U, V	0% (0)	35.71% (5)	26.9% (6)	25.58% (11)
R, U	0% (0)	35.71% (5)	69.59% (16)	48.84% (21)
U, V	16.67% (1)	35.71% (5)	26.09% (6)	27.91% (12)
R, V	0% (0)	42.86% (6)	26.09% (6)	27.91% (12)

Tabla 2.3. Porcentaje de ítems con índices iguales o superiores a 0.5.

Como puede observarse, los ítems incluidos en la dimensión "**características extrínsecas**" obtuvieron índices bastante bajos, ya que sólo uno superó el 0.5 en utilidad y viabilidad. Con esto se puede concluir que esta dimensión no fue valorada como importante para determinar la calidad del diseño. No existen diferencias en los datos independientemente de que se tengan en cuenta los tres aspectos (representatividad, utilidad y viabilidad) o sólo dos.

Aunque tampoco existen diferencias importantes en la distribución de frecuencia de ítems con índices mayores o iguales que 0.5 (en torno al 36%) entre los aspectos valorados en las **características sustantivas**, sí se podría decir que, al menos, fueron considerados, en su conjunto, más importantes para determinar la calidad de un diseño que las características extrínsecas.

Respecto a las **características metodológicas**, cabe destacar que diez ítems (lo que supone más del 25% del total y casi el 50% de los ítems referidos a este dominio) obtuvieron valoraciones positivas en representatividad y utilidad, pero no en viabilidad; concretamente, obtuvieron índices positivos el 69.59% de los ítems incluidos en esta dimensión. Si, sin embargo, se tiene en cuenta el aspecto de viabilidad del dato, el porcentaje de ítems con índices positivos desciende a un 26.09%. Esto nos lleva a deducir que, a pesar de que los aspectos metodológicos son considerados importantes para determinar la calidad del diseño de las intervenciones, no se suelen facilitar.

Esto es corroborado atendiendo a la última columna, donde se estudia el porcentaje de ítems con índices óptimos sin diferenciar en función del dominio en el que estén incluidos: el porcentaje más elevado se encuentra cuando se obvian los resultados obtenidos en viabilidad del dato y se atiende únicamente a la **representatividad** y la **utilidad**.

4. TERCERA FASE: ELABORACIÓN DE UNA VERSIÓN DEPURADA DE LA ESCALA Y APLICACIÓN EXPLORATORIA.

En esta tercera fase se llegó a la elaboración de una versión depurada de la escala para medir la calidad de estudios primarios tomando como base los resultados encontrados en el estudio de validez de contenido previamente descrito y otros ítems propuestos por expertos en medición de la calidad y meta-análisis (principalmente, por la Unidad de Meta-análisis de la Universidad de Murcia).

Por otro lado, se realizó un nuevo estudio exploratorio para probar el funcionamiento de dicha escala y conocer los resultados preliminares acerca de las características relacionadas con la calidad del diseño que suelen presentar los estudios primarios, esta vez en el ámbito concreto de la formación continua.

4.1. Elaboración de la escala (versión depurada).

4.1.1. Método.

4.1.1.1. Muestra.

Para la **elaboración de la versión depurada** de la escala para medir la calidad de los estudios primarios se utilizaron como muestra los 43 ítems que se presentaron en la fase anterior (ver anexo II, pág. vii) junto con los resultados obtenidos en el estudio de validez de contenido y otros ítems propuestos por expertos.

4.1.1.2. Instrumentos.

Se tomó, de los 43 ítems iniciales, los 23 que cumplieron el criterio de inclusión (al menos un valor de índice de Osterlind de 0.5 en dos de los tres conceptos estudiados). Hay que tener en cuenta que los ítems 26 y 27 ("mortalidad experimental" y "sin mortalidad") se recogieron en sólo uno (el nuevo ítem 6 "mortalidad global") y que el ítem 36 "técnicas de control" se desglosó en tres: los nuevos ítems 13, 14 y 15, referidos al "enmascaramiento del evaluador, del usuario y del profesional que realiza la intervención", respectivamente. Por otro lado, se incluyeron otros ítems propuestos por expertos a través del cuestionario de validez de contenido y en reuniones.

4.1.1.3. Procedimiento.

El procedimiento seguido fue el siguiente:

1. **Recopilación de ítems que cumplieron el criterio de inclusión**: en primer lugar, se recogieron aquellos ítems que en el estudio de validez de contenido obtuvieron un índice de un valor de 0.5 ó mayor en al menos dos de los tres conceptos estudiados (representatividad, utilidad y viabilidad).

2. **Recogida de otros ítems propuestos por expertos**: a continuación, se recogieron aquellos ítems que fueron propuestos por los expertos participantes en reuniones realizadas o a través de la prueba de validez de contenido (concretamente, en los apartados de respuesta abierta).

3. **Comparación**: finalmente, se tomaron como base los ítems resultantes del estudio de validez de contenido y se añadieron nuevos ítems propuestos por expertos; en ocasiones, en ítems ya presentes, se crearon nuevas categorías. Se diferenció además entre los ítems que medían la calidad del reporte (referidos a si el artículo contenía todos los apartados propuestos por la APA, si provenía de una revista con alto índice de impacto, etc.) y los que se dirigían a la calidad del estudio en sí, tomándose únicamente los incluidos en este último caso.

4.1.2. Resultados.

En la tabla 2.4 que a continuación se presenta, se muestran los 33 ítems que conformaron finalmente dicho instrumento con sus diferentes opciones (columna de la izquierda) y las razones por las que se incluyó (columna de la derecha).

1. Grupo control: 0-inactivo; 1-activo	Usado en estudios meta-analíticos (Sánchez –Meca, Rosa y Olivares, 1999)
2. Criterios de selección de la muestra (inclusión/exclusión) (Assignment criteria; en Shadish, Cook y Campbell, 2002b, p.323): 0-no especificados; 1-especificados	Estudio validez de contenido (ítem 21; R, U y V ≥ 0.5)
3. Azar: 0- pre-experimentales y cuasiexperimentales (sin control de vvee); 0.5-cuasiexprimentales con control de vvee (balanceo, bloqueo, estratificación); 1-experimentales (asignación aleatoria de las unidades a grupos)	Recomendación de expertos y estudio validez de contenido (ítem 22; R, U y V ≥ 0.5)
4. Tipo de metodología/ diseño: 0-pre-experimental; 0.5-cuasi-experimental; 0.75-serie temporal (obs pre ≥30 y post ≥30) y discontinuidad en regresión; 1-experimental-aleatorio	Usado en estudios meta-analíticos (Sánchez –Meca, Rosa y Olivares, 1999) y estudio

	validez de contenido (ítem 23; R, U y V ≥ 0.5)
5. Muestra: 0 - n<12; 0.5 - n=[12-40]; 1- n>40	Usado en estudios meta-analíticos (Sánchez –Meca, Rosa y Olivares, 1999) y estudio validez de contenido (ítem 24; R, U y V ≥ 0.5)
6. Mortalidad global: 0- ≥20%; 0.5 - %=]0-20[; 1- 0%	Usado en estudios meta-analíticos (Sánchez –Meca, Rosa y Olivares, 1999) y estudio validez de contenido (ítems 26 y 27; R y U ≥ 0.5)
7. Mortalidad diferencial: 0-≥20%; 0.5 - %=]0-20[; 1- 0%	Recomendación de expertos y estudio validez de contenido (ítem 28; R y U ≥ 0.5)
8. Exclusiones posteriores a la asignación aleatoria (post-assignment attrition; en Shadish, Cook y Campbell, 2002b, p.323): 0- ≥20%; 0.5 - %=]0-20[; 1- 0%	Estudio validez de contenido (ítem 29; R y U ≥ 0.5)
9. Seguimiento: 0 – no se da seguimiento; 0.3< 6 meses; 0.6 – meses = [6-11]; 1- ≥12 meses	Recomendación de expertos y estudio validez de contenido (ítem 31; R y U ≥ 0.5)
10. Momentos de medida: 0-posterior; 1- previo y posterior	Recomendación de expertos y estudio validez de contenido (ítem 32; R, U y V ≥ 0.5)
11. Las medidas del pre-test aparecen en el pos-test: 0- falta más de uno; 0.5- falta 1; 1- todas aparecen en todos los momentos	Recomendación de expertos y estudio validez de contenido (ítem 33; R y U ≥ 0.5)
12. Variables dependientes normalizadas: 0- sólo autoinformes sin estandarizar; 0.5-ninguno normalizado, pero al menos uno es cuestionario o autoinforme estandarizado; 1-al menos un instrumento es objetivo o normalizado (p.ej. registro fisiológico, pruebas baremadas)	Recomendación de expertos y estudio validez de contenido (ítem 34; R y U ≥ 0.5)
13. Enmascaramiento del evaluador: 0-no; 1-sí	Recomendación de expertos y estudio validez de contenido (ítem 36 desglosado; R y U ≥ 0.5)
14. Enmascaramiento del usuario: 0-no; 1-sí	Recomendación de expertos y estudio validez de contenido (ítem 36 desglosado; R y U ≥ 0.5)
15. Enmascaramiento del profesional que realiza la intervención: 0-no; 1-sí	Recomendación de expertos y estudio validez de contenido (ítem 36 desglosado; R y U ≥ 0.5)
16. Homogeneidad de la intervención: 0- no todos los sujetos han recibido la misma intensidad de intervención, duración y profesionales; 1- todos los sujetos han recibido la misma intensidad de intervención, duración y profesionales	Recomendación de expertos y estudio validez de contenido (ítem 35, no valorado)
17. Definición del constructo: 0-no; 0.5-no operativa; 1-operativa	Estudio validez de contenido (ítem 37; R y U ≥ 0.5)
18. Métodos estadísticos para inferir los valores perdidos: 0-no; 1-sí	Estudio validez de contenido (ítem 38; R y U ≥ 0.5)
19. Tamaño de efecto y valor: 0-no; 1-sí	Estudio validez de contenido (ítem 40; R, U y V ≥ 0.5)
20. Índice de calidad metodológica: suma de puntuaciones: 0-19	Usado en estudios meta-analíticos (Sánchez –Meca, Rosa y Olivares, 1999)
21. Índice estadístico calculado	Descriptivo, de interés para

	cálculo posible meta-análisis
22. Los resultados mostraron diferencias estadísticamente significativas (0-no; 1-sí)	Descriptivo, de interés para concluir acerca de la eficacia del programa
23. Índice de variabilidad facilitado	Descriptivo, de interés para posible meta-análisis
24. Número de participantes en cada grupo	Descriptivo, de interés para posible meta-análisis
25. Número de grupos en el estudio	Descriptivo, de interés para posible meta-análisis
26. Exclusiones tras medidas posteriores: 0-no; 1-sí	Descriptivo, de interés para posible meta-análisis
27. Rango de edad especificado: 0-no; 1-sí	Estudio validez de contenido (ítem 7; R, U y V ≥ 0.5)
28. Media de edad (valor concreto)	Usado en estudios meta-analíticos (Sánchez –Meca, Rosa y Olivares, 1999) y estudio validez de contenido (ítem 8; R, U y V ≥ 0.5)
29. Periodo de tratamiento: 0.5- ≤6 meses; 1->6 meses	Usado en estudios meta-analíticos (Sánchez –Meca, Rosa y Olivares, 1999) y estudio validez de contenido (ítem 17; R, U y V ≥ 0.5)
30. Intensidad del tratamiento: número de sesiones a la semana	Usado en estudios meta-analíticos (Sánchez –Meca, Rosa y Olivares, 1999) y estudio validez de contenido (ítem 18; R, U y V ≥ 0.5)
31. Unidades de intervención: 0-individual; 1-grupal	Usado en estudios meta-analíticos (Sánchez –Meca, Rosa y Olivares, 1999) y estudio validez de contenido (ítem 19; R, U y V ≥ 0.5)
32. Área formativa	Descriptivos, de interés para posible meta-análisis en el área concreta de la formación
33. Campo de intervención (destinatarios)	Usado en estudios meta-analíticos (Sánchez –Meca, Rosa y Olivares, 1999) y estudio validez de contenido (ítem 13; R y V ≥ 0.5)
34. Tipo de publicación (1-Revista; 2-Libro; 3-Tesis; 4-Congreso; 5-Otros	Estudio validez de contenido (ítem 1; U y V ≥ 0.5)

Tabla 2.4. Proveniencia de cada uno de los ítems que compusieron la versión depurada de la escala de calidad.

La razón preponderante de la inclusión de los ítems en esta versión depurada de la escala fue la obtención en el estudio de validez de contenido de índices considerados positivos en al menos dos de los tres aspectos estudiados (representatividad, utilidad y viabilidad) y recomendados por expertos en diversas reuniones convocadas para tratar de hacer más exhaustivos el conjunto de ítems recogidos. En muchos casos, además, eran ítems utilizados frecuentemente en estudios meta-analíticos (por ejemplo, en Sánchez Meca, Rosa y Olivares, 1999). Los ítems 1, 16, del 20 al 26 y el 32 se incluyeron por otras razones:

- El **ítem 1** se incluyó por haber sido utilizado frecuentemente en los estudios meta-analíticos publicados.

- El **ítem 16** que, a pesar de haber sido incluido en el estudio de validez de contenido (ver el ítem 35 de la tabla 2.2), sólo superó el valor 0.5 en representatividad, fue incluido por recomendación de expertos en la materia.

- El **ítem 20** hace referencia a un índice global de calidad que se obtendría sumando los ítems del 1 al 19, que son aquellos que realmente se consideran indicadores de calidad metodológica de un programa. Fue incluido por su extendido uso en estudios meta-analíticos como, por ejemplo, en Sánchez-Meca, Rosa y Olivares (1999).

- El resto de ítems (del 21 en adelante), no puntúa para la obtención del índice de calidad mencionado, al ser únicamente descriptivos. Algunos de ellos se han incluido a posteriori, sin haber pasado previamente por el estudio de validez, por ser considerados de interés a nivel metodológico y/o sustantivo y por tratarse de posibles variables moduladoras a la hora de realizar un meta-análisis; así:

- Los **ítems 21, 22** y **23** versan sobre los datos aportados y la significación estadística de los resultados.

- Los **ítems 24 y 25** hacen referencia al número de participantes en cada grupo y al número de grupos en los estudios.

- El **ítem 26** responde a si existen exclusiones después de haber recogido alguna medida posterior.

- El **ítem 32** se incluyó por ser considerado de interés en el área concreta de evaluación de programas de formación continua.

4.2. Estudio exploratorio.

4.2.1. Método.

4.2.1.1. Muestra.

Para la **realización del estudio bibliográfico** (Sanduvete, Chacón, Holgado, Gómez y Sánchez, 2006), se hizo una búsqueda de lo publicado acerca de programas de formación continua. De los **2379** estudios encontrados en las bases de datos utilizadas en la búsqueda, tras la lectura de los resúmenes, se incluyeron únicamente aquéllos que versaban sobre formación continua en las organizaciones, desechando por tanto los programas de educación familiar, la formación ocupacional (dirigida a desempleados), las acciones formativas en las que se trabajaba con cuidadores no formales (como pueden ser los familiares de las personas mayores con dependencia) o los cursos realizados por la universidad para sus estudiantes, por ejemplo. Cumplieron este criterio de inclusión **287** resúmenes de trabajos, de los que finalmente se estudiaron **95** por ser los únicos de los que se consiguió obtener el texto completo.

4.2.1.2. Instrumentos.

Los instrumentos utilizados fueron la escala resultante tras ser depurada; el software "Procite-5" para el tratamiento de la información obtenida; el programa estadístico "SPSS 12.0" para la codificación de los datos y su posterior análisis; y Microsoft Excel para el cálculo del índice de calidad (ítem 20).

4.2.1.3. Procedimiento.

Se realizó una búsqueda en todas las bases de datos a las que se tenía acceso en la Universidad de Sevilla, que guardaran relación con el tema a estudiar. Concretamente, fueron EBSCO Online, Medline, Serfile, CABHealth, CINAHL, PsycINFO, Econlit, ERIC, MathSci, Current Contents y Humanities Index. Se recogieron artículos publicados hasta septiembre de 2006.

Las **palabras clave** introducidas fueron tres de manera conjunta; es decir, en búsquedas avanzadas, unidas por el nexo "AND" para que los resultados obtenidos contuvieran los tres términos. La búsqueda se solicitó en todos los posibles campos: título, resumen, palabras clave, texto completo, etc. Concretamente, las palabras clave fueron:

- **Training programs** (programas de formación): se eligieron estas palabras porque la intención era encontrar los programas de formación publicados para posteriormente estudiar su calidad a través del instrumento construido para tales fines.

- **Evaluation** (evaluación): al incluir esta palabra se intentó restringir la búsqueda para encontrar aquellos programas que mostraran los resultados encontrados y, posiblemente, algunas características metodológicas de diseño.

- **Work** (trabajo): incluyendo esta palabra en la búsqueda se intentó restringir los estudios encontrados al área concreta de las organizaciones, disminuyendo así la probabilidad de encontrar acciones formativas dirigidas a personas fuera del ámbito laboral.

Estas palabras clave no sirvieron para obtener unos resultados muy específicos, encontrándose **2379** estudios. Se procedió a la lectura de todos los resúmenes para tratar de acotar y estudiar únicamente lo que interesara, aplicando para ello un **criterio de exclusión**: se tomó como requisito imprescindible que la acción fuera dirigida a empleados y empleadas de alguna organización; se excluyeron, por tanto, aquellos trabajos que versaban sobre acciones formativas no dirigidas a empleados o empleadas como podían ser la formación ocupacional y los programas dirigidos a estudiantes o a cuidadores informales. Tras seguir este criterio, el número de estudios a incluir se redujo considerablemente, quedando únicamente **287**.

Tras esta criba, se procedió a la búsqueda de los textos completos a través de internet. Se encontraron definitivamente **95** estudios, que fueron codificados con la escala para medir la calidad de los estudios primarios previamente elaborada por tres codificadores independientes, que obtuvieron un grado de acuerdo bajo, con un índice de correlación intraclase de 0.5.

4.2.2. Resultados.

Los resultados se presentan en el anexo VI (pág. xxxi) (Sanduvete, Chacón, Holgado, Gómez y Sánchez, 2006). Se constató en el ámbito de la formación continua las mismas conclusiones obtenidas para los estudios psicológicos con el primer estudio exploratorio realizado: suele darse **poca especificación** de las características metodológicas (sin especificación de los criterios de selección de la muestra, definición no operativa de los constructos, sin inferencia de valores perdidos, sin especificación del tamaño de efecto y rango de edad no especificado); y un **grado medio de control y estandarización** (uso de diseños pre-experimentales y cuasiexperimentales, mortalidad global y diferencial, exclusiones posteriores a la asignación aleatoria y tras tomar una medida, sin seguimiento, sólo medida posterior, variables dependientes semiestandarizadas, sin uso del enmascaramiento y corto periodo de intervención).

Pero quizá más interesante que estos resultados en sí, fue el hecho de que este estudio exploratorio permitió detectar algunos aspectos relacionados con el funcionamiento de la escala que habrían de ser cambiados. Concretamente, se constató lo siguiente:

- No todos los ítems eran aplicables a todos los tipos de diseño de intervención por lo que, en la práctica, sólo los diseños experimentales podían alcanzar la puntuación máxima (19) en el índice de calidad metodológica. Además, los diseños cuasiexperimentales podían alcanzar una puntuación máxima mayor que aquellos estudios que no presentaban intervención. Se consideró que este trato desigual a favor de los diseños experimentales y, en segundo lugar, de los cuasiexperimentales, impedía la posibilidad de comparar resultados entre los distintos diseños.

- Algunas categorías e ítems resultaban algo ambiguos por falta de definición operativa de los constructos de los que se trataban o por poca especificación a la hora de marcar los intervalos, con lo que podrían surgir dudas acerca de la categoría concreta en que encuadrar un estudio determinado.

Por ello, se hizo necesario afinar más y hacer modificaciones. A continuación, se presenta la última modificación realizada a la escala de medición de la calidad de los estudios primarios, con la que se trató de lograr dos objetivos (Chacón, Sanduvete, Sánchez y Sánchez-Meca, 2007b):

- Homogeneizar su uso para todos los diseños, de tal manera que cualquier estudio pudiera puntuar de 0 a 19, independientemente del tipo de diseño que presentara en lugar de que, por el hecho de ser un diseño experimental, partiera de ventajas en la máxima puntuación posible a alcanzar en comparación con los demás diseños.

  Se trató por tanto de adaptar los ítems a cualquier tipo de diseño, intentando así que el instrumento fuera flexible y que permitiera valorar de manera igualitaria la calidad de cualquier estudio, independientemente del diseño que presentara.

- Concretar, especificar y operativizar aquellos ítems o categorías de ítem que provocaron ambigüedades a los codificadores en la versión anterior.

5. DOBLE PROCESO DEDUCTIVO-INDUCTIVO: ELABORACIÓN DE UNA VERSIÓN INTEGRADORA DE LA ESCALA.

A continuación, se explica el método seguido para obtener la hasta ahora última versión de la escala.

5.1. Método.

5.1.1. Muestra.

La muestra utilizada fueron los 34 ítems que conformaron la escala de medición de la calidad en estudios primarios en su versión previa (tal y como se presenta en el apartado anterior).

5.1.2. Instrumentos.

Aparte de la escala previamente elaborada, no se utilizó ningún otro instrumento.

5.1.3. Procedimiento.

Se partió de los 34 ítems de los que constaba la escala en la versión previa. Tras ser aplicada en los estudios que versaban sobre formación continua, se prestó atención a cómo se comportaba cada ítem en los distintos estudios. Se detectaron:

- Aquellos ítems que mostraron diferencias en función del diseño que los estudios presentaban.

- Aquellos ítems que mostraron alguna dificultad de comprensión a la hora de decidir qué categoría escoger, una de las posibles causas por la que se obtuvo un bajo índice de acuerdo intercodificadores.

Una vez detectados estos ítems, se pasó a la realización de los cambios oportunos para lograr:

- La comparabilidad de los datos obtenidos en cada estudio, sin que el tipo de diseño presentado supusiera un aspecto diferenciador. Se logró que todos los ítems fueran aplicables a todos los tipos de diseño y que todos pudieran sumar en el índice de calidad de 0 a 19 (compuesto por 19 ítems, concretamente del 1 al 19, según la tabla 2.5 que a continuación se presenta).

- Claridad conceptual de los ítems y operacionalización para facilitar la elección de la categoría adecuada a la hora de la codificación.

5.2. Resultados.

A continuación, en la tabla 2.5, se presenta la escala resultante (Chacón, Sanduvete, Sánchez y Sánchez-Meca, 2007b). Además, para mayor claridad, en el anexo VII (pág. xxxvii) se muestran tres variantes de la misma escala: la primera adaptada a estudios observacionales y de encuesta (en aquellos casos donde no hay intervención, sino sólo recogida de la respuesta que da el sujeto); la segunda, a estudios cuasiexperimentales (donde sí hay intervención, pero la selección y/o la asignación a los distintos grupos no se hace al azar); y la última, a estudios experimentales (donde hay intervención y la selección y asignación se realizan aleatoriamente).

Para facilitar la interpretación, a continuación se explican las partes que componen la escala:

- La primera columna muestra el **ítem** que se está valorando.

- La segunda columna (**V**) presenta el **valor** que obtiene cada una de las opciones en cada ítem. Sólo los valores obtenidos en los ítems del **1 al 19** (en casillas sombreadas) son sumados para el cálculo del **índice global de calidad** (ítem 20); los demás, son únicamente descriptivos. Las valoraciones de estos 19 ítems oscilan entre 0 como menor puntuación y 1 como mayor por lo que, consecuentemente, este índice oscilará entre el valor 0 como menor y el 19 como más alto (los valores **8** y **9** no suman en ningún caso y son tratados en los análisis de manera diferente por hacer referencia a datos no disponibles y no aplicables respectivamente). Las diferencias de puntuación entre las distintas opciones son iguales (por ejemplo, 0, 0.5 y 1 para ítems con tres opciones de respuesta; 0, 0.33, 0.66 y 1 para ítems con cuatro opciones de respuesta), a excepción del ítem 4 referido al diseño donde, entre el 0.5 y el 1, se incluyeron con un 0.75 las "series temporales", consideradas con más calidad que los "diseños cuasiexperimentales" pero no con tanta calidad como los "diseños de discontinuidad en la regresión" y los "diseños experimentales" (Shadish y Myers, 2004). No se pusieron las cuatro opciones con distancias iguales por considerarse que los "diseños cuasiexperimentales" estaban más cercanos a los "diseños de series temporales" que a los "pre-experimentales". La explicación de este proceder, que en principio puede parecer arbitrario por lo novedoso, es que los diseños pre-experimentales presentan muchos problemas de validez que los "cuasiexperimentales" solventan (lo cual supone una gran diferencia entre ambos), mientras que las ventajas diferenciales entre las otras tres opciones son significativas pero más graduales: los diseños cuasiexperimentales no llegan a presentar la ventaja de los "diseños de series temporales" que suponen estudiar la evolución de un dato durante un largo periodo de tiempo (detección de posibles tendencias cíclicas, mayor seguridad sobre la consistencia de los datos, etc.); pero éstos últimos no llegan a tener la ventaja del conocimiento del criterio de asignación que los "diseños de discontinuidad en la regresión" presentan (Chacón, Shadish y Cook, en prensa).

- La tercera columna muestra la etiqueta de cada una de las **categorías**.

- Los ítems son diferenciados en función de las características a las que atiende; así, del ítem **0 al 23** se tratan las **características metodológicas**; y del **24 al 34** se hace referencia a **características sustantivas**.

ÍTEM	V	CATEGORÍAS
CARACTERÍSTICAS METODOLÓGICAS		

0. Tipo de estudio	**1**	Teórico (se describen modelos o no hay datos)
	2	Observacional
	3	Encuesta
	4	Cuasiexperimental
	5	Experimental
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio)
1. Grupo de comparación	**0**	No hay
	0.5	Inactivo
	1	Activo; en **diseños observacionales y de encuesta**, cuando sí hay
	8	No se especifica el dato
	9	No aplicable (teórico)
2. Criterios de selección de la muestra (inclusión/exclusión)	**0**	No especificados
	1	Especificados
	9	No aplicable (por el objeto de estudio o por ser teórico)
3. Azar	**0**	**Observaciones o encuestas** sin selección aleatoria de la muestra ni delimitación de los criterios de inclusión **Pre-experimentales** y **cuasiexperimentales** sin control de variables extrañas ni delimitación de los criterios de inclusión y asignación **Experimentales** con asignación y/o selección aleatorias inadecuados
	1	**Observaciones o encuestas** con selección aleatoria de la muestra o delimitación de los criterios de inclusión **Pre-experimentales** y **cuasiexperimentales** con control de variables extrañas o delimitación de los criterios de inclusión y asignación **Experimentales** con asignación y selección aleatorias adecuados
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
4. Diseño	**0**	**Observaciones o encuestas** con una o dos medidas **Pre-experimentales** **Experimentales** con un solo momento de medida
	0.5	**Observaciones o encuestas** con medidas entre 3 y 29 **Cuasiexperimentales** pre-post con grupo control no equivalente con medidas entre 2 y 29
	0.75	**Series temporales:** cuasiexperimentales con 30 o más medidas
	1	**Observaciones o encuestas** con 30 ó más medidas **Discontinuidad en la regresión** **Experimentales** con al menos dos momentos de medida
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
5. Muestra	**0**	N<12
	0.5	12≤N≤40
	1	N>40
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
6. Mortalidad global	**0**	≥20%
	0.5	0<N<20%
	1	0%
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
7. Mortalidad diferencial	**0**	≥20%
	0.5	0<n<20%
	1	0%
	8	No se especifica el dato
	9	No aplicable (teórico; un solo grupo; objeto de estudio incompatible)
8. Exclusiones posteriores	**0**	≥20%

a la agrupación de la muestra a las distintas condiciones	**0.5**	0<n<20%
	1	0%
	8	No se especifica el dato
	9	No aplicable (teórico; sólo un grupo; objeto de estudio incompatible)
9. Seguimiento	**0**	Nada
	0.3	<6 meses
	0.6	[6-11] meses
	1	≥12 meses
	8	No se especifica el dato
	9	No aplicable (teórico; no es posible recoger el dato en más de una ocasión; objeto de estudio incompatible)
10. Momentos de medida	**0**	Sólo posterior; una medida cuando no hay intervención
	1	Previo y posterior; más de una medida cuando no hay intervención
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
11. Medidas que aparecen en todos los momentos de registro	**0**	Más de una medida no aparece en todos los momentos de registro; también cuando ocurre esto en la única variable que se mide
	0.5	Una medida no aparece en todos los momentos de registro (siempre que se mida más de una variable)
	1	Todas las medidas son tomadas en todos los momentos de registro
	8	No se especifica el dato
	9	No aplicable (teórico; sólo un momento de registro; objeto de estudio incompatible)
12. Variables dependientes normalizadas (uso de instrumentos normalizados)	**0**	Sólo autoinformes sin estandarizar
	0.5	Ninguno normalizado pero al menos uno es cuestionario o autoinforme estandarizado (explícitamente expresado)
	1	Al menos uno es objetivo o normalizado
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
13. Enmascaramiento del evaluador	**0**	No: conoce la hipótesis del estudio
	1	Sí: desconoce la hipótesis del estudio
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
14. Enmascaramiento del usuario	**0**	No: conoce la hipótesis del estudio
	1	Sí: desconoce la hipótesis del estudio
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
15. Enmascaramiento del formador (como evaluador interno cuando no hay intervención)	**0**	No: conoce la hipótesis del estudio
	1	Sí: desconoce la hipótesis del estudio
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
16. Homogeneidad de la intervención o del proceso de registro cuando no hay intervención	**0**	No mismos proceso, intensidad (nº sesiones), duración y/o profesionales
	1	Mismos proceso, intensidad, duración y profesionales
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
17. Definición de los constructos	**0**	Ninguno definido conceptual y/o empíricamente
	0.5	Al menos 1 definido conceptual y/o empíricamente
	1	Todos definidos conceptual y empíricamente
	8	No se especifica el dato
18. Métodos estadísticos para inferir los valores perdidos	**0**	Ninguno. Sólo se analizaron los datos completos
	1	Sí o análisis "por intención de tratar"
	8	No se especifica el dato
	9	No aplicable (teórico; no se dieron valores perdidos; objeto de estudio incompatible)
19. Tamaño de efecto y	**0**	No se especifica

valor	**1**	Se especifica (o índices derivables)
	9	No aplicable (teórico; sólo una medida en un grupo; objeto de estudio incompatible)
20. Índice de calidad	**0-19**	Suma
21. Índice estadístico calculado	Anotar valor concreto	
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
22. Diferencias estadísticamente significativas entre medidas (explicitar qué se compara)	**0**	No
	1	Sí
	8	No se especifica el dato
	9	No aplicable (teórico; una sola medida; objeto de estudio incompatible)
23. Índice de variabilidad	Anotar valor concreto	
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
CARACTERÍSTICAS SUSTANTIVAS		
24. Número de participantes en cada grupo	Anotar valor concreto (la suma ha de dar el valor de la muestra)	
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
25. Número de grupos en el estudio	Anotar valor concreto	
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
26. Exclusiones tras medidas posteriores	**1**	No
	2	Sí
	8	No se especifica el dato
	9	No aplicable (teórico; una sola medida; objeto de estudio incompatible)
27. Rango de edad especificado	**1**	No se especifica
	2	Sí se especifica
	9	No aplicable (por el objeto de estudio o por ser teórico)
28. Media de edad	Anotar valor concreto	
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
29. Periodo de estudio	**1**	≤ 6 meses
	2	> 6 meses
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
30. Intensidad del tratamiento/registro	nº de horas/ periodo de tiempo	
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
31. Unidades de intervención o de registro	**1**	Individual (o grupo considerado como unidad)
	2	Grupal
	8	No se especifica el dato
	9	No aplicable (por el objeto de estudio o por ser teórico)
32. Área formativa	Anotar valor concreto	
33. Campo de intervención: destinatarios	Anotar valor concreto	
34. Tipo de publicación	**1**	Revista
	2	Libro
	3	Tesis
	4	Congreso
	5	Otras publicaciones
	6	Trabajos no publicados

Tabla 2.5. Versión integradora de la escala de medición de la calidad en estudios primarios.

Una diferencia de la última versión respecto a la previa es que, anteriormente, los estudios sin intervención no eran tenidos en cuenta con lo que, aplicando la escala, dichos estudios habrían sido no calificados en muchos de los ítems presentados; y los diseños mejor valorados eran los experimentales, seguidos de los cuasi y, finalmente, los que no mostraban intervención.

Partiendo de la idea de que el tipo de diseño no necesariamente correlaciona con la calidad metodológica, se trató de ampliar las categorías de cada ítem para adaptarlas a cualquier tipo de diseño de tal manera que el índice de calidad pudiera oscilar entre 0 y 19 independientemente del tipo de diseño que presentara el estudio (Chacón, Sánchez-Meca y Sanduvete, 2007).

En el apartado de discusión, se presenta una descripción detallada de los ítems que fueron modificados, los cambios concretos que se realizaron y las razones de dichos cambios.

A continuación, al entenderse que la definición clara y concisa de cada uno de los ítems y sus categorías es fundamental pues, para optimizar su utilidad, es muy importante que cualquier persona que quiera utilizar esta escala sepa a qué se refiere cada apartado y que todos tengan el mismo concepto respecto a lo que se está midiendo en cada momento, se presenta el **manual de codificación** que pretende aportar claridad al significado de cada uno de los conceptos para aumentar así la fiabilidad entre codificadores cuando se aplique esta escala.

CARACTERÍSTICAS METODOLÓGICAS

Ítem 0. Tipo de estudio. Es una categoría ómnibus en la que se trató de etiquetar de manera genérica el trabajo a categorizar tras hacer una primera lectura superficial, con la intención de facilitar la codificación del resto de variables.

a) Teórico: se consideraron teóricos todos aquellos trabajos en los que únicamente se describían modelos y/o donde no existían datos empíricos directos del contexto de intervención.

b) Observacional: fueron aquellos estudios en los que no se llevó a cabo ninguna intervención, sólo se registró, el contexto de los sujetos apenas fue modificado y la respuesta registrada no fue provocada.

c) Encuesta: fueron aquellos estudios en los que no se llevó a cabo ninguna intervención pero, para recoger información, se requirió a las personas que respondieran a determinados estímulos (en general, ítems por escrito o entrevistas).

d) Cuasi-experimental: fueron aquellos trabajos en los que se manipuló al menos una variable independiente para estudiar su efecto en una dependiente (por tanto, hubo intervención) y no hubo selección y/o asignación aleatoria de las personas que conformaron la muestra a las distintas condiciones.

e) Experimental: fueron aquellos estudios en los que se manipuló al menos una variable independiente para estudiar su efecto en una dependiente (hubo intervención) y las personas que conformaron la muestra fueron seleccionadas y posteriormente asignadas a los distintos grupos de manera aleatoria.

f) Cuando no se dio la información, se asignó un 8.

g) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 1: Grupo de comparación. Con este ítem se determina si existía un grupo aparte del considerado foco de atención del estudio y, en caso de que lo hubiera, qué nivel de participación tiene.

a) Sin grupo de comparación: cuando no hubo grupo de comparación, se asignó un 0.

b) Grupo de comparación inactivo: Los sujetos que conformaron el grupo de comparación no recibieron ninguna formación alternativa al grupo de intervención, bien porque estaban esperando a que comenzara la acción formativa o porque sencillamente no estaba prevista su realización. También se consideró grupo control inactivo al que no se le cambió la forma habitual de formarse (Shadish y Ragsdale, 1996). En este caso, se asignó un 0.5.

c) Grupo de comparación activo: Los sujetos del grupo control recibieron algún tipo de formación alternativa al grupo de intervención. Se valoró con un 1. En caso de que no se llevara a cabo intervención en el estudio, se valoró con esta máxima puntuación la existencia de un grupo de comparación observado o encuestado, aunque no fuera posible catalogarlo de activo.

d) Cuando no se dio la información, se asignó un 8.

e) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 2. Criterios de selección de la muestra (inclusión/exclusión): hace referencia a si se explicitaron los criterios de decisión por los que algunas personas participaron en el estudio, mientras que otras quedaron fuera. Es lo que Shadish, Cook y Campbell (2002b) denominan "assignment criteria".

a) **No se especificaron:** en los casos en los que no se explicitaron dichos criterios de selección, se valoró con un 0.

b) **Se especificaron**: en los casos en los que se explicitaron dichos criterios, se valoró con un 1.

c) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 3. Azar: hace referencia a si la selección de la muestra fue aleatoria y si los participantes fueron o no asignados al azar a las distintas condiciones en los casos en los que había más de una.

a) Se valoraron con un 0 los siguientes casos: las **observaciones y encuestas** cuya selección de la muestra **no** fue **aleatoria**; los **diseños pre-experimentales** y **cuasiexperimentales** (del tipo "pretest-postest con grupo control no equivalente") **sin control de variables extrañas**; los **diseños experimentales** con un **proceso de aleatorización inadecuado** en la selección de la muestra o la posterior asignación a los diferentes grupos.

b) Se valoraron con un 1 los casos siguientes: los diseños **observacionales y de encuestas** en los que la selección de la muestra se hizo **aleatoriamente**; los **diseños pre-experimentales** y **cuasiexperimentales** (del tipo "pretest-postest con grupo de control no equivalente") **con control de variables extrañas** como, por ejemplo, balanceo, bloqueo y estratificación; los **diseños experimentales** en los que la selección de la muestra y la posterior asignación de las personas a los distintos grupos se hizo con un sistema de aleatorización adecuado.

c) Cuando no se especificó el dato, se asignó un 8.

d) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 4. Diseño: Se valoró en este ítem la validez interna, la calidad y la estabilidad del dato (Cronbach, 1982).

a) Se asignó un 0 a las **observaciones** o **encuestas** con 1 ó 2 medidas; a los **diseños pre-experimentales**, caracterizados por la ausencia de selección y/o asignación aleatoria y porque sólo existe un grupo de personas que es medido transversalmente; por último, también a los **diseños experimentales** (en los que existe más de un grupo formado aleatoriamente) cuando sólo se da un momento de medida.

b) Se asignó un 0.5 cuando se trató de **metodología observacional** o de **encuestas**, con un número de medidas entre 3 y 29 (ambos valores inclusive); y los **diseños cuasiexperiementales** cuando se toman varias medidas a lo largo del tiempo, pero menos de 30.

c) Se asignó 0.75 a las **series temporales**, un tipo de diseño cuasiexpermental en el que un grupo es medido en 30 o más ocasiones a lo largo del tiempo en el pretest y el postest (Shadish y Myers, 2004).

d) Fueron valorados con un 1 los estudios **observacionales** y de **encuestas** con 30 o más medidas; los **diseños de discontinuidad en la regresión**, que son diseños cuasiexperiementales en los que se establece un punto de corte a partir del cual las personas van a recibir la intervención, por lo que se puede decir que los criterios de asignación a los grupos son totalmente conocidos; y los **diseños experimentales** con al menos dos momentos de medida.

e) Cuando no se especificó el dato, se asignó un 8.

f) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 5. Muestra: Se valora en este ítem el tamaño total de la muestra del estudio (N). Se encuadró cada caso en el siguiente sistema de puntuación:

a) **N<12:** cuando participaron en el estudio menos de doce personas, se asignó un 0.

b) **12≤N≤40:** cuando participaron entre 12 y 39 personas (ambos valores inclusive), se asignó un 0.5.

c) **N>40:** cuando participaron más de cuarenta personas, se asignó un 1.

d) En las ocasiones en las que no se especificó este dato, se asignó un 8.

e) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 6. Mortalidad global (Shadish, Cook y Campbell, 2002b). Se valora el número de personas que comienzan el estudio pero que, por diversas razones, no llegan a concluirlo. Las distintas categorías son:

a) **≥20%:** cuando el 20% o más de la muestra inicial no llegó a concluir el estudio, se asignó un 0.

b) **0<N<20%:** cuando entre el 0 y el 20% de la muestra inicial no llegó a concluir el estudio, se asignó un 0.5.

c) **0%:** cuando no se dio mortalidad (es decir, cuando todas las personas que comenzaron el estudio lo finalizaron), se asignó un 1.

d) Cuando no se especificó el dato, se asignó un 8.

e) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 7. Mortalidad diferencial: Este ítem valora la mortalidad diferencial sufrida entre dos grupos de comparación en un segundo momento de medida o posterior, con el siguiente sistema de puntuación:

a) **≥20%:** cuando la diferencia de mortalidad experimental entre grupos fue del 20% o más, se valoró este ítem con un 0.

b) **0<n<20%:** cuando la diferencia de mortalidad experimental entre grupos fue de entre el 0 y el 20%, se valoró con un 0.5.

c) **0%:** Cuando la mortalidad experimental fue igual en todos los grupos existentes, se valoró con un 1.

d) Cuando no se especificó el dato, se asignó un 8.

e) Cuando se consideró este ítem como no aplicable por tratarse de un estudio teórico o por tener un solo grupo, se asignó un 9.

Ítem 8. Exclusiones posteriores a la agrupación de la muestra a las distintas condiciones: este ítem hace referencia al porcentaje de sujetos que, habiendo sido asignados (o perteneciendo por sus propias características) a una de las condiciones de estudio, posteriormente son excluidos. Las puntuaciones son las siguientes:

a) **≥20%:** cuando el 20% o más de la muestra fue excluida tras la agrupación en las distintas condiciones, se valoró este ítem con un 0.

b) **0<n<20%:** cuando entre el 0 y el 20% de la muestra fue excluida tras la agrupación en las distintas condiciones, se asignó un 0.5.

c) **0%:** cuando no se excluyó posteriormente a ninguna de las personas agrupadas en las distintas condiciones, se valoró este ítem con un 1.

d) Cuando no se especificó el dato, se asignó un 8.

e) En las ocasiones en las que el estudio era teórico o sólo existía un grupo de personas o de medidas, se consideró este ítem como no aplicable, asignándosele un 9.

Ítem 9. Seguimiento: Se determina con este ítem durante cuánto tiempo se toman medidas, ya sea de la variable dependiente *una vez finalizada* la intervención, o de las variables registradas en los estudios en los que no hay intervención. Si se presentara más de un periodo de seguimiento, se tomaría el más amplio para responder a este ítem. Se distingue entre las siguientes posibilidades:

a) **0:** No se tomó ninguna medida de seguimiento. En estos casos, se asignó un 0.

b) **<6 meses:** Se tomaron medidas durante menos de 6 meses. En este caso, se asignó un 0.3.

c) **[6-11] meses:** Se tomaron medidas entre los 6 y 11 meses**,** ambos incluidos. En este caso, se asignó un 0.6.

d) **≥12 meses:** Se tomaron medidas durante un año o más, en cuyo caso se valoró este ítem con un 1.

e) Cuando no se especificó el dato, se asignó un 8.

f) Cuando se consideró este ítem como no aplicable por tratarse de un estudio teórico, se asignó un 9.

Ítem 10. Momentos de medida: Se pretende concretar cuándo se tomaron las medidas. Las categorías son:

a) **Sólo posterior o sólo una medida cuando no hay intervención:** se consideraron parte de esta categoría aquellos estudios en los que todas las medidas se tomaron tras la intervención y aquellos en los que, no habiendo intervención, se daba una sola medida. En estas ocasiones, se valoró con un 0.

b) **Previo y posterior o más de una medida cuando no hay intervención:** se eligió esta categoría cuando las medidas fueron tomadas antes y después de la intervención o cuando, sin haber intervención, se daba más de una medida de tal manera que se posibilitaba su comparación. En estas ocasiones, se valoró con un 1.

c) Cuando no se especificó el dato, se asignó un 8.

d) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 11. Medidas que aparecen en todos los momentos de registro: se hace en este ítem el conteo de medidas que, siendo tomadas en el primer momento, también aparecen en todos los demás momentos de registro. Se distingue entre:

a) **Más de una no se mide en todos los momentos de registro**: más de una medida inicial no se repite en todos los registros posteriores. Estos casos se valorarán con un 0, al igual que aquellos casos en los que **sólo se mide una variable y ésta no se registra en todos los momentos de medida**.

b) **Todas menos una:** una medida inicial no se repite en todos los registros posteriores, siempre que se mida más de una variable; de lo contrario, formaría parte del apartado a) previamente comentado. Estos casos se valoraron con un 0.5.

c) **Todas:** todas las medidas iniciales se repiten en todos los sucesivos registros. Estos casos fueron valorados con un 1.

d) Cuando no se especificó el dato, se asignó un 8.

e) Cuando se consideró este ítem como no aplicable por tratarse de un estudio teórico o por tener sólo un momento de registro, se asignó un 9.

Ítem 12. Variables dependientes normalizadas: se pretende determinar qué grado de estandarización y objetividad poseen las variables medidas en función del instrumento que se utiliza para su registro (Anguera, Chacón, Holgado y Pérez, en prensa). Se diferencian las siguientes categorías:

a) Únicamente **no estándares:** hace referencia a aquellas medidas tomadas con instrumentos elaborados para una situación de evaluación concreta y no validados ni baremados, por lo que no pueden asegurar la validez de la medida y no existe una población de referencia con la que comparar el dato obtenido en una situación concreta. Algunos instrumentos con estas características son los registros de conducta, las fuentes documentales y los autoinformes sin estandarizar (para que se consideren estandarizados, se ha de comentar este hecho explícitamente en el estudio). Estos casos se valoraron con un 0.

b) Ninguno normalizado, pero al menos uno es **estándar**: aunque no hubo ninguna variable dependiente medida con instrumento normalizado, al menos una de ellas fue medida con instrumentos que, aunque no estaban baremados ni validados, sí estaban estructurados, por lo que se pudo afirmar con cierto grado de seguridad que se estaba midiendo, mediante un procedimiento reglado y homogéneo para todos los usuarios, aquello que se pretendía medir. Algunos ejemplos de estos instrumentos son las entrevistas estructuradas, los cuestionarios, las encuestas y los autoinformes estandarizados (dicho explícitamente, pues de lo contrario se consideró un autoinforme sin estandarizar y formó parte de la opción a). Estos casos se valoraron con un 0.5.

c) Al menos uno es **normalizado:** se incluyen en esta categoría aquellos casos en los que, al menos, una variable dependiente fue medida con pruebas objetivas, baremadas (con una población de referencia con la que comparar el dato obtenido) y/o validadas. Unos ejemplos de estos instrumentos son los registros fisiológicos, los tests y los registros de observación. Estos casos se valoraron con un 1.

d) Cuando no se especificó el dato, se asignó un 8.

e) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 13. Enmascaramiento del evaluador (codificador): se valora si los evaluadores desconocían las hipótesis del estudio, con el siguiente sistema de puntuación:

a) **No:** los evaluadores conocían las hipótesis del estudio. En estos casos, se asignó un 0.

b) **Sí:** los evaluadores desconocían las hipótesis del estudio. En estos casos, se asignó un 1.

c) Cuando no se especificó el dato, se asignó un 8.

d) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 14. Enmascaramiento del usuario: se valora si las personas que participaron en el estudio desconocían las hipótesis de partida.

a) **No:** el alumnado conocía las hipótesis del estudio. En estos casos, se asignó un 0.

b) **Sí:** el alumnado desconocía las hipótesis del estudio. En estos casos, se asignó un 1.

c) Cuando no se especificó el dato, se asignó un 8.

d) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 15. Enmascaramiento del profesional que realiza la intervención o del evaluador interno cuando no hay intervención: se valora si las personas encargadas de implementar la intervención desconocían las hipótesis del estudio; en los casos en los que no hubo intervención, se valoró con este ítem el conocimiento o desconocimiento de los objetivos por parte de un evaluador interno (en caso de que lo hubiera) que sería una persona que, formando parte de los sujetos estudiados, recoge datos para dicho estudio.

a) **No:** los formadores o evaluadores internos conocían las hipótesis del estudio. En estos casos, se asignó un 0.

b) **Sí:** los formadores o evaluadores internos desconocían las hipótesis del estudio. En estos casos, se asignó un 1.

c) Cuando no se especificó el dato, se asignó un 8.

d) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 16. Homogeneidad de la intervención o proceso de registro: se valora la integridad de la intervención o proceso de registro; es decir, el grado en que todas las personas del estudio participaron en intervenciones con las mismas condiciones (intensidad -nº sesiones-, horas de cada acción formativa y formador) y, en caso de que no hubiera intervención, se valora si el proceso de registro fue para todos igual.

a) **No**: las personas participaron en intervenciones con distinta intensidad, duración y/o profesionales. Cuando no hubo intervención, el proceso de registro varió. En estos casos, el ítem fue valorado con un 0.

b) **Sí**: Todas las personas participaron en intervenciones con las mismas características en intensidad, duración y profesionales. Cuando no hubo intervención, todos los datos se registraron con el mismo proceso. En estos casos, se valoró con un 1.

c) Cuando no se especificó el dato, se asignó un 8.

d) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 17. Definición del constructo: se valora si los constructos de interés en la investigación se han definido con claridad y de forma operativa. En concreto, debe hacerse referencia, al menos, al tipo de intervención implementada (variable/s independiente/s) si la hubiese, y a la medida de las variables dependientes o de respuesta, con el siguiente sistema de puntuación:

a) **Ninguno definido conceptual y/o empíricamente:** Cuando ningún constructo implicado en el estudio se definió de forma conceptual ni empírica (operativa), se valoró con un 0.

b) **Al menos uno definido conceptual y/o empíricamente:** cuando al menos uno de los constructos implicados en el estudio se definió de forma conceptual y/o empírica (operativa), se valoró con un 0.5.

c) **Todos definidos conceptual y empíricamente:** los casos en los que todos los constructos implicados en el estudio se definieron tanto conceptual como empíricamente (operativamente) se valoraron con un 1.

d) Cuando no se especificó el dato, se asignó un 8.

Ítem 18. Métodos estadísticos para inferir los valores perdidos: Se valora si se ha realizado algún tratamiento estadístico de los valores perdidos.

a) **No**: los casos en los que no se aplicó ningún procedimiento estadístico de imputación de valores perdidos, o no se realizaron los análisis estadísticos incluyendo los datos "por intención de tratar", sino sólo los de aquellas personas cuyo registro incluía todas las medidas posibles ("completer analysis"), se valoró con un 0.

b) **Sí**: cuando se aplicó algún procedimiento estadístico de imputación de valores perdidos o se realizaron los análisis estadísticos sobre los datos "por intención de tratar" ("intention-to-treat analysis"), este ítem fue valorado con un 1. Se anotó el nombre del procedimiento concreto que se siguió.

c) Cuando no se especificó el dato, se asignó un 8.

d) Cuando se consideró el ítem como no aplicable por tratarse de un estudio teórico o por no haber valores perdidos, se asignó un 9.

Ítem 19. Tamaño de efecto y valor: consiste en discernir si los resultados del estudio fueron dados mediante un estadístico que muestra el tamaño de efecto.

a) **No**: cuando los resultados del estudio no fueron dados mediante un estadístico que mostraba el tamaño de efecto, se valoró con un 0.

b) **Sí**: cuando los resultados del estudio fueron dados mediante un estadístico que mostraba el tamaño de efecto como, por ejemplo, la diferencia de medias estandarizada "d"; o mediante índices derivables como "O Ratio" o la media, desviación tipo y tamaño de la muestra (son necesarios los tres valores para calcular el tamaño de efecto), se asignó un 1.

c) Si se consideró este ítem como no aplicable por tratarse de un estudio teórico o en el que sólo existiera una medida en un solo grupo (hecho que impediría cualquier posible comparación), se asignó un 9.

Ítem 20. Índice de calidad: en esta variable se calculó la suma de las puntuaciones que el estudio alcanzó en los ítems anteriores (excluyendo el "ítem 0" considerado meramente descriptivo), con los que se evaluó la calidad metodológica. Las puntuaciones mínima y máxima en este ítem, por tanto, fueron 0 y 19 respectivamente.

Los ítems que a partir de ahora se presentan (3 referidos a características metodológicas y 11 a sustantivas) tienen como finalidad la realización de un estudio descriptivo de las características que suelen mostrar los trabajos que versan sobre formación continua para empleados y empleadas. Por esta razón, no son contabilizados de manera cuantitativa y no aportan valor al índice de calidad (ítem 20). Esta es también la causa por la que no se aporta un valor numérico a cada categoría, como aparecía en los 20 ítems anteriores (en la escala aparece por facilitar la codificación, pero son números que no aportan valor respecto al constructo "calidad").

Ítem 21. Índice estadístico calculado y valor por grupo: se presenta como ítem de formato abierto, y hace referencia a los índices estadísticos que se presentaron en el estudio y su valor. En los casos en los que existía más de un grupo, o más de una medida en el tiempo, se especificaron los valores en cada uno de éstos. No se trata de una categoría mutuamente excluyente, por lo que para un solo estudio se pueden haber marcado varios estadísticos. Cuando no se especificó el dato, se asignó un 8. Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 22. Diferencias estadísticamente significativas entre medidas (de grupos y/o registros): este ítem responde a si las diferencias entre medidas (de varios grupos y/o varios momentos) son estadísticamente significativas o no. Se explicitó en cada caso qué se estaba comparando concretamente.

a) **No**: se recogieron en este primer apartado los estudios que no mostraron diferencias estadísticamente significativas.

b) **Sí**: los estudios en los que se encontraron diferencias estadísticamente significativas marcaron en esta categoría.

c) Cuando no se especificó el dato, se asignó un 8.

d) Cuando se consideró que este ítem no era aplicable por tratarse de un estudio teórico o con una sola medida, se asignó un 9.

Ítem 23. Índice de variabilidad facilitado y valor por grupo: con este ítem de formato abierto se recogió el índice de variabilidad que se mostraba en el estudio y su valor concreto. Cuando se hallaron varios grupos o medidas con este dato especificado, se recogieron todos ellos por separado. Cuando no se especificó el dato, se asignó un 8.

Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

CARACTERÍSTICAS SUSTANTIVAS

Ítem 24. Nº de participantes por grupo: hace referencia al número de personas que conformaban un grupo de tal manera que, al sumar todos los componentes de todos los grupos, el valor obtenido coincidirá con el tamaño de la muestra. Se recogió el valor concreto y, posteriormente, se crearon intervalos para agrupar dichos valores. Además, se asignó un 8 en los casos en que no se especificaba el dato y si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 25. Nº de grupos en el estudio: en este ítem se determinó el número de grupos que aparecían en el estudio. Al igual que en el ítem anterior, se recogió el valor concreto y, posteriormente, se crearon intervalos para agrupar dichos valores. Se asignó un 8 en los casos en que no se especificaba el dato. Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 26. Exclusiones tras medidas posteriores: en este ítem se constató si alguna persona era excluida del estudio una vez iniciado éste con, al menos, una medida ya tomada.

a) **No:** Se incluyeron aquellos casos en los que no se excluyó ninguna persona del estudio en ningún momento una vez iniciada la recogida de información.

b) **Sí:** En esta categoría se incluyeron aquellos casos en los que se daban exclusiones a partir de la segunda medida.

c) Se asignó un 8 en los casos en que no se especificaba el dato.

d) Se otorgó un 9 cuando se consideró un ítem no aplicable porque fuera un estudio teórico o porque se recogiera una única medida.

Ítem 27. Rango de edad especificado: este apartado se centra en una característica de la muestra; concretamente, el rango de edad de los participantes.

a) **No**: En la primera categoría, se incluyeron aquellos estudios en los que no se especifica el rango de edad de los participantes.

b) **Sí**: En esta categoría, se incluyeron aquellos estudios en los que sí se especificó rango de edad. Se anotó además el valor concreto.

c) Se otorgó un 9 cuando se consideró un ítem no aplicable porque fuera un estudio teórico o porque se recogiera una única medida.

Ítem 28. Media de edad (valor concreto): al igual que el anterior, este ítem también hace referencia a una característica de la muestra, concretamente la media de edad de

los participantes en el estudio. Se anotó el dato concreto y posteriormente se crearon rangos con los que se aunaron los resultados en grupos. Cuando no se dio el dato, se asignó un 8; si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 29. Periodo de estudio: este ítem hace referencia a la duración total del estudio, desde la primera medida o inicio de la intervención hasta la última medida incluyendo, por tanto, periodo de seguimiento si lo hubiera.

a) **≤ 6 meses**: el primer apartado recogió a aquellos estudios en los que la duración fue igual o inferior a 6 meses.

b) **> 6 meses**: en este apartado se incluyeron aquellos estudios cuya duración fue mayor a 6 meses.

c) Cuando no se especificó el dato, se asignó un 8.

d) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 30. Intensidad del tratamiento o del registro cuando no hay intervención: este ítem, de formato abierto, pone en relación el número de horas de duración de la acción formativa con el periodo de tiempo durante el que ésta se llevó a cabo, preferiblemente en semanas si así estuviera especificado, aunque quedó abierto para recoger aquellos valores expresados en otras unidades. Además, cuando no hubo intervención, se recogió información acerca de cuántas medidas, de qué duración, se hacían en qué periodo de tiempo concreto. Cuando no se especificó este dato, se asignó un 8. Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 31. Unidades de intervención o de registro: consiste en especificar si la intervención se realiza a una única persona (de manera individual) o a varias (a un grupo); en caso de no darse intervención, se respondería a si se toma medidas de una sola persona o de un grupo de ellas.

a) **Individual**: la primera categoría recogió a aquellos casos en los que la intervención o la medida se realizaba a un individuo o a un grupo considerado como una unidad.

b) **Grupal**: esta categoría recogió a aquellos casos en los que en la intervención participaba un grupo de personas o se tomaban medidas de más de una persona.

c) En los casos en los que no se especificó el dato, se asignó un 8.

d) Si no fue posible valorar esta categoría por no ser aplicable, ya sea por tratarse de un estudio teórico o por la naturaleza del objeto de estudio, se asignó un 9.

Ítem 32. Área formativa: este ítem de formato abierto hace referencia a los contenidos sustantivos sobre los que versaba el estudio.

Ítem 33. Campo de intervención: en este apartado de respuesta abierta se hace referencia a los destinatarios de los que hablaba el estudio.

Ítem 34. Tipo de publicación: se recoge, de manera genérica, la fuente de información de la que se tomaron los estudios incluidos. Las diferentes opciones son las siguientes:

a) **Revista**

b) **Libro**

c) **Tesis**

d) **Congreso**

e) **Otras publicaciones**

f) **Trabajos no publicados**

6. DISCUSIÓN.

A continuación, en la tabla 2.6 se resume el procedimiento utilizado en cada una de las fases de elaboración de la escala de medición de la calidad de estudios primarios y sus resultados, hasta llegar a la hasta ahora última versión del instrumento (Chacón, Sanduvete, Sánchez y Sánchez-Meca, 2007b). Las columnas hacen referencia a cada una de las fases realizadas, en orden cronológico, empezando por la de más antigüedad; y las filas diferencian entre los distintos aspectos de la metodología y los resultados.

Y, posteriormente, en la tabla 2.7 se presenta un resumen de la evolución por la que han pasado los distintos ítems hasta llegar a la última versión de la escala de calidad de estudios primarios. Concretamente:

- La primera columna muestra una **etiqueta** general para reconocer el contenido del ítem rápidamente, con pocas palabras.

- En la segunda columna se muestran los ítems que componían el **cuestionario exploratorio** con el que se llevó a cabo el estudio de validez de contenido (Chacón, Sanduvete y Alarcón, 2005).

- En la tercera columna, se presentan los ítems de la versión depurada de la escala, creada a partir de los **resultados del estudio de validez de contenido** y tras la fusión con algunos ítems que generalmente son usados en los **estudios meta-analíticos** (Chacón, Sánchez-Meca, Sanduvete y Alarcón, 2006).

- La cuarta columna muestra los ítems definitivos tras **concretar algunas definiciones**, **operativizar** más los conceptos y realizar una **adaptación** para generalizar el uso de la escala a cualquier estudio, independientemente del **diseño** que presente (Chacón, Sánchez-Meca y Sanduvete, 2007; Chacón, Sánchez-Meca, Sanduvete y Alarcón, en elaboración).

- La quinta columna muestra una breve explicación de las **razones** por las que se hicieron los cambios, detallados con mayor detenimiento más adelante.

En negrita se marcan las diferencias en un mismo ítem entre los diferentes momentos.

Y tras la presentación de la tabla, se describen los cambios acaecidos, las razones que llevaron a ellos y las consecuencias que sucedieron a dichos cambios.

	1ª FASE. ELABORACIÓN DEL CUESTIONARIO PARA EL ESTUDIO DE VALIDEZ DE CONTENIDO	2ª FASE. ESTUDIO DE VALIDEZ DE CONTENIDO	3ª FASE. ELABORACIÓN DE LA VERSIÓN DEPURADA DE LA ESCALA	4ª FASE. ELABORACIÓN DE LA VERSIÓN INTEGRADORA DE LA ESCALA
MUESTRA	1. Elaboración de la escala: 27 documentos disponibles acerca de la medición de la calidad en estudios primarios. 2. Estudio bibliográfico: 1899 resúmenes sobre estudios psicológicos, sociales o de educación.	30 expertos en meta-análisis y revisiones sistemáticas, calidad, evaluación y diseño.	1. Elaboración de la escala: - Los 23 ítems que cumplieron el criterio de inclusión en el estudio de validez de contenido. - Otras escalas disponibles (ej., Sánchez-Meca, 1998) 2. Análisis bibliográfico: 95 trabajos completos acerca de formación continua.	Escala depurada (34 ítems)
INSTRUMENTOS	1. Elaboración de la escala: - Bases de datos electrónicas. - Procite. 2. Estudio bibliográfico: - Bases de datos electrónicas. - Procite. - SPSS 12.0. - Sistema de categorías (19 ítems).	- Cuestionario exploratorio con 43 ítems, tres opciones de respuesta (de -1 a +1) y 3 conceptos a evaluar (representatividad, utilidad y viabilidad del dato). - Internet para envío y recogida de los datos. - Microsoft Excel para el análisis de datos.	1. Elaboración de la escala: - Ítems que cumplieron el criterio de inclusión. - Ítems de otras escalas. 2. Análisis bibliográfico: - Escala, versión depurada. - Bases de datos informatizadas. - Procite. - SPSS 12.0.	Escala depurada (34 ítems)
PROCEDIMIENTO	1. Elaboración de la escala: - Búsqueda de artículos. - Recogida de ítems. - Estructuración en dominios y subdominios. 2. Análisis bibliográfico: - Búsqueda de resúmenes. - Elección de los ítems que se codificarían. - Codificación (3 codificadores).	- Selección de la muestra. - Distribución del instrumento. - Análisis de datos.	1. Elaboración de la escala: - Recopilación de ítems con la aplicación del criterio de inclusión. - Recopilación de ítems de otras escalas disponibles. - Comparación entre ambos resultados. - Modificación para crear un instrumento más completo. 2. Análisis bibliográfico: - Búsqueda de artículos completos referidos a la formación continua. - Codificación.	Cambios en los términos utilizados y algunas categorías para lograr dos objetivos: - Comparaciones homogéneas entre diseños (la calidad de cualquier tipo de diseño podría oscilar entre 0 y 19). - Mayor concreción y operacionalización.

RESULTADOS	1. Cuestionario exploratorio (43 ítems) con tres dimensiones: características extrínsecas, sustantivas y metodológicas. 2. Estudio exploratorio: características más frecuentes respecto a la calidad del diseño en los estudios publicados.	Estudio de validez de contenido: cada ítem presentó tres índices que podían oscilar entre -1 y +1 en los tres conceptos estudiados: representatividad, utilidad y viabilidad del dato. Criterio de inclusión: un valor de 0.5 o mayor en al menos dos de los tres conceptos.	1. Escala con 34 ítems. 2. Estudio exploratorio: características más frecuentes respecto a la calidad del diseño en los estudios publicados sobre formación continua.	- La nueva escala resultante (35 ítems), instrumento que se usará para conocer las características de calidad que suelen presentar los estudios en formación continua y proponer mejoras (capítulo 3).

Tabla 2.6. Método y resultados en cada fase de elaboración de la escala de medición de la calidad en estudios primarios.

IDENTIFICADOR	EXPLORATORIO (2005)	V. DEPURADA (2006)	V. INTEGRADORA (2007)	RAZONES DE CAMBIOS
Tipo de estudio	---	---	**0. Tipo de estudio: 0-teórico (se describen modelos o no hay datos); 1-observacional; 2-encuesta; 3-cuasi-experimental; 4-experimental**	**Incorporación.** Descriptivo, a cumplimentar a posteriori, ayuda a tener una visión general
Grupo de comparación	---	1. Grupo control: 0-inactivo; 1-activo	1. Grupo de **comparación**: **0-no hay**; 0.5-inactivo; 1-activo o **sí hay (en diseños sin intervención)**	**Incorporación, 1º total y después parcial.** 1º, necesidad de determinar la existencia o no de grupo control y el tipo; 2º, adaptación a diseños sin intervención
Criterios de selección de la muestra	21. Criterios de inclusión y exclusión de las unidades de la muestra explicitados: 1-no;2-sí	2. Criterios de selección de la muestra (inclusión/exclusión) (Assignment criteria; en Shadish, Cook y Campbell, 2002b, p.323): 0- no especificados; 1-especificados	2. Criterios de selección de la muestra (inclusión/exclusión) (Assignment criteria; en Shadish, Cook y Campbell, 2002b, p.323): 0- no especificados; 1-especificados	---
Azar	22. Asignación aleatoria de las unidades a los grupos: 1-no y sin control de variables extrañas; 2-no pero con control de variables extrañas; 3-sí	3. Azar: **0- pre-experimentales y cuasiexperimentales sin control de vvee; 0'5-cuasiexprimentales con control de vvee (balanceo, bloqueo, estratificación); 1-experimentales (asignación aleatoria de las unidades a grupos)**	3. Azar: **0- experimentales con asignación y/o selección aleatoria inadecuados;** pre-experimentales y cuasiexperimentales sin control de vvee; **observaciones o encuestas sin selección aleatoria de la muestra:** 1-experimentales **(selección de la muestra** y asignación aleatoria a grupos adecuados); **cuasiexprimentales y pre-experimentales con control de vvee (balanceo, bloqueo, estratificación); observaciones o encuestas con selección aleatoria de la muestra**	**Incorporación parcial:** 1º, definición más concreta; 2º asignación + selección y diferentes grados de calidad en cada diseño **Consecuencias:** - Cualquier tipo de diseño puede puntuar del 0 al 1 en función de su calidad - Anteriormente, sólo se tenía en cuenta el método de asignación; posteriormente, también la selección

Diseño	23. Tipo de metodología/ diseño: 1-pre-experimental (sólo un grupo; una medida); 2-cuasi-experimental (dos grupos sin asignación aleatoria o con grupo control no equivalente con pre-test y post-test; 3-experimental (aleatorio); **4-otros (cuestionarios, observación)**	4. Tipo de metodología/ diseño: 0-pre-experimental; 0.5-cuasi-experimental; 0.75-serie temporal (obs pre ≥30 y post ≥30) y discontinuidad en regresión; 1-experimental-aleatorio	4. Diseño: 0-observaciones o encuestas con una o dos medidas; pre-experimentales; **experimentales con un solo momento de medida**; 0.5-Pre-post **con grupo control no equivalente con medidas entre 2 y 29: observaciones o encuestas con medidas entre 3 y 29**; 0.75-series temporales; 1- **observaciones o encuestas con 30 ó más medidas; discontinuidad en la regresión;** experimentales **con al menos dos momentos de medida**	**Incorporación parcial en cada una de las opciones.** **Consecuencias:** - Cualquier tipo de diseño puede puntuar del 0 al 1 en función de tres criterios: 1. Si se conoce el método de asignación 2. Número de medidas 3. Número de grupos
Muestra	24. Tamaño de la muestra: 1-n ≤5; 2-5 <n <10; 3-n ≥10	5. Muestra: 0 - n<**12**; 0.5 - n=[**12-40**]; 1- n>**40**	5. Muestra: 0 - n<12; 0.5 - n=[12-40]; 1- n>40	**Modificación del punto de corte** por no resultar informativo: antes, poca sensibilidad; no se encontraba una muestra menor de 5 **Consecuencias:** Ahora es más informativo y es más fácil obtener un valor medio en este ítem, pero más complicado obtener el valor máximo.
Mortalidad global	26. Mortalidad experimental: 1- ≤30%; 2->30%	6. Mortalidad global: 0- ≥**20**%; 0.5 - %=]**0-20**[; 1- **0**%	6. Mortalidad global: 0- ≥20%; 0.5 - %=]0-20[; 1- 0%	**Modificación del punto de corte:** antes, poca sensibilidad; no se encontraba una mortalidad mayor de 30 **Consecuencias:** Mayor concreción y más informativo
Mortalidad diferencial	28. Mortalidad experimental entre grupos: 1- homogéneo; 2- no homogéneo	7. Mortalidad diferencial: **0- ≥20%; 0.5 - %=]0-20[; 1- 0%**	7. Mortalidad diferencial: 0- ≥20%; 0.5 - %=]0-20[; 1- 0%	**Modificación**, por creación del punto de corte. **Consecuencia:** mayor concreción y operacionalización
Exclusiones posteriores	29. Exclusiones tras la asignación aleatoria (por ejemplo, no codificados): 1- no; 2-sí (número)	8. Exclusiones posteriores a la asignación aleatoria **(post-assignment attrition; en Shadish, Cook y Campbell, 2002b, p.323): 0- ≥20%; 0.5 - %=]0-20[; 1- 0%**	8. Exclusiones posteriores a la **agrupación de la muestra** (post-assignment attrition; en Shadish, Cook y Campbell, 2002b, p.323): 0- ≥20%; 0.5 - %=]0-20[; 1- 0%	**1º, modificación** por creación de puntos de corte; 2º, **modificación terminológica** **Consecuencia:** 1º, mayor concreción y operacionalización; 2º, adaptación del ítem a los diseños sin intervención

Seguimiento	31. Periodo de seguimiento: 1 < 5 meses; 2-[6-11] meses; 3- > 12 meses	9. Seguimiento: **0 – no se da seguimiento; 0.3< 6 meses**; 0.6 – meses = [6-11]; 1- ≥12 meses	9. Seguimiento: 0 – no se da seguimiento; 0.3< 6 meses; 0.6 – meses = [6-11]; 1- ≥12 meses	**Modificación de los puntos de corte.** **Consecuencias:** - En primer lugar, un estudio de 5 meses de seguimiento puntuaría como si no lo tuviera; ahora puntúa 0.6 más - Mayor concreción: se pasó de dos a cuatro posibilidades
Momentos de medida	32. Momentos de medida: 1- post intervención; 2-pre y post intervención	10. Momentos de medida: 0- posterior; 1- previo y posterior	10. Momentos de medida: 0- sólo posterior **o una medida cuando no hay intervención**; 1- previo y posterior **o más de una medida si no hay intervención**	**Incorporación parcial.** **Consecuencia:** adaptación del ítem a los diseños sin intervención
Medidas que aparecen en los distintos momentos de registro	33. Las medidas del pre-test aparecen en el pos-test: 1- ninguna; 2-algunas; 3-todas	11. Las medidas del pre-test aparecen en el pos-test: **0- falta más de uno; 0.5- falta 1**; 1- todas aparecen en todos los momentos	11. Medidas que aparecen **en todos los momentos de registro** : 0->1 no aparece en todos los momentos; 0.5-1 no aparece en todos los momentos de registro **(siempre que se mida más de una variable)**; 1-todas aparecen en todos los momentos	**1º, modificación parcial de las categorías; 2º, modificación terminológica; 3º, ampliación de la explicación de la 2ª categoría.** **Consecuencia:** 1º, más concreción y operatividad; 2º, adaptación del ítem a los diseños sin intervención; 3º, atención especial cuando sólo se mide una variable
Variables dependientes normalizadas	34. Variables dependientes normalizadas: 1-no hay (autoinformes y medidas a posteriori); 2-cuestionarios o autoinformes estandarizados; 3-al menos una es objetiva (medidas psicofisiológicas)	12. Variables dependientes normalizadas: 0- sólo autoinformes sin estandarizar; 0.5-ninguno normalizado, pero al menos uno es cuestionario o autoinforme estandarizado; 1-al menos un instrumento es objetivo o normalizado (p.ej. registro fisiológico, pruebas baremadas)	12. Variables dependientes normalizadas: 0- sólo autoinformes sin estandarizar; 0.5-ninguno normalizado, pero al menos uno es cuestionario o autoinforme estandarizado; 1- al menos un instrumento es objetivo o normalizado (p.ej. registro fisiológico, pruebas baremadas)	---
Enmascaramiento del evaluador	---	**13. Enmascaramiento del evaluador: 0-no; 1-sí**	13. Enmascaramiento del evaluador: 0-no; 1-sí	**Incorporación total.** **Consecuencias:** mayor concreción: diferenciación entre quienes realizan la intervención y quienes evalúan

Enmascaramiento del usuario	36. Técnicas de control: 1-ciego (de beneficiarios); 2-ciego (de implementadores); 3-doble ciego (ambos); 4-otros (especificar)	**14. Enmascaramiento del usuario: 0-no; 1-sí**	14. Enmascaramiento del usuario: 0-no; 1-sí	**Modificación por desagregación.** **Consecuencias:** un solo ítem pasó a dos
Enmascaramiento del profesional	---	**15. Enmascaramiento del profesional que realiza la intervención: 0-no; 1-sí**	15. Enmascaramiento del profesional que realiza la **intervención (del evaluador interno cuando no hay intervención)**: 0-no; 1-sí	**1º, modificación por desagregación; 2º, incorporación parcial** para generalizar su uso a todo tipo de diseño **Consecuencias:** 1º, un solo ítem pasó a dos; 2º, adaptación para su uso en todo tipo de diseños
Homogeneidad	35. Homogeneidad de la intervención/estudio: 1-las personas no reciben el tratamiento en las mismas condiciones contextuales; 2-las personas reciben el tratamiento en las mismas condiciones contextuales	16. Homogeneidad de la intervención: 0- no todos los sujetos han recibido la misma **intensidad de intervención, duración y profesionales**; 1-todos los sujetos han recibido la misma **intensidad de intervención, duración y profesionales**	16. Homogeneidad de la intervención **o proceso de registro**: 0- no todos los sujetos han recibido la misma intensidad de intervención, duración y **profesionales o proceso de medición cuando no hay intervención**; 1- todos los sujetos han recibido la misma intensidad de intervención, duración y profesionales **o proceso de medición cuando no hay intervención**	**Incorporación, 1º parcial** para definición más concreta **y 2º incorporación parcial** para generalizar el uso también a los diseños sin intervención **Consecuencias:** - Mayor especificación del constructo y mayor operacionalización. - Adaptación del ítem a los diseños sin intervención
Definición del constructo	37. Definición del constructo del resultado: 1-replicable por el lector en su propio contexto; 2-definición vaga; 3-no definición	17. Definición del constructo: **0-no; 0.5-no operativa; 1-operativa**	17. Definición del/los constructo/s **(variable/s dependiente/s e independiente/s cuando hay intervención): 0-ninguno definido conceptual y/o empíricamente; 0.5-sí, al menos uno definido conceptual y/o empíricamente; 1-todos definidos conceptual y empíricamente**	**2 modificaciones** para mayor especificación. **Consecuencias:** mayor especificación y operacionalización
Inferencia de valores perdidos	38. Métodos estadísticos para tratar los valores perdidos: 1-no; 2-sí	18. Métodos estadísticos para inferir los valores perdidos: 0-no; 1-sí	18. Métodos estadísticos para inferir los valores perdidos: 0-no; 1-sí	---

Tamaño de efecto y valor	40. Tamaño de efecto y valor	19. Tamaño de efecto y valor: **0-no; 1-sí**	19. Tamaño de efecto y valor: 0-no; 1-sí	**Incorporación parcial:** especificación de dos opciones para que pueda puntuar entre 0 y 1 **Consecuencias:** inclusión para el cálculo del índice de calidad
Índice de calidad	---	**20. Índice de calidad: suma de puntuaciones: 0-19**	20. Índice de calidad: suma de puntuaciones: 0-19	**Incorporación total:** índice cuantitativo que determina la calidad de cada estudio **Consecuencia:** información cuantitativa integradora
Índice estadístico	---	**21. Índice estadístico calculado**	21. Índice estadístico calculado	**Incorporación total:** para estudiar la posibilidad de integración en meta-análisis **Consecuencia:** recogida de dato relevante para meta-análisis
Diferencias significativas	---	**22. Los resultados mostraron diferencias estadísticamente significativas: 0-no; 1-sí**	22. Las medidas mostraron diferencias estadísticamente significativas: 0-no; 1-sí	**1º, incorporación total:** descriptivo para estudiar la posibilidad del "sesgo de publicación" de cara a realizar un meta-análisis; y **2º, adaptación** del ítem a los diseños sin intervención **Consecuencia:** recogida de dato relevante para meta-análisis
Índice de variabilidad	---	**23. Índice de variabilidad facilitado**	23. Índice de variabilidad facilitado	**Incorporación total:** para estudiar la posibilidad de integración en meta-análisis **Consecuencia:** recogida de dato relevante para meta-análisis
Nº participantes en cada grupo	---	**24. Número de participantes en cada grupo**	24. Número de participantes en cada grupo	**Incorporación total:** especificación mayor de la muestra **Consecuencia:** mayor especificación
Nº grupos	---	**25. Número de grupos en el estudio**	25. Número de grupos en el estudio	**Incorporación total:** especificación mayor del diseño **Consecuencia:** mayor especificación
Exclusiones tras medidas posteriores	---	**26. Exclusiones tras medidas posteriores: 0-no; 1-sí**	26. Exclusiones tras medidas posteriores: 1-no; 2-sí	**Incorporación total:** especificación mayor del proceso de recogida de datos **Consecuencia:** mayor especificación

Rango de edad	7. Edad (rango) referido: 1-no; 2-sí	27. Rango de edad especificado: 0-no; 1-sí	27. Rango de edad especificado: 1-no; 2-sí	---
Media de edad	8. Edad (media)	28. Media de edad (valor concreto)	28. Media de edad (valor concreto)	---
Periodo	17. Periodo de tratamiento	29. Periodo de tratamiento: **0.5-≤6 meses; 1->6 meses**	29. Periodo de **estudio**: 1- ≤6 meses; 2->6 meses	**1º, incorporación parcial:** puntos de corte; **2º**, **Modificación terminológica** **Consecuencia:** 1º, facilitación de recogida del dato; 2º, adaptación del ítem para su uso con diseños sin intervención
Intensidad	18. Intensidad del tratamiento (p.ej. número de dosis)	30. Intensidad del tratamiento: **número de sesiones a la semana**	30. Intensidad del tratamiento **o del registro cuando no hay intervención**: número de sesiones a la semana	**Incorporación, 1º parcial** para definición más concreta **y 2º incorporación parcial** para generalizar el uso también a los diseños sin intervención **Consecuencias:** 1º, mayor operacionalización; 2º, adaptación del ítem a los diseños sin intervención
Unidad	19. Unidades: 1-en grupo; 2-individual	31. Unidades de intervención: 0-individual; 1-grupal	31. Unidades de intervención **o de registro**: 1-individual; 2-grupal	**Incorporación parcial** por generalización a diseños sin intervención **Consecuencias:** Adaptación del ítem a los diseños sin intervención
Área formativa	12. Contexto de intervención: 1-urbano; 2-rural; 3-mixto	**32. Área formativa**	32. Área formativa	**Modificación terminológica y eliminación parcial:** se pasó de hablar de "contexto" a "área formativa" y de opción múltiple a formato abierto. **Consecuencia:** concreción en el ámbito sustantivo y mayor especificación.
Campo de intervención	13. Campo de intervención: 1-sanitario; 2-educacional; 3-social; 4-clínico; 5-organizacional; 6-otros	33. Campo de intervención **(destinatarios)**	33. Campo de intervención (destinatarios)	**Eliminación parcial:** se pasó de opción múltiple a formato abierto. **Consecuencia:** mayor especificación; posibilidad de adaptación a distintos contextos
Tipo de publicación	1. Tipo de publicación: 1-revista, 2-libro, 3-tesis, 4-congreso, 5-otros	34. Tipo de publicación: 1-Revista; 2-Libro; 3-Tesis; 4-Congreso; 5-Otros	34. Tipo de publicación: 1-Revista; 2-Libro; 3-Tesis; 4-Congreso; **5-Otras publicaciones; 6-Trabajos no publicados**	**Modificación parcial por desagregación:** la última opción "otros" fue dividida en dos: "otras publicaciones" y "trabajos no publicados" Consecuencia: mayor especificación

Año de publicación	**2. Año de publicación**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Índice de impacto	**3. Índice de impacto (en revistas)**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Base de datos	**4. Base de datos (especificar)**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Entrenamiento de los investigadores	**5. Entrenamiento de los investigadores: 1-especificado; 2-no hay datos suficientes**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Estructura APA	**6. Estructura del artículo recomendada por la APA**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables

Edad, desviación típica	**9. Edad (desviación típica)**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Origen cultural	**10. Origen cultural: 1-sólo uno; 2-más de uno; 3-no hay datos suficientes**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Nivel socioeconómico	**11. Nivel socioeconómico: 1-bajo; 2-medio; 3-alto**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
País	**14. País**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Orientación teórica	**15. Orientación teórica: 1-especificada; 2-inferida; 3-no hay datos suficientes**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables

Evidencia empírica previa	**16. Evidencia empírica previa: 1-especificada; 2-no hay datos suficientes**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Discusión	**20. Los puntos fuertes y débiles son discutidos: 1-no; 2-sí**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Cálculo del tamaño de la muestra	**25. Cálculo estadístico del tamaño de la muestra (magnitud del error): 1-sí; 2-no**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Sin mortalidad	**27. Sin mortalidad: 1-no; 2-sí**	---	---	**Fusión** con el ítem referido a la "mortalidad experimental" (número 26 del cuestionario exploratorio para el estudio de validez de contenido) **Conclusiones**: simplificación de la escala por fusión de los ítems que se refieren al mismo constructo
Periodo de línea base	**30. Periodo de línea base: 1-< 5 meses; 2- 6-11 meses; 3-> 12 meses**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables

Intervalos de confianza	**39. Especificación de los intervalos de confianza en los análisis estadísticos: 1-no; 2-sí**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Efectividad	**41. Otros datos además de los objetivos marcados: 1-efectos positivos; 2-efectos negativos; 3-ambos; 4-ninguno**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Interpretación de los resultados	**42. Interpretación de los resultados: 1-todos; 2-algunos de ellos; 3-ninguno**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables
Interpretación de los sesgos en los resultados	**43. Interpretación de los sesgos en los resultados: 1-todos; 2-algunos de ellos; 3-ninguno**	---	---	**Eliminación total** por incumplimiento del criterio de inclusión (estudio de validez de contenido) **Conclusiones**: simplificación de la escala por eliminación de los ítems menos representativos, útiles y/o viables

Tabla 2.7. Evolución acaecida hasta llegar a la versión integradora de la escala de calidad.

Como puede apreciarse echando un vistazo a la última columna de la tabla 2.7, las principales razones por las que introdujeron cambios fueron las siguientes:

- **Eliminación**: consistió en la exclusión de ítems o categorías anteriormente presentados. Generalmente, con estas eliminaciones se consiguió hacer de la escala un instrumento más concreto y preciso, en tanto se descartó aquello que no fue considerado representativo, útil y/o viable; en definitiva, aquello que no aportaría información relevante. La eliminación pudo ser:

    - **Total** (de un ítem completo): fueron aquellos casos en los que el ítem fue descartado en su totalidad.

    - **Parcial** (de alguna categoría): se trataron de aquellos casos en que no se eliminó un ítem al completo, sino sólo alguna categoría (el ítem se mantendría, con diferente número de opciones).

- **Incorporación**: supuso incluir ítems o categorías que no aparecían al principio. Generalmente, con esto se consiguió mayor concreción de los constructos en estudio, mayor utilidad al incorporarse aspectos relevantes anteriormente omitidos y, por tanto, lograr un instrumento más completo. La incorporación pudo ser:

    - **Total** (de un ítem completo): se incluyó un ítem completo que no aparecía al principio.

    - **Parcial** (de alguna categoría): se incluyó alguna categoría que anteriormente no se daba.

- **Modificación**: supuso el cambio de algo que ya estaba incluido, ya fuera la redacción de un ítem o de alguna categoría dentro de éste. Generalmente, con esto se logró la adaptación de la escala para su aplicación en todo tipo de diseños (anteriormente, estaba principalmente pensada para su aplicación a diseños con intervención, especialmente experimentales); mayor concreción al definir más claramente y de manera más generalizada los conceptos; en consecuencia, mayor operacionalización; y más utilidad por haberse logrado un uso más generalizado. Fundamentalmente, se dieron tres tipos de modificaciones:

    - **Cambio terminológico**: consistió en el cambio de un concepto por otro, generalmente para lograr la generalización del uso de los ítems en todo tipo de diseño.

    - **Fusión/desagregación**: consistió en unir varios ítems en uno o, al contrario, formar de un solo ítem varios.

    - **Cambio en puntos de corte**: supuso la modificación de los intervalos de algunos ítems cuantitativos, generalmente para lograr una información más útil, a la vez de una codificación sencilla.

A continuación se concreta, para cada ítem que fue modificado, las razones de dicho cambio y las consecuencias que acarrearon.

- En primer lugar, cabe destacar el hecho de que lo que en la escala depurada se codificó como "categoría 9, valor perdido", en la versión integradora se cambió por dos posibles opciones: "**8**, no se especifica el dato" y "**9**, no aplicable", con lo que se especificó en mayor grado la causa por la que no pudo recogerse el dato concreto.

- El **ítem 0**, de carácter descriptivo referido al **tipo de estudio**, se incluyó por facilitar el trabajo de codificación tras una primera lectura; para mayor claridad, se crearon diferentes modalidades de la escala de calidad, pensando en utilizar una u otra en función del tipo de estudio (ver anexo VII, pág. xxxvii). En definitiva, la idea de este primer ítem 0 fue la de clasificar cada trabajo según el tipo de diseño tras una primera lectura superficial y, a partir de ahí, decidir qué modalidad de escala utilizar para su codificación.

- El **ítem 1**, referido al **grupo de comparación**, se incluyó después del estudio de validez de contenido, al completar el resultado con otras escalas que previamente no estaban disponibles. Tras su inclusión, se modificó para generalizar su uso en todos los tipos de diseño: en primer lugar, se incluyó una tercera categoría para aquellos casos en que no había grupo de comparación; y, además, se amplió la categoría que puntuaba uno para que se incluyeran también los casos en que no había intervención, ya que en estos casos no es posible distinguir entre un control inactivo o activo, puesto que en ningún grupo se introducen cambios en la vida cotidiana de las personas que son estudiadas.

- Las modificaciones realizadas en el **ítem 3**, referido al **azar**, responden a la necesidad de no considerar los diseños experimentales como los de mayor calidad sin estudiar mínimamente si el proceso de aleatorización fue o no adecuado (León y Pedromingo, 2001). Por otro lado, se considera que si no se utiliza azar pero el método de selección y asignación es conocido, el diseño resultante será de mayor calidad que cuando se desconoce. Finalmente, para que los estudios en los que sólo había un grupo también pudieran ser puntuados, se tuvo en cuenta no sólo la asignación, sino también el método de selección de la muestra. De este modo, el resultado fue pasar de tres opciones en que se graduaba la asignación aleatoria a dos opciones, donde se valoraron de igual modo los diseños experimentales con selección y asignación aleatorias adecuados, los pre-experimentales y cuasiexperimentales con control de variables extrañas y los diseños observacionales o de encuestas con selección aleatoria de la muestra.

- El **ítem 4** referido al **diseño**, en la primera versión tenía 4 opciones: 1. Pre-experimental (solo un grupo o una medida); 2. Cuasiexperimental (dos grupos sin asignación aleatoria) o grupos no equivalentes con pretest y postest; 3. Experimental; aleatorio; y 4. Otros (estudios observacionales, naturalistas y de encuestas). Sin embargo, las 4 opciones elaboradas posteriormente para el mismo ítem fueron distintas: 1. Pre-experimental, 1 ó 2 medidas cuando no hay intervención y experimental con un solo momento de medida; 2. Cuasiexperimental (pre-post con grupo control no equivalente) entre 2 y 29; observaciones y encuestas con medidas entre 3 y 29; 3. Series temporales; 4. Experimental con al menos dos momentos de medida, diseño de discontinuidad en la regresión y observaciones o encuestas con más de 30 medidas. Estas modificaciones se realizaron con la intención de que la calidad del diseño fuera puntuada gradualmente de 0 a 1, independientemente del tipo de diseño que tuviera el estudio. Las **observaciones y**

encuestas fueron consideradas de mayor calidad a medida que aumentaba el número de medidas; este mismo criterio se siguió para graduar la calidad de los **diseños experimentales**; respecto a los **diseños cuasiexperimentales**, los de discontinuidad a la regresión fueron los más valorados, seguidos de los de series temporales, los cuasi (al menos dos momentos con grupo control no equivalente) y, finalmente, los pre-experimentales fueron los valorados en nulo (siguiendo varios criterios, concretamente, el número de grupos, el número de medidas y el conocimiento o no del método de asignación).

- En el **ítem 5**, referido al tamaño de la **muestra**, se mantuvo con tres opciones, pero los puntos de corte variaron tratando así de lograr mayor sensibilidad, más información; anteriormente, los puntos de corte fueron 5 y 10; posteriormente, pasaron a 12 y 40, debido a que en muy pocas ocasiones el número de personas era menor que 5 y en la mayoría de las ocasiones era mayor a 10, por lo que prácticamente todos los casos se encuadraban en la tercera opción (mayor o igual que 10); poniendo el primer punto de corte en 12 y el segundo en 40, se consiguió que los casos se distribuyeran entre los tres intervalos. Realmente, la creación de intervalos óptimos para este ítem variará en función del contexto de intervención que se esté tratando; una posible solución sería anotar los valores concretos de la muestra y crear los intervalos a posteriori.

- En el **ítem 6**, referido a la **mortalidad global**, también se modificaron los puntos de corte con la intención de hacer más informativa la recogida del dato: en muy pocas ocasiones la mortalidad experimental era mayor del 30%, por lo que se bajó el primer punto de corte al 20%; además, se dio la máxima puntuación a aquellas ocasiones en que no había mortalidad experimental. Al igual que se comentó en el ítem anterior, el punto de corte adecuado variará según el contexto de intervención, por lo que un procedimiento a seguir sería anotar el porcentaje concreto en cada caso y crear los intervalos a posteriori.

- En el **ítem 7**, referido a la **mortalidad diferencial**, se pasó de valores dicotómicos (sí o no homogeneidad) a tres intervalos para lograr así una mayor concreción y operacionalización.

- El **ítem 8**, referido a las **exclusiones posteriores**, pasó por dos transformaciones: en primer lugar, hubo un cambio terminológico con la intención de generalizar la aplicación de este ítem a casos en que la agrupación de la muestra no se realizaba aleatoriamente, de tal manera que se pasó de decir "exclusiones posteriores a la asignación aleatoria" a decir "posteriores a la agrupación de la muestra"; por otro lado, se pasó de opciones dicotómicas (si/no) a tres intervalos, con lo que se aumentó la concreción de la información obtenida.

- El **ítem 9**, referido al **seguimiento** tuvo, al igual que otros ítems anteriores, una modificación en sus puntos de corte para ganar en concreción y especificación; así, por ejemplo, antes de la modificación entraban en la misma categoría un estudio con cuatro meses de seguimiento y otro sin seguimiento; tras la modificación, sin embargo, puntuaría más bajo la segunda situación.

- En el **ítem 10**, referido a los **momentos de medida**, se hizo una inclusión en cada una de las categorías para adaptarlo así a las situaciones en que no se daba

intervención. Así, anteriormente se diferenció entre "posterior" y "previo y posterior"; más tarde, se incluyó a la categoría posterior "o una medida cuando no hay intervención", y a la categoría de previo y posterior, "o más de una medida si no hay intervención".

- Al **ítem 11**, referido a las **medidas que aparecen en los distintos momentos de registro**, por un lado se le modificaron algunas opciones para lograr mayor concreción; así, por ejemplo, se eliminaron términos poco específicos como "algunos" por valores concretos; por otro lado, se le modificó la terminología usada para así adaptarlo a los diseños sin intervención: como hablar de pre-test y pos-test implicaba la existencia de una intervención como punto de separación entre el antes y el después, se optó por cambiar esta nomenclatura por "medidas que aparecen en todos los momentos de registro", con lo que ya no necesariamente tenía que haber intervención; por último, se especificó que un estudio puntuaría en la categoría intermedia siempre y cuando la variable que no aparece en todos los momentos no fuera la única, en cuyo caso puntuaría "0".

- El **ítem 13**, que hace referencia al **enmascaramiento del evaluador**, se incorporó tras el estudio de validez de contenido pues, proveniente de escalas que previamente no se encontraron disponibles, fue considerado de interés ya que recogía información acerca del enmascaramiento del evaluador, diferenciando así a esta figura de quien realiza la intervención; en ocasiones, ambas funciones pueden recaer en la misma persona o grupo de personas, pero no siempre tiene que ser así.

- Los **ítems 14 y 15**, referidos al **enmascaramiento del usuario y del profesional** respectivamente, previamente estaban fusionados en uno solo. De este modo, se hizo un estudio diferenciado del enmascaramiento de estas dos figuras y del evaluador, previamente comentado. Además, al ítem 15 se le añadió la posibilidad de codificar el enmascaramiento del evaluador interno cuando no había intervención, haciendo así que el ítem fuera codificable ante cualquier tipo de diseño.

- El **ítem 16**, referido a la **homogeneidad**, sufrió dos modificaciones en dos momentos distintos: en primer lugar, se incluyó aquellos aspectos en los que era preciso fijarse para concluir si el estudio era homogéneo para todas las personas participantes, concretamente la intensidad, duración y profesionales, con lo que se ganó en operacionalización; en segundo lugar, para lograr la aplicabilidad en todos los diseños, se concretó que la homogeneidad podía hacer referencia también al proceso de registro, con lo que se logró adaptar el ítem a las situaciones en las que no se daba intervención.

- El **ítem 17**, referido a la **definición del constructo**, pasó por dos modificaciones para tratar de conseguir mayor especificación y operacionalización. Así, se pasó de explicitar si se definía el constructo y si esta definición era o no vaga a determinar si se definía y si era o no operativa, para finalmente distinguir entre la definición conceptual y la empírica y, a partir de ahí, hacer algunas posibles combinaciones.

- En el **ítem 19**, referido al **tamaño de efecto**, en principio sólo se recogía el valor concreto; posteriormente, además, se codificó de manera dicotómica (si se daba o

no), para tener la posibilidad de contabilizar el resultado de 0 a 1 y así poder incluirlo en el cálculo del índice de calidad, ítem que a continuación se comenta.

- El **ítem 20**, referido al **índice de calidad**, se incluyó después de obtener los resultados del estudio de validez de contenido con la intención de calcular un índice cuantitativo que diera una visión general del nivel de calidad que aportaba cada estudio.

Es preciso recordar que los ítems que a partir de ahora se presentan son únicamente descriptivos; no forman parte del cálculo del índice de calidad, pero aportan información interesante a nivel sustantivo. Hay que tener en cuenta, por tanto, que en función del problema que se esté tratando, quizá podrían eliminarse ítems que aquí aparecen e incluirse otros que no se encuentran. Sea como sea, a continuación se siguen comentando los cambios que presentaron cada uno de los ítems y las razones de tales modificaciones.

- Los **ítems 21** y **23**, referidos al **índice estadístico calculado y** al **índice de variabilidad** respectivamente, se aportaron después de haber realizado el estudio de validez de contenido con la intención de ir tanteando qué estudios podrían ser incluidos en un meta-análisis y cuantos de ellos no eran útiles para ello por carecer de datos.

- El **ítem 22**, referido a la **significación de las diferencias**, se incluyó de igual manera después del estudio de validez de contenido para, de cara a la posibilidad de realizar un meta-análisis, determinar si el "sesgo de publicación" puede estar influyendo en los resultados. Este sesgo puede darse porque actualmente se tiende a publicar aquellos estudios que obtuvieron resultados significativos de tal manera que, al recopilar los estudios disponibles acerca de una temática para la realización de un meta-análisis, es posible que se esté tomando una muestra no significativa, con más resultados positivos de los que en la realidad se dan. Por otra parte, se realizó un cambio terminológico: concretamente, se sustituyó la palabra "resultados" por "medidas", por si la primera pudiera parecer demasiado dirigida a los diseños con intervención exclusivamente.

- El **ítem 24**, referente al **número de participantes que hay en cada grupo**, se incluyó también después del estudio de validez de contenido para obtener una mayor especificación de la muestra, aspecto imprescindible para estudiar la posibilidad de generalización de los resultados.

- El **ítem 25**, referido al **número de grupos**, también se incluyó a posteriori para recoger datos más concretos acerca del diseño.

- El **ítem 26**, referente a las **exclusiones que se realizaron tras medidas posteriores**, también se incluyeron después para lograr mayor especificación acerca de los datos recogidos.

- El **ítem 29**, referido al **periodo de estudio**, pasó por dos cambios. En primer lugar, se pasó del formato abierto a la inclusión de dos intervalos, con lo que se trató de agilizar su cumplimentación. Como en otras ocasiones, es interesante recordar que la elección del intervalo óptimo dependerá en gran medida del ámbito sustantivo,

por lo que es lógico tomar dicha decisión tras haber recogido la información concreta. En segundo lugar, se modificó la palabra "tratamiento" por "estudio", para así hacer aplicable este ítem a aquellos casos en los que no se daba intervención.

- El **ítem 30**, referido a la **intensidad**, se cambió en dos ocasiones: en primer lugar, se dio una definición de "intensidad del tratamiento" más acorde a todos los posibles contextos, ya que en el cuestionario con el que se realizó el estudio de validez de contenido se hablaba de "número de dosis", con una orientación claramente clínica; en segundo lugar, se incluyó la posibilidad de estudiar la "intensidad del registro" para aquellos casos en los que no había "tratamiento", con lo que se logró generalizar el uso de este ítem a los diseños sin intervención.

- Al **ítem 31**, referido a las **unidades de intervención**, se le añadió en último término la posibilidad de que las unidades fueran de registro para así adaptarlo a cualquier tipo de diseño.

- El **ítem 32**, referido al **contexto**, se cambió en dos sentidos. Por un lado, se especificó que el aspecto que se recogía iba a ser el área formativa. Esta modificación se realizó en busca de una mayor especificación, teniendo en cuenta el ámbito de estudio, por lo que no sería generalizable a otros contextos. Por otro lado, se pasó de un formato de opción múltiple a un formato abierto, con lo que se trató de recoger la respuesta de la manera más concreta posible. Esto mismo ocurrió en el **ítem 33**, referido al **campo de intervención**.

- Al **ítem 34**, referido al **tipo de publicación**, se le incluyó en último término una categoría más. Concretamente, la opción "otros" se desagregó en "otras publicaciones" y "trabajos no publicados", para distinguir así entre aquellos estudios no publicados y los que sí lo estaban.

- Los siguientes ítems del cuestionario utilizado para la validez de contenido fueron eliminados por no cumplir el criterio de inclusión a partir de los índices obtenidos en el estudio de validez de contenido, ya que no alcanzaron un índice del 0.5 en al menos dos de los tres conceptos estudiados (representatividad, utilidad y viabilidad del dato): **ítem 2** (referido al **año de publicación**), **3** (**índice de impacto**), **4** (**base de datos**), **5** (**entrenamiento de los investigadores**), **6** (**estructura del artículo recomendado por la APA**), **9** (**desviación típica de la edad**), **10** (**origen cultural**), **11** (**nivel socioeconómico**), **14** (**país**), **15** (**orientación teórica**), **16** (**evidencia empírica previa**), **20** (**discusión de puntos fuertes y débiles**), **25** (**cálculo estadístico del tamaño de la muestra**), **27** (**sin mortalidad**), **30** (**periodo de línea base**), **39** (**intervalos de confianza de los análisis estadísticos**), **41** (**otros datos además de los objetivos marcados**), **42** (**interpretación de los resultados**) y **43** (**interpretación de los sesgos en los resultados**).

7. CONCLUSIONES.

Según estudios bibliográficos realizados acerca de la temática que aquí nos ocupa, los programas de formación continua suelen presentar algunas características que podrían estar afectando negativamente a la calidad de su diseño, como puede ser la baja

especificación de cómo se realizan dichos programas o la poca operacionalización y concreción de aquello en lo que se trata de intervenir y del procedimiento que se lleva a cabo.

Para determinar empíricamente en qué medida esto es así, en este capítulo se ha tratado de aportar una escala para medir la calidad de los estudios primarios. A pesar de que son comunes este tipo de instrumentos en la medición de la calidad, esta propuesta presenta algunos **beneficios** con respecto a otras versiones encontradas en la literatura:

- Se pretendió elaborar un instrumento que no sólo aportara un **índice cuantitativo** acerca del nivel de calidad sino que, teniendo en consideración sus distintos **ítems por separado**, aportara ideas sobre qué puntos débiles se pueden estar presentando y cómo podrían ser éstos solventados. En definitiva, se intentó que el instrumento sirviera en un primer momento para evaluar, pero también que fuera útil para mejorar los programas, aportando para ello información concreta acerca de los aspectos en los que incidir para provocar dicha mejora, e incluso proponiendo ideas sobre el sentido en el que intervenir concretamente.

- Se intentó crear un instrumento que fuera **aplicable a cualquier tipo de diseño** por diversas razones, entre las que destacan las siguientes:

  - Generalmente las escalas están optimizadas para ser aplicadas con diseños experimentales, cuando éstos no suelen ser usados en los ámbitos de la psicología, la educación y la intervención social. Creando esta escala aplicable a todos los diseños, se aumentó la **utilidad** del instrumento.

  - Hay que acabar con el mito de que el **diseño**, simplemente por ser experimental, ya presentará mejor índice de calidad que un diseño de otro tipo. La aplicación de esta nueva escala quiso poner en tela de juicio esta idea preconcebida.

  - Realmente, el **límite** entre los distintos tipos de diseño puede ser muy **difuso**; así, por ejemplo, cuando una intervención se prolonga durante un largo periodo de tiempo, ¿no se está pasando a realizar un diseño de baja intervención? Puede que la intervención comenzara siendo un elemento extraño en la vida de las personas participantes pero, pasados unos meses, ¿esta intervención no pasa a formar parte de la vida cotidiana de estas personas? En este sentido, no parece justificable el hecho de puntuar como de mayor calidad unos estudios u otros tomando como criterio el "supuesto diseño" que presentan.

- Se intentó crear un instrumento que fuera **aplicable a todo tipo de estudio, independientemente de la temática sobre la que versara**. Al tratarse de un índice de calidad metodológica, se hizo posible su aplicación a distintos ámbitos, con lo que se logró aportar propuestas de mejora de las intervenciones en distintos ámbitos. Así, por ejemplo, en este capítulo se describió un estudio exploratorio aplicado a los ámbitos psicológico, social y educacional; pero, además, se aplicó el instrumento al ámbito de intervención con personas mayores (Sanduvete, 2004) y, en el próximo capítulo, se presentará su aplicación en formación continua.

El **procedimiento** concreto que se siguió para elaborar esta escala se caracterizó por ser inductivo-deductivo ya que se fueron alternando principalmente dos fuentes de

información: por un lado, la bibliografía referida a la evaluación de la calidad y, por otro lado, los datos obtenidos con cada las distintas versiones de la escala en diferentes contextos de intervención; concretamente, constó de cuatro **fases** que a continuación se resumen:

- En la **primera fase**, se **elaboró un cuestionario** tras recopilar todos los ítems encontrados acerca de la medición de la calidad en estudios primarios en todos los documentos encontrados que versaban sobre esta temática.

  En este momento, a modo de estudio piloto, se llevó a cabo la **aplicación de la escala**. Concretamente, se recogieron resúmenes referidos a estudios aplicados en ciencias sociales, psicología y educación. Este estudio, además de dar una visión general del estado de poca especificación y grado medio de control que presentan generalmente las cuestiones psicológicas en los trabajos que se realizan, sirvió para ir definiendo más concretamente los conceptos utilizados y para hacer una exploración acerca del nivel de viabilidad de la información necesaria para responder a cada ítem.

- En la **segunda fase**, se realizó un **estudio de validez de contenido** para determinar cuáles de éstos ítems eran considerados por los expertos como representativos de la dimensión en la que se ubicaba, útiles y cuáles, según su opinión, solían estar disponibles en los informes y publicaciones. De este modo, se hizo una criba para reducir el amplio número de ítems del que constaba el cuestionario a sólo aquellos que resultaban de interés.

  La aparente importancia otorgada a las cuestiones metodológicas mostrada en los resultados de este estudio de validez de contenido puede ser contraria a lo que asiduamente se encuentra en la práctica profesional: en la mayoría de las ocasiones, los distintos profesionales desarrollan su actividad en una dinámica tan rápida y cambiante que no tienen los recursos suficientes para dedicar tiempo a los aspectos relacionados con la planificación y el diseño. Ocurre que los problemas a resolver surgen de manera imprevisible y son de tal envergadura que requieren una respuesta rápida. Se ven, de este modo, inmersos en una práctica constante donde hay poca cabida a la reflexión y a la profundización de aspectos tan fundamentales como es la elaboración de un buen diseño previo a la intervención. Esta discrepancia entre los resultados obtenidos en el estudio de validez de contenido de los ítems y la realidad profesional pone de manifiesto la falta de conexión entre el desarrollo de la "academia" y el ejercicio profesional.

  De todos modos, la mayor importancia dada a los ítems asignados al dominio de las características metodológicas puede tener una doble explicación: en primer lugar, puede haber habido un sesgo en la muestra de expertos, inclinada a temas metodológicos; pero también es cierto que difícilmente se pueden generalizar elementos de calidad en el terreno de las dimensiones extrínsecas y sustantivas, ya que suelen venir determinadas por el ámbito de intervención concreto (Chacón, Sánchez-Meca, Sanduvete y Alarcón, 2006).

- La **tercera fase** sirvió básicamente para detallar más la definición de los ítems encontrados y para realizar algunas otras inclusiones propuestas por expertos. De

este modo, se trató de que ningún ítem útil para medir la calidad de los estudios primarios quedara sin contemplarse en el instrumento elaborado.

Llegados a este punto, se desarrolló una **nueva aplicación de la escala**, donde se tomaron los textos completos que versaban sobre la formación continua. Los resultados obtenidos siguieron en la línea de lo encontrado en el primer estudio exploratorio realizado con los estudios psicológicos en general (poca especificación y grado medio de control y estandarización). Además, se detectó el sesgo que presentaba el instrumento a favor de los diseños experimentales y se vio que algunos ítems provocaban dudas a la hora de realizar la codificación. Para solventar estas deficiencias, se pasó a la cuarta e integradora versión de la escala.

- La propuesta de versión integradora de la escala para la medición de la calidad en estudios primarios obtenida en esta **cuarta fase** presentó diferencias respecto a la versión anterior que se centraron en la búsqueda de mayor concreción y operacionalización y en la necesidad de que todos los ítems fueran aplicables a los distintos tipos de diseño para que los resultados obtenidos fueran comparables (para que el valor del índice de calidad pudiera oscilar de 0 a 19 indistintamente del tipo de diseño que presentara el estudio).

Esta última versión de la escala no se considera como algo cerrado y finalizado, sino que se seguirá perfilando en el futuro, pues actualmente presenta ciertas **debilidades** que habrían de ser solventadas:

- Aún no se ha probado en profundidad la capacidad de **generalización** del instrumento. Se intentará en el futuro realizar nuevas aplicaciones a programas en distintos ámbitos de intervención y hacer los ajustes necesarios para que su uso sea recomendable en cualquier contexto.

- No se ha comprobado las diferencias existentes entre los resultados encontrados con esta versión de la escala y otras escalas encontradas en la literatura. En el futuro se realizará un estudio de **validez convergente**.

Esta versión integradora de la escala se utiliza para determinar en el próximo capítulo, de manera exploratoria, el grado de calidad metodológica que presentan los estudios relacionados con la formación continua, concretar las características principalmente metodológicas que suelen presentar estos trabajos y, a partir de esta información, proponer mejoras.