DISEÑO DE CUESTIONARIOS DE CONTEXTO PARA LA EVALUACIÓN DE SISTEMAS EDUCATIVOS: OPTIMIZACIÓN DE LA MEDIDA DE CONSTRUCTOS COMPLEJOS

Context questionnaire design for the evaluation of educational systems: optimization of complex constructs measurement

JESÚS M. JORNET, JOSÉ GONZÁLEZ SUCH Y Mª JESÚS PERALES
Universitat de València

En este artículo se realiza una revisión acerca de un posible modelo de diseño y desarrollo de cuestionarios de contexto para la evaluación de sistemas educativos. Se revisan las fases generales, a la par que se identifica la posible composición de indicadores (simples y complejos) de dichos instrumentos. En el diseño de cuestionarios de contexto hay que considerar diversos aspectos, entre ellos son clave: a) la selección de variables e indicadores que se deben incluir como parte del modelo y b) el modo en que se evalúan los indicadores complejos, es decir, aquellos que son resultado de una escala o de la combinación de indicadores simples (representados por un solo ítem). Para evaluar indicadores complejos (por ejemplo, clima social del aula) existen escalas con buena validez de constructo; sin embargo, no pueden utilizarse completamente, al estar compuestas por un elevado número de ítems. En este último caso, incluimos un procedimiento que hemos desarrollado para diseñar micro-instrumentos que estén compuestos por los mejores ítems de la escala (es decir, los que mejor predicen la puntuación total, manteniendo la estructura dimensional de la escala). Estos micro-instrumentos deben ser capaces de representar indicadores complejos para que puedan ser incluidos en los cuestionarios de contexto. Se trata de un procedimiento de reducción de escalas, dirigido a identificar los elementos o ítems que mejor predicen la puntuación global de una escala, pero manteniendo su estructura teórica.

Palabras clave: *Evaluación de sistemas educativos, Cuestionarios de contexto, Metodología de diseño y construcción de instrumentos, Indicadores educativos.*

Introducción

Las evaluaciones de sistemas educativos, sean las que se realizan en cada país o las que se desarrollan a nivel internacional (como, por ejemplo, los proyectos PISA, TIMSS, PIRLS…) se han convertido en una línea de investigación y análisis de gran relieve, con amplia repercusión social y con capacidad de influencia en las políticas educativas a nivel local, nacional e internacional.

Fecha de recepción: 8-2-2012 • Fecha de aceptación: 7-3-2012

Los elementos considerados clave en estas evaluaciones, sobre los que se han centrado los esfuerzos técnicos y las discusiones políticas, han sido las pruebas estandarizadas de rendimiento. Estas pruebas son las que permiten ofrecer información sobre los niveles de rendimiento en cada una de las áreas fundamentales de contenido para cada una de los países o regiones que participan en las evaluaciones.

La aplicación de estas pruebas de rendimiento se acompaña de los llamados *cuestionarios de contexto*. Son procedimientos de recogida de información sobre los sujetos, grupos y centros a los que se aplican las pruebas y en algunos casos sobre el propio proceso de enseñanza aprendizaje. La información que aportan, analizada en paralelo con las pruebas de rendimiento, puede resultar de gran utilidad para explicar los resultados obtenidos, y, por tanto, para sustentar las decisiones sobre políticas educativas. Sin embargo, como señalábamos en De la Orden y Jornet (2012), existe un problema fundamental que está en la base de la limitada utilidad de estos análisis: el diseño deficiente de los denominados cuestionarios de contexto.

En los últimos años, los institutos y agencias de evaluación que realizan evaluaciones de sistemas educativos han incrementado su interés y atención por ajustar mejor este tipo de instrumentos. Sin embargo, dichos esfuerzos generalmente se han basado en la identificación de los reactivos que mostraban asociaciones significativas para explicar el desempeño (Willms, 2006; Backhoff *et al*., 2008; Miranda, 2008; Murillo, 2009), aportando, sin duda, informaciones de gran valor. No obstante, se ha priorizado el análisis empírico de la utilidad de los reactivos como indicadores de factores asociados al desempeño, sin realizar una incidencia especial en las bases teóricas o racionalidad de elaboración de los cuestionarios como instrumento de medida.

En consecuencia, se han producido innovaciones, en ocasiones escasas y en cualquier caso parciales, que han mejorado el rol de este tipo de instrumentos en el conjunto de la investigación educativa en general y de los planes de evaluación en particular, pero que no han llegado a satisfacer las expectativas acerca de su utilidad global en las evaluaciones de sistemas. De este modo, las mejoras han sido menores que los esfuerzos realizados. El hecho de apoyar la mejora principalmente sobre análisis estadísticos más potentes de la información disponible, normalmente de carácter causal y multinivel, aunque haya proporcionado indudables avances probablemente no sea el único camino para mejorar el uso de la información evaluativa.

Desde nuestra perspectiva, y tal como se indica en De la Orden y Jornet (2012), el incremento de la utilidad de este tipo de informaciones debe apoyarse en un mejor diseño de los instrumentos y elaborarlos con la atención que merecen: desde la descripción del modelo explicativo de referencia para dar respuesta al concepto de calidad que pretende evaluarse, pasando por la definición de los constructos implicados, hasta la selección de los indicadores que los hacen operativos en forma de medida.

En este trabajo presentamos una propuesta de diseño y desarrollo de cuestionarios de contexto. El origen de esta línea de investigación se sitúa en el Proyecto *Análisis de Variables de Contexto: Diseño de cuestionarios de contexto para la evaluación de sistemas educativos* (Proyecto AVACO, I+D+I, 2006-2008. Código SEJ 2005-05 923 —financiado por el MICINN—), y que ha sido comprobado y validado en el Proyecto *Modelos de Análisis de Variables de Contexto* (Proyecto M-AVACO, I+D+I. 2009-2012. Código EDU 2009-13485 —asimismo financiado por el MICINN—)[1]. La propuesta incluye dos grandes conjuntos de acciones:

a) Proceso general de diseño y desarrollo de los cuestionarios de contexto para la evaluación de sistemas educativos.
b) Optimización de la medida de constructos complejos para el diseño de micro-instrumentos que integrar en dichos cuestionarios.

La presentación de la propuesta se estructura en los dos ámbitos de trabajo señalados, con el fin de ofrecer una muestra global y a ser posible coherente del tipo de procesos que pueden desarrollarse.

Componentes para el diseño de cuestionarios de contexto para la evaluación de sistemas educativos

Cuando nos referimos a cuestionarios de contexto, lo estamos haciendo en relación a un sistema de instrumentos que pueden ir dirigidos a diferentes fuentes. Es decir, no se trata de un instrumento aislado, sino de varios que se dirigen de forma integrada a recabar información a partir de diferentes audiencias o partes interesadas y/o implicadas en la evaluación y que solo puede ser recogida a partir de ellas.

Los cuestionarios de contexto tienen sentido en el marco de las evaluaciones de sistema, porque a partir de ellos se supone que puede abordarse la explicación del rendimiento. Por tanto, en primer lugar es necesario definir los elementos que definen la evaluación de sistema (fases 1 y 2) para después, de una forma coherente con estas decisiones, diseñar los mencionados cuestionarios de contexto.

Fase 1. Definir el concepto de calidad por evaluar

El diseño de los cuestionarios de contexto está condicionado por el sentido global de la evaluación de sistemas en que se enmarca. Así, el elemento inicial que debe orientar todo el proceso es la definición del concepto de calidad que se desea comprobar. La calidad, como tal, es un constructo teórico que debe ser definido como origen del diseño del plan de evaluación. En este sentido, en De la Orden y Jornet (2012) hacíamos referencia a diferentes enfoques de este concepto. Cada enfoque orientará a seleccionar —o priorizar— diferentes tipos de información. Por este motivo, es necesario basarse en un modelo de referencia, de carácter sistémico —como, por ejemplo, el descrito por De la Orden (1997, 2007)—, para identificar de manera adecuada qué efectos se desean evaluar, el tipo de informaciones que se requieren y orientar de manera precisa la planificación de la evaluación.

Asimismo, hay que tener en cuenta la utilidad que se persigue en la evaluación: a) si se trata de poder coadyuvar a la orientación de políticas educativas o socio-educativas, a nivel de macro-sistema, o b) si se trata de extraer información que sea útil para otros niveles de intervención, como por ejemplo, el diseño curricular, las instituciones escolares o la organización escolar (Lukas y Santiago, 2004), los procesos de enseñanza-aprendizaje que hay que desarrollar en las aulas o el papel de la comunidad y/o las familias en el proceso educativo (Cardona, Perales y Gómez-Costa, 2009). Una evaluación que pueda dar información útil para todos los propósitos, con toda seguridad es un *desiderátum* si se mantienen los modos de hacer evaluación de sistemas que se dan en la actualidad. Los niveles de análisis y de intervención, necesariamente diferenciales, requieren de información asimismo diferencial[2]. Ello debe contemplarse inicialmente en la definición del concepto de calidad a evaluar, y el nivel de intervención en el que se debe impactar con la evaluación. En definitiva se trata de responder a una doble cuestión: *¿para qué y a quién debe servir la información evaluativa?*

De este modo, y en síntesis, entendemos que no pueden diseñarse de igual manera evaluaciones que pretendan comprobar un tipo de calidad u otro, ni que pretendan extraer información útil para diferentes niveles de intervención (nacionales o transnacionales; macro-analíticos o micro-analíticos; internos al sistema —instituciones educativas y/o aulas—, o externos al mismo —por ejemplo, el papel de las familias en el acompañamiento del proceso de aprendizaje de sus hijos/as—). Por ello es muy importante considerar las necesidades de información que puedan tener, según el plan de evaluación diseñado,

las diferentes partes involucradas en la evaluación. Ello ayuda a especificar las preguntas de evaluación a que debe responderse y orienta de forma más concreta la fase siguiente. La definición precisa del plan de evaluación está en la base de la validez última que pueda exigirse al mismo.

Fase 2. Definir los elementos del plan de evaluación

Definido el concepto de calidad que se desea evaluar y los efectos en que se concreta[3], en esta segunda fase se trata de hacerlos operativos como elementos susceptibles de medición/evaluación. En este caso, nos referimos a determinar variables e indicadores que formarán parte del plan de evaluación y, en concreto, del sistema de cuestionarios de contexto, así como las fuentes de recogida de información.

En primer lugar hay que tener en cuenta que deben diferenciarse *variables e indicadores*. Las variables, como es obvio, hacen referencia a unidades de información que poseen variabilidad, mientras que consideramos indicadores a aquellas variables cuya variación es concomitante de manera sistemática con el fenómeno global sobre el que se pretende informar, bien de manera individual, bien por su relación con otros indicadores. Entre los indicadores, tal como puede observarse en la figura 1, diferenciamos entre:

- *Indicadores simples*: aquellos que están representados por un único reactivo.
- *Indicadores complejos*: aquellos que provienen de alguna combinación de varias informaciones, sean ratios o síntesis numéricas de un conjunto de variables y/o indicadores, o resultados de una escala que mide un constructo teórico complejo (como por ejemplo, clima social del aula, auto-concepto, o metodología docente).

Para la selección de variables e indicadores, es necesario tener cuenta los siguientes criterios:

a) *Sobre la adecuación y el rol de la información en el plan de evaluación*:

   - Pertinencia y relevancia de la información respecto al objeto y finalidad de la evaluación
   - Rol de la información dentro del plan de evaluación. Pueden identificarse dos vectores de clasificación:
     - Información descriptiva/explicativa, en función del uso final que se vaya a realizar de la información
     - Tipología de información: contexto, entrada, proceso y producto (ver figura 2 y De la Orden y Jornet, 2012).

b) *Sobre la calidad de la información*:

   - La *evaluabilidad* hace referencia al grado en que la información que se va a recoger puede ser entendida en términos de elementos observables —o inferibles a partir de comunicaciones verbales— (al menos extraíble a partir de percepciones, opiniones, actitudes, intereses…).
   - La *interpretabilidad* se refiere al grado en que pueda ser interpretada de manera clara, sea por procedimientos cuantitativos o cualitativos.
   - Los *criterios de bondad*, que se refieren a la calidad métrica (o fáctica, relativa a su representatividad) con que los instrumentos o técnicas utilizadas para recabar la información ofrecen garantías, como la *fiabilidad y la validez*.

En segundo lugar, se trata de *identificar cuáles son los mejores informantes* —o las mejores fuentes de información— para recabar la información requerida. En la figura 3 se presenta un posible esquema de decisión para la selección de fuentes o agentes de información. En la tabla 1 se muestra un resumen que estructura la información, según las fuentes desde las que se va a recoger.

FIGURA 1. Esquematización del proceso general de diseño de cuestionarios de contexto

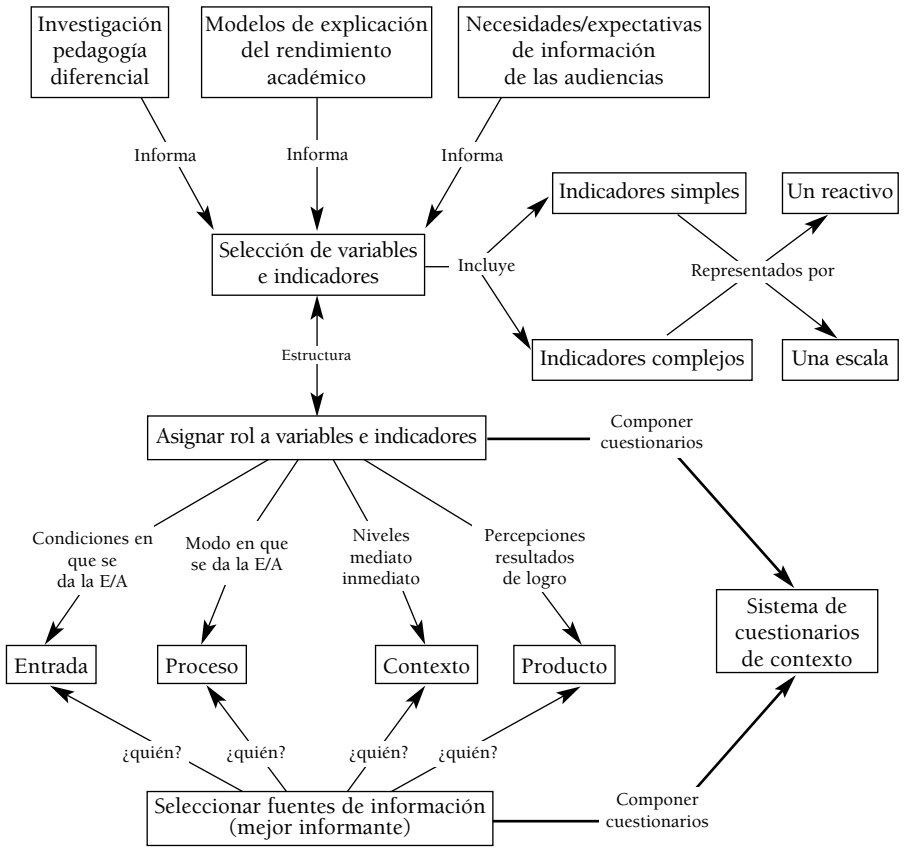
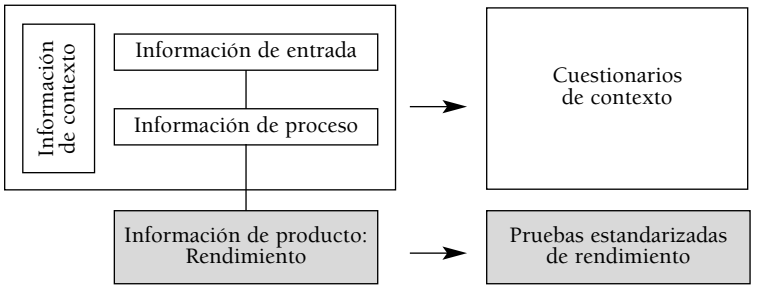


FIGURA 2. Rol de la información e instrumentos para recogerla



Los criterios que se pueden considerar para la selección de la fuente de información se sintetizan en los siguientes:

a) *Objetividad*. Es necesario asegurar que la información recogida proviene de la fuente o agente que tiene una mejor posición para aportarla. Las informaciones de entrada, proceso y contexto son más susceptibles de subjetividad, por lo que cabe extremar las precauciones cuando seleccionamos la fuente o agente más adecuado. En muchas evaluaciones de sistemas educativos se incluyen en los cuestionarios de contexto informaciones que, cada vez con más frecuencia, pueden formar parte de bancos de información (por ejemplo, estudios o situación laboral de padres/madres, pueden formar parte de los registros escolares). Es posible, por tanto, extraerlas de los cuestionarios de contexto y no hace falta preguntarles a los alumnos que en muchos casos no disponen de la certeza suficiente acerca de los estudios que poseen sus progenitores o de su situación laboral. Otro ejemplo que puede aportarse en esta misma línea es el modo en que diseña la programación el profesorado. Normalmente se incluyen cuestiones al respecto en los cuestionarios dirigidos a docentes, cuando sería más objetiva la información que puede aportar un observador externo (por ejemplo, la inspección educativa), tras la observación de las evidencias documentales que pueda aportar el profesorado acerca de la programación. Ejemplos como estos pueden aportarse muchos. Por ello, entendemos que es necesario un esfuerzo previo para orientar los sistemas de cuestionarios estrictamente hacia las informaciones en las que cada agente sea la mejor opción.

Figura 3. Esquema de decisión para la selección de fuentes y/o agentes de información

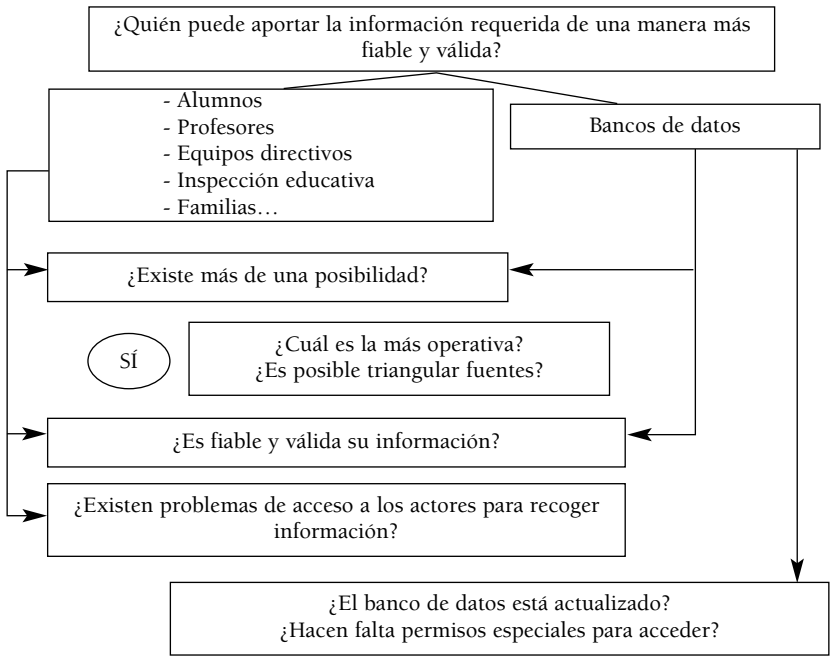


TABLA 1. Ejemplo de tabla de estructuración de información

Síntesis de variables/indicadores y fuentes de información			Fuentes de información					
Tipo de indicador		**Indicadores**	**Alumnos**	**Profesores**	**Director**	**Observador externo**	**Familias**	**Otras fuentes**
Entrada	Características de alumnos	X		X			X	
	Características de los profesores		X	X				
	Sistema de selección de alumnos			X	X			
	Infraestructura (aulas teóricas, prácticas…)	X	X	X	X			
	Medios disponibles en las aulas	X	X		X			
Proceso	Actuación del profesorado (aula)	X		X	X			
	Actuación del profesorado (tutorías)			X	X	X		
	Metodología didáctica	X	X	X				
	… … …							
	Producto … … …							
	Contexto … … …							

b) *Validación/ triangulación*. En cualquier caso, el proceso de enseñanza-aprendizaje —y en general la educación— se produce en una relación dialéctica entre diversos actores (profesorado, alumnado, equipos directivos, inspección educativa, familias…). Por ello, es preciso identificar las informaciones en las que, incluso habiendo un *mejor* informante, sea conveniente corroborar la información a partir de la aportada por otros actores. Como principio general, asumimos que siempre que sea posible es necesario verificar la información del fenómeno evaluado a partir de la concurrencia/divergencia de las informaciones aportadas por los diferentes agentes involucrados.

c) *Nivel de análisis*. Un aspecto primordial que no es frecuente considerar es el nivel de análisis sobre el que se pretende tratar la información recogida (*macro-analítico vs. micro-analítico*): alumnado y/o profesorado, aula, escuela o institución educativa, zona geo-política o administrativa (estado o comunidad autónoma, provincia, nación…). Esta identificación debe ser coherente con el plan de evaluación global (objeto y finalidad) y coadyuva a determinar también las características de los instrumentos de recogida de información que se pretende utilizar, así como los análisis que pueden ser pertinentes para su tratamiento. Las decisiones sobre el nivel de análisis pretendido se traducen en el diseño (tamaño y estructuración) de la muestra, pues en función del nivel sobre el que se pretenda dar información habrá unos u otros requerimientos para su configuración (De la Orden y Jornet, 2012).

En las evaluaciones de sistemas, por su carácter muestral y por el tipo de instrumentos utilizados, es difícil que se den las condiciones técnicas necesarias para poder interpretar la información a nivel individual, en ocasiones incluso hay razones técnicas suficientes para que *no se interpreten* a nivel de aula y escuela, y habitualmente los niveles asumibles son los que podríamos denominar macro-analíticos (como por ejemplo, basados en variables demográficas: región geo-política, variables de estratificación, como por ejemplo, las demográficas de clasificación de colectivos o grupos —sexo, titularidad del centro, etc.—). Es fundamental que los análisis que se realicen sean respetuosos con esto, y coherentes con el diseño del plan de evaluación y del muestreo, evitando ofrecer una información que, en realidad, el estudio no permite dar con garantías, o que incluso ni se lo haya propuesto como finalidad.

Fase 3. Diseñar los cuestionarios de contexto

En esta fase nos referimos de manera exclusiva al diseño de los sistemas de cuestionarios de contexto. Definido y concretado el plan de evaluación en las fases 1 y 2, se trata ahora de configurar los instrumentos específicos para cada uno de los colectivos implicados o fuentes (por ejemplo, alumnado, profesorado, equipos directivos…). Podemos identificar las siguientes etapas:

a) El punto de partida es la *figura de estructuración de información* para cada audiencia o colectivo implicado en la evaluación, mostrada en la tabla 1, completándolo en cada caso con el tipo de indicador (simple/complejo). Leyendo esta figura por columnas, se identifica la síntesis de información que hay que recoger a partir de cada uno de los colectivos o fuentes.
b) La batería de cuestionarios incluirá uno por fuente o colectivo. Cada uno de ellos integrará, finalmente, los indicadores simples (representados por un reactivo) y los complejos. Estos últimos estarán representados por micro-instrumentos desarrollados por el proceso de reducción —u optimización de la medida— que describiremos en apartado 2 de este trabajo.

La medida de los indicadores complejos se basa en escalas o instrumentos ya elaborados o que se elaboran ex profeso como escala original desde la que posteriormente se debe derivar el micro-instrumento. Debe tenerse en cuenta que una debilidad importante de la investigación psico-socio-educativa es que en muchas ocasiones, bajo una misma denominación de un constructo, se derivan soluciones teóricas y métricas muy diferenciadas. Por ejemplo, bajo el término *clima social del aula*, se encierran diferentes concepciones que, si bien parten de un enfoque de medida común (basado en la percepción de los sujetos), difieren en los componentes teóricos que integran (por ejemplo: ¿clima social y clima de aprendizaje forman parte del mismo constructo? ¿son constructos independientes?...) (Ramos y Pérez-Carbonell, 2008, 2009; Pérez-Carbonell, Ramos y López-González, 2009; López-González, Pérez-Carbonell y Ramos, 2011; Murillo y Hernández-Castilla, 2011). Es frecuente encontrar en la literatura cuestiones de este tipo. Por ello, si se pretende basar la medida en instrumentos ya elaborados (a partir de los cuales se derive el micro-instrumento) hay que ser muy cauteloso, extremando el análisis documental y el análisis lógico del constructo que se evalúa (llevado a cabo por un comité de expertos) y las soluciones métricas disponibles. Por otra parte, si se aborda el diseño de un instrumento capaz de medir determinado constructo es conveniente realizarlo con un enfoque diagnóstico, es decir, definiendo el constructo y el instrumento que lo represente en toda su extensión, sin limitar previamente la longitud del instrumento: hay que priorizar la representación sustantiva, teórica, o lo que es lo mismo la validez de constructo y contenido (ver como ejemplos, González-Barbera *et al.*, 2009; Chiva y Moral, 2009; Gómez-Costa y Cardona, 2009; Biencinto *et al.*, 2009). En

cualquier caso, un instrumento que tenga una adecuada validez de constructo (sea un instrumento ya existente o creado *ex profeso*) será sin duda el mejor punto de partida para la identificación de micro-instrumentos que puedan integrarse posteriormente como componentes métricos de un cuestionario de contexto.

Fase 4. Validar el modelo

Una vez diseñada la batería de cuestionarios de contexto incluyendo los diferentes constructos y las diferentes fuentes es necesario validar el modelo. En realidad, los cuestionarios de contexto se han construido a modo de un plan *integrado* y *coherente*, que incluye los indicadores simples y otros más difíciles de conceptualizar —los indicadores complejos, que miden un constructo con micro instrumentos que son en realidad escalas reducidas— e integrando *versiones diferenciales* de los mismos en función del colectivo al que se dirigen (la batería de cuestionarios de contexto incluye, en realidad, tantos cuestionarios como colectivos o fuentes vayan a ser consultados en la evaluación de sistema de que se trate: alumnado, profesorado, directores/as, familias…).

La validación de este sistema de recogida de información encuentra en las ecuaciones estructurales una metodología con el potencial de recoger la complejidad de niveles y tipos de información, para esclarecer el peso diferencial con que contribuye a la explicación del desempeño educativo cada uno de los indicadores pertenecientes al sistema de cuestionarios de contexto (González-Montesinos y Backhoff, 2010).

Descripción del procedimiento R-AVACO[4]para la elaboración de micro-instrumentos: optimización de la medida

Tradicionalmente, la identificación de los mejores predictores de una escala se ha apoyado en el uso de modelos de regresión, principalmente de la regresión paso a paso. La dificultad que entraña este tipo de aplicación para el propósito que nos ocupa es que la selección de los mejores predictores de la puntuación total de la escala se realiza basándose únicamente en criterios de tipo estadístico (como, por ejemplo, la eliminación de reactivos que por tener una elevada correlación con la puntuación total y a su vez con otros reactivos se entienden como información redundante). Este tipo de selección, en múltiples ocasiones, puede conllevar una distorsión en cuanto a la definición real del constructo que se pretende evaluar con la escala original y el que queda representado en el micro-instrumento. Por ello, el proceso que se presenta a continuación pretende ser una ayuda para la identificación de micro-instrumentos que mantengan su capacidad de representar el constructo original.

El objetivo principal de los estudios realizados en los proyectos mencionados (AVACO y M-AVACO) ha sido identificar una estrategia metodológica que permitiera derivar micro-instrumentos de medida a partir de instrumentos diseñados para medir/evaluar variables de entrada, proceso y contexto, utilizables en la elaboración de cuestionarios de contexto para la evaluación de sistemas educativos.

Como objetivos implicados tuvimos en cuenta:

- Definición de la estrategia metodológica de reducción de instrumentos a partir de la identificación de reactivos clave (los mejores predictores de las puntuaciones totales o dimensionales de instrumentos ya desarrollados).
- Validación de la estrategia tomando como referencia ensayos piloto con diferentes variables usuales en la evaluación de sistemas educativos.

Así, el procedimiento que aquí presentamos está dirigido a la identificación de los ítems o reactivos clave que permitan, para un constructo dado y que pueda ser medido por una escala determinada, aportar información suficiente

para orientar, al menos, una clasificación de sujetos respecto al constructo, con un elevado grado de fiabilidad y validez, similar a la que se produciría con la totalidad de la escala.

Se trata pues, de seleccionar los elementos de la escala objeto de estudio que mejor permitan mantener las características métricas de la misma. Se pretende, de este modo, obtener un micro-instrumento que pueda ser considerado como parte de un cuestionario de contexto.

Por ello, el procedimiento que hay que seguir es básicamente técnico —estadístico/ psicométrico— y se aplica sobre la información recabada con una escala ya existente, que permita medir o evaluar un constructo determinado. Por ejemplo, en este mismo número monográfico se presentan aplicaciones basadas en esta propuesta: López-González, Tourón y Tejedor (2012) y Joaristi, Lizasoain y Gamboa (2012).

Asumimos que el micro-instrumento obtenido no permitirá el diagnóstico individual —no es lo que se pretende—, aunque la escala origen de la que parte evidentemente sí lo permitía. Asumimos esta premisa porque la finalidad de la evaluación en la que se integran estos cuestionarios de contexto —la evaluación de sistemas educativos—, no es el diagnóstico individual, sino el análisis meso o macro, como se ha indicado anteriormente

La figura 4 recoge el procedimiento general de elaboración de micro-instrumentos.

Fase 0. Análisis de dimensionalidad de la escala

Para la medición de cada uno de los constructos identificados como indicadores complejos en el diseño de los cuestionarios de contexto partimos, como se ha indicado, de una escala existente y validada (cuando el análisis de contenido garantiza que se trata del mismo constructo) o de una escala diseñada y validada *ex profeso*. Pre-existente o diseñada para esa finalidad, esta va a ser la que denominamos en lo sucesivo escala original, o la versión 1, cuyo número de ítems es el número original de la escala (N. orig. en la figura 4).

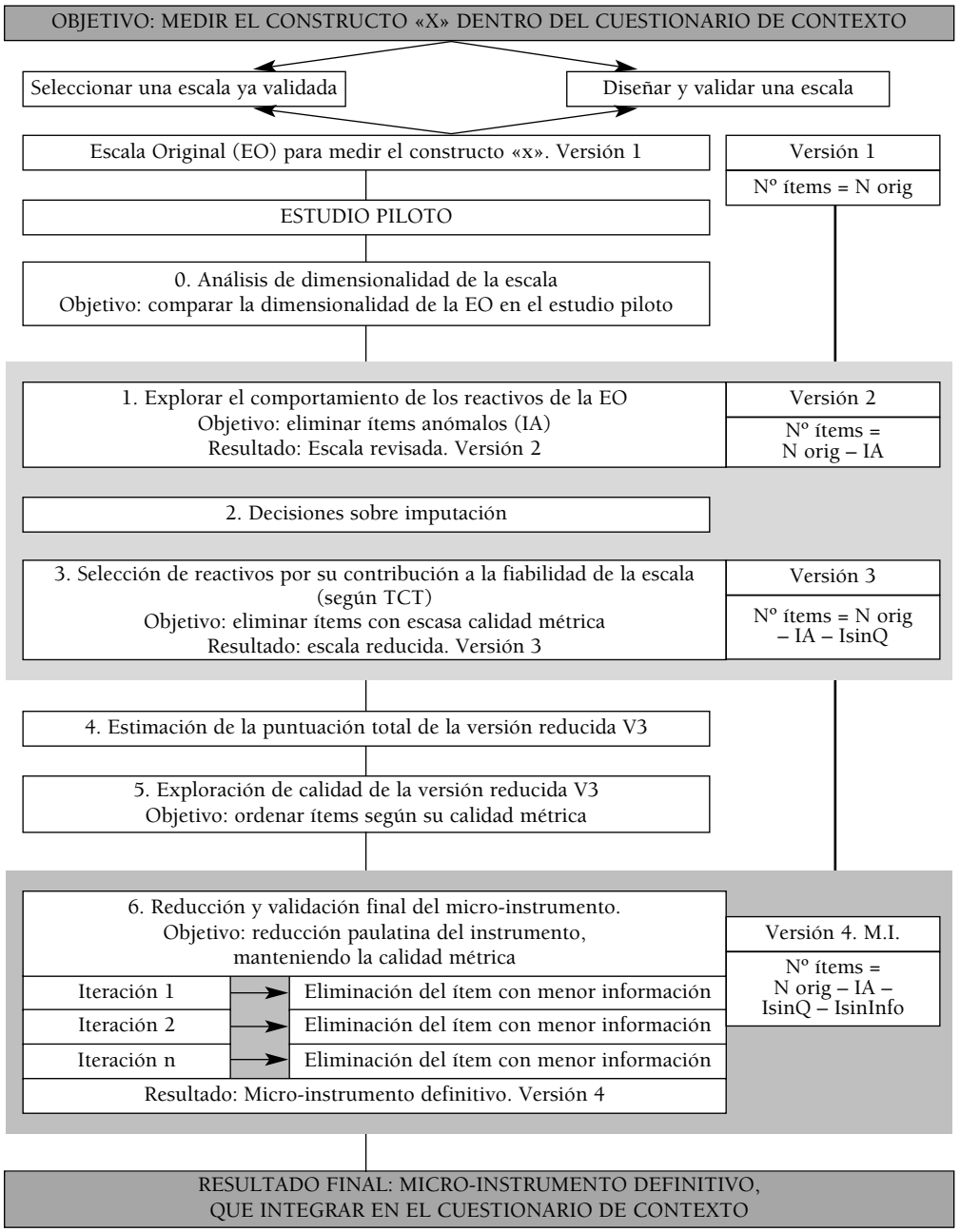
Para realizar el proceso de reducción de la escala original a un micro-instrumento partimos de un ensayo piloto: una aplicación de la escala original a un grupo similar al que será destinatario final de la prueba, con cuyos datos se va a trabajar.

La fase de análisis de dimensionalidad de la escala no forma parte del procedimiento de reducción propiamente dicho sino que se trata, como cuestión previa, de asegurarnos de que la escala original con la que vamos a trabajar funciona de manera similar (unidimensional o multidimensional) con el grupo con el que se ha llevado a cabo el ensayo piloto para proceder al estudio de reducción a como funcionaba en los estudios ya realizados por los autores que la diseñaron.

Como es sabido, el problema de análisis de la dimensionalidad es una cuestión recurrente, dada su complejidad. En el caso de escalas (sean de percepción, o de actitudes), habitualmente se comprueba mediante análisis factorial —AF— (con diferentes soluciones para la extracción de factores y para la rotación). Sin embargo, generalmente, la métrica de las variables no corresponde a la adecuada para este tipo de análisis. Las escalas Likert de instrumentos que miden constructos de percepción (como por ejemplo, clima social del aula, auto-concepto académico, etc.) podríamos clasificarlas como *ordinales-de intervalo*; es decir, al menos se puede asegurar la ordinalidad en la medida, si bien no podemos asegurar que las unidades sean iguales, con lo que su identificación como medidas de intervalo queda comprometida.

Por ello, aunque en la tradición investigadora se ha utilizado con profusión el AF para este propósito, en nuestro caso estimamos más oportuno proceder mediante análisis de conglomerados

FIGURA 4. Procedimiento R-AVACO para la elaboración de micro-instrumentos



jerárquicos, tomando como medida la *distancia euclídea*, que se ajusta mejor a la métrica de las variables a analizar. Como método de conglomeración, el tradicional método de Ward, asumiendo distancia 5 para la identificación de dimensiones. En cualquier caso, la

asunción de distancia 5 es arbitraria, si bien la idea es intentar una clasificación de ítems en conglomerados combinando al mismo tiempo el criterio lógico-sustantivo (o teórico) con el criterio estadístico (es decir, menor distancia re-escalada). Es probable que en cada constructo haya que *modular* esa distancia, pudiendo ser menor o algo mayor. No obstante, lo recomendable es que sea menor.

Esta exploración nos permitirá comprobar si las dimensiones que se identifican corresponden con las teóricas, descritas y comprobadas para la escala original. En el caso en que se dé más de una dimensión, el proceso se realizaría para cada una de ellas, con el fin de mantener la estructura sustantiva, teórica, de la escala tal como fue diseñada.

Existe la alternativa de emplear las técnicas derivadas de la modelización Rasch para establecer las propiedades de métricas de las escalas identificadas en las fases previas de la reducción (González-Montesinos, 2012). Siguiendo a este autor, se puede señalar que este tipo de procedimientos es apropiado, particularmente para escalas con reactivos de respuesta graduada (Likert). Disponemos de una serie de procedimientos desarrollados como extensiones del modelo Rasch original. Estos procedimientos alternativos son: el Graded Response Model (GRM) (Samejina, 1969), el Rating Scale Model (RSM) (Andrich, 1978) y el Partial Credit Model (PCM) (Masters, 1982). Estas técnicas tienen en común la ventaja de aportar elementos para establecer la dimensionalidad de escalas compuestas por reactivos politómicos (Likert) y, además, establecen las propiedades métricas de las categorías internas de los reactivos. Para ello se calibran las dificultades de los umbrales de paso entre cada categoría interna, de manera tal que se asegura que las respuestas a las partes bajas o altas de la estructura de los reactivos representan en efecto un menor o mayor grado de posesión del rasgo que se pretende medir con la escala (Embretson y Reise, 2000).

También está vigente la muy potente alternativa de someter a prueba los modelos de medida de las escalas identificadas en la reducción a través de procedimientos de análisis factorial confirmatorio (AFC), que forman parte de los modelos de ecuaciones estructurales (SEM) (González-Montesinos y Backhoff, 2010; Backhoff y González-Montesinos, 2012). De hecho, una combinación de procedimientos de escalamiento Rasch y confirmación de modelos de medida a través de CFA-SEM es la alternativa ideal para las fases intermedias y finales de consolidación de las escalas e instrumentos de contexto. Esta combinación de técnicas psicométricas para validación de escalas se ha empleado ya con mucho éxito en evaluaciones nacionales e internacionales de gran alcance, tanto para ítems de dominios cognitivos, actitudinales y de percepción (Schultz y Sibberns, 2004).

Fase 1. Explorar el comportamiento de los reactivos de la escala original

En la fase 1 se inicia un primer ciclo de reducción de la escala original, basado en la eliminación de los ítems con un comportamiento menos adecuado (partiendo de que todos ellos fueron considerados adecuados en el estudio de validación de la escala original, y por eso forman parte de ella). Dado que el uso final del micro-instrumento no será el diagnóstico individual, sino el análisis meso y macro, dentro de instrumentos de contexto integrados en evaluaciones de sistema, el objetivo es ir reduciendo el instrumento, disminuyendo paulatinamente el número de ítems, eliminando aquellos que tengan un comportamiento menos claro o aporten menor información.

La exploración debe constituir el primer paso de todo el proceso, de forma que la primera reducción de datos se realice a partir de la misma.

a) El *objetivo* de esta fase es eliminar aquellos reactivos que tienen comportamientos anómalos (ítems anómalos, IA), tales como, los que presentan:
   - Ausencia total de varianza.

- Escasa variabilidad. Tomamos, en este caso, como criterio de decisión que se eliminarían aquellos elementos que tengan más del 80% en un extremo de la escala (por ejemplo, si trabajamos con una escala Likert de 4 puntos, en valoraciones 1-2, o bien, 3-4), lo que sería concurrente con asimetrías muy marcadas.
- Existencia de un porcentaje excesivo de casos extremos (*outliers*). Como criterio entendemos que este no debería exceder del 5%.

b) El procedimiento se basa en una exploración de datos (fundamentalmente en procesos gráficos y la distribución de frecuencias de la escala[4]). Se trata, pues, de analizar las formas de la distribución, de manera que aquellos elementos que tengan una distribución atípica se eliminan.

Con la aplicación de estos criterios, se elimina un número *N* de ítems, aquellos considerados anómalos. Tenemos la versión 2 del instrumento, cuyo número de ítems será el *N* original de la escala menos los considerados ítems anómalos (Nº ítems = N. orig – IA).

El análisis de datos exploratorio (exploratory data análisis o EDA) fue originalmente propuesto por J. W. Tukey (1977) y proporciona todos los elementos necesarios para lograr la primera reducción de reactivos. Los procedimientos EDA corresponden a una aplicación completa de elementos clave de la estadística descriptiva (frecuencias, medias, desviaciones típicas y gráficos) y viene implementado en la mayoría de los paquetes estadísticos comúnmente disponibles.

Fase 2. Valores perdidos y decisiones sobre imputación

El procedimiento continúa con un *análisis de valores perdidos* para comprobar la hipótesis de aleatoriedad. En términos generales, se puede afirmar que a menor número de valores perdidos la escala es *factible*. Se trata de comprobar que la presencia de valores perdidos es aleatoria y no responde a un patrón sistemático de no respuesta. Por ello, aunque no existe un criterio prefijado al que podamos aludir como referencia para la toma de decisiones, estimamos que cualquier ítem con una presencia de valores perdidos superior al 30% puede contener algún problema de formulación importante, o bien, abordar un contenido no adecuado para el grupo al que se dirige el instrumento, por lo que sería conveniente su eliminación. En caso de que el número de casos perdidos sea mínimo y no se mantenga un patrón sistemático, se pueden asumir dichos casos como mortalidad experimental y eliminarlos del grupo sobre el que se realiza el estudio, o bien pensar en algún procedimiento robusto o de recorte como, por ejemplo, la media recortada (*trimmed mean*).

En caso de que exista un número considerable de valores perdidos (pero siempre inferior al 30%), se analizan otras opciones, como la imputación o la interpolación, para recuperar esos casos.

La imputación es un proceso habitual cuando se trata de trabajar con escalas de opinión, actitudes, etc. El motivo fundamental reside en que en este tipo de instrumentos suelen darse bastantes reactivos omitidos, de forma que los sujetos, cuando no tienen formada una respuesta, prefieren dejar en blanco, no contestando el elemento. El efecto sobre los análisis es muy negativo, pues en todas aquellas aproximaciones en que se requiera que los registros individuales estén completos se elimina el registro en su totalidad, aunque sea tan solo un ítem el que está en blanco. Obviamente, si se trata de análisis en los que se relacionan variables, la eliminación de casos se incrementa, pues se requiere que todos los casos tengan respuesta en las variables que se analizan (así, por ejemplo, si se trata de una correlación bivariada, debe haber respuesta en las dos variables; en caso contrario, todos los registros que no

tengan respuesta en ambas variables se eliminan). Todo esto puede implicar un grave proceso de reducción de la muestra.

Respecto a las soluciones de imputación, pasan sobre todo por imputar o interpolar. Existen diversos procedimientos que pueden aplicarse en función de la métrica de las variables, las características de la distribución de las mismas, el tipo de muestreo, la cantidad de casos perdidos y las relaciones que pueden establecerse entre las variables que se deben imputar con otras variables a partir de las cuales se pueda reducir la incertidumbre (Särndal, Swensson y Wretman, 1991; Muñoz y Álvarez, 2009).

En nuestro caso, hemos optado por un proceso simple de imputación: la sustitución del caso perdido por la mediana del ítem —en el caso de variables en las que al menos se pueda asegurar la ordinalidad, pero no más allá—, o por la moda —en el caso de variables nominales—, considerando todas las respuestas existentes. Así, el proceso de imputación que hemos seleccionado tiene las siguientes características:

I. El *objetivo* en este caso es mantener el tamaño de la muestra. Por lo que siempre que se haya cumplido el criterio señalado en la fase anterior, mantendríamos los sujetos imputando los valores perdidos.
II. En cuanto al *procedimiento*, en nuestro caso, procedemos por la sustitución de valores perdidos por la mediana (considerando todas las respuestas al ítem). Si se trata de variables nominales, la opción es la moda.

Adicionalmente al proceso de imputación propiamente dicho, estimamos que se requiere una *validación del proceso de imputación*, de forma que se estime si esta ha producido diferencias entre la distribución de los datos originales y la serie ya imputada. El problema metodológico aquí no es menor; de hecho buena parte de los procedimientos para analizar la relación entre ambas series no es aplicable, dado que eliminaría los sujetos que contuvieran valores perdidos en la variable original. Por ejemplo, una correlación entre ambas series, que sería lógicamente un indicador de referencia, no es aplicable. Siempre nos ofrecería como resultado +1 dado que, en definitiva, se acabarían correlacionando las dos series únicamente con los sujetos completos.

La prueba *T* de Wilcoxon, para muestras relacionadas puede constituir un apoyo simple de corroboración de la imputación, dirigido a contrastar si los rangos de ambas series de la variable al menos se mantienen sin distorsiones debidas la imputación[6].

Fase 3. Selección de reactivos a partir de su contribución a la fiabilidad de la escala

La tercera fase del procedimiento se dirige a la selección de reactivos que puedan formar parte del micro-instrumento, según su calidad métrica. Así, partimos desde la teoría clásica de construcción de tests (TCT) de un análisis de fiabilidad, basado en el modelo alfa de Cronbach (1951). Los objetivos de esta fase son:

- Seleccionar los elementos que mejor se relacionan con el puntaje total del cuestionario.
- Eliminar elementos redundantes (optimizar información).

Para proceder a la selección de ítems, se toman del análisis los siguientes criterios de forma secuencial:

- En primer lugar, eliminar los elementos que presenten en el indicador «*alfa si se elimina el elemento*» un incremento de la fiabilidad.
- A continuación, eliminar los elementos utilizando el índice de homogeneidad corregido: $r_{it\text{-}i} \leq 0.30$.
- Finalmente, eliminar los elementos que, presentando en la matriz de correlaciones

entre los ítems (preferiblemente utilizando la medida de correlación de Spearman, dado que al menos podemos asegurar la ordinalidad, pero no más allá) una correlación con otro elemento $r_{xy} \geq 0.50$, a su vez, tengan peor índice de homogeneidad corregido. No obstante, este criterio se aplicará revisando el contenido de los ítems, de forma que aquellos que claramente midan aspectos diferentes del constructo y sean característicos de diferentes dimensiones se mantendrán, aunque se cumpla el criterio numérico.

En este caso, se eliminan los ítems que no satisfagan alguno de los criterios mencionados: aquellos que tengan una menor calidad métrica (*IsinQ*, en la figura 4). Concluida esta fase, se dispone de la versión 3, que estará compuesta por los reactivos seleccionados como ítems-clave de la escala objeto de estudio. Su número de ítems, por tanto, será el *N* de la escala original, menos los ítems considerados anómalos en la fase 1 (IA) y los ítems considerados con menos calidad métrica en la fase 3 (IsinQ).

Fase 4. Estimación del puntaje total de la versión reducida

Se estima como la suma total de puntuaciones a los ítems. Si bien no tiene un uso directo para la selección inicial de ítems con la versión reducida, se utiliza como referente del nuevo instrumento. Es necesaria para el proceso de validación final del micro-instrumento.

Fase 5. Exploraciones de la calidad de la versión reducida

Para concluir el proceso de selección de elementos y, en consecuencia, ajustar la versión reducida final, se realizan diversas exploraciones a partir de las cuales conformamos la decisión final.

En primer lugar, se analiza la versión 3 del instrumento a partir del modelo alfa de Cronbach. Con ello, disponemos del indicador de fiabilidad global para poder así corroborar el tamaño de la pérdida en fiabilidad desde la escala original. El criterio en este caso es obvio: a mayor valor en el coeficiente alfa, mejor. No obstante, hay que tener en cuenta el punto de partida de alfa en la escala original y tomamos como criterio global que la pérdida en fiabilidad sea ≤10%, siempre y cuando se mantenga en valores altos.

Tras esta comprobación global de la escala, que ofrece además la puntuación alfa de referencia, iniciamos la exploración de los ítems de la versión 3. Estas exploraciones, que nos van a permitir una nueva selección de ítems, se realizan sobre la versión ya reducida del instrumento (versión 3), y se apoyan en los siguientes indicadores:

- Se identifican grupos extremos (27% superior e inferior) a partir de la puntuación total de la versión reducida (versión 3). Posteriormente, se contrasta la media de cada ítem entre ambos grupos, mediante la prueba *t* de Student. En el caso en que las pruebas *t* no fueran significativas (hecho que puede resultar extraño dado el proceso de reducción anterior), se utilizarían para seleccionar ítems, de forma que estos serían eliminados. Se trataría, pues, de una segunda depuración. Ello nos permite utilizar la discriminación de cada ítem como criterio de ordenación de los mismos, de forma que constituye el primer indicador para la confirmación de la versión reducida (versión 3).
- Asimismo, se analizan los ítems considerando la pérdida en el coeficiente de fiabilidad si se elimina el ítem. Como en casos anteriores, siempre y cuando se observe un incremento en el coeficiente alfa al eliminar el ítem, este debe ser eliminado. En cualquier caso, como criterio, nos permite ordenar los ítems en función de las aportaciones a la fiabilidad total de la versión reducida (versión 3).
- El tercer indicador para cada reactivo será nuevamente el índice de homogeneidad

corregido, si bien en esta ocasión se estima para el ítem en el micro-instrumento. Se entiende que a mayor intensidad de r_{it-i}, positivo, el ítem muestra un mejor comportamiento como representación del puntaje global de la versión reducida (versión 3).

- El cuarto indicador es el valor T_{ij} (suma de las covarianzas de cada ítem con todos los demás). Se puede entender como un indicador complementario al anterior. Así, el T_{ij} puede entenderse como la contribución que un ítem determinado realiza al conjunto de la variabilidad de la puntuación total del micro-instrumento considerando su relación con el conjunto de reactivos que lo componen. Recuérdese que:

$$\sigma_t^2 = \Sigma\sigma_i^2 + 2\sigma_{ij}$$

Donde:

σ_t^2: Varianza de la puntuación total en el micro-instrumento
$\Sigma\sigma_i^2$: Sumatorio de las varianzas de los ítems
σ_{ij}: Sumatorio de las covarianzas entre los ítems, es decir: $\sigma_{ij} = \Sigma\ T_{ij}$

La selección cuantitativa final de ítems se fundamentará, además de en el proceso de validación que describiremos a continuación, en la ordenación de los ítems en función de estos cuatro indicadores. De esta forma, los resultados se trasladan a una hoja de cálculo y en ella se establecen los rangos con cada indicador, así como el rango promedio. Estos resultados serán un elemento de referencia para el proceso iterativo de validación posterior. Con todo lo anterior, dispondremos de la información de calidad de la versión de instrumento reducido disponible para el proceso de validación.

Fase 6. Reducción final y validación final del micro instrumento

Se inicia aquí una segunda etapa de reducción. Partiendo de la versión reducida (versión 3) y tomando como criterio la ordenación de ítems por calidad métrica resultante de la fase 5, en esta etapa se reduce paulatinamente el instrumento en un ítem para proceder paralelamente a su validación en un proceso iterativo en el que se va comprobando la calidad global de la información que ofrece el instrumento reducido en tres niveles:

1. Uso de la puntuación total.
2. Calidad del instrumento para clasificar tipologías de centros, aulas o alumnos.
3. Mantenimiento de las características sustantivas (teóricas y de calidad métrica) del instrumento.

Así, tenemos:

- *Objetivos*: comprobar si la puntuación global de ambas versiones (la versión 3, y con un ítem menos) es equivalente, y si la clasificación de sujetos, aulas y/o, centros que producen ambas versiones del cuestionario son equivalentes.
- *Procedimiento de validación*: en este caso, se entiende que el proceso de validación se dirige a corroborar si el micro-instrumento permite clasificaciones similares a las que se pudieran establecer con la escala original. Teniendo como referencia la información de la fase 5 acerca de la versión reducida (versión 3), se comienza el proceso de iteraciones, con el número de ítems seleccionado. Para cada iteración, se tiene en cuenta:

  - Correlación de Pearson entre los totales de ambas versiones.
  - Clasificación de los sujetos a partir del puntaje total en tres niveles (alto, medio y bajo) tomando como referencia los grupos extremos (27% superior e inferior).
  - Comparación de ambas clasificaciones mediante Ji-cuadrado entre escala original *vs.* micro-instrumento.
  - Comparación de la clasificación de centros (aulas y/o sujetos) que producen ambos instrumentos (escala original *vs.*

micro-instrumento). Así, se observa mediante Ji-cuadrado si las clasificaciones de las diferentes unidades muestrales (centros o aulas) son equivalentes. Como criterio para representar al instrumento reducido tomaremos el porcentaje de unidades en que se produzca una clasificación concurrente entre ambas versiones del instrumento.

- Criterio teórico sustantivo. Asimismo, se tendrá en cuenta si están representadas en el instrumento reducido todas las dimensiones identificadas mediante el análisis de dimensionalidad inicial al menos por un ítem. O, cuanto menos, si estos pertenecen mayoritariamente a la dimensión más general y no se produce pérdida de información teórica sustancial que distorsione la cualidad del constructo que se evalúa.

Las siguientes iteraciones se establecerán disminuyendo un ítem en cada una de ellas, utilizando como referencia la ordenación de ítems estimada en la fase 5.

El criterio de detención de las iteraciones se vincula con la calidad métrica de la reducción. Se detendrán las iteraciones cuando se produzca alguna de las siguientes situaciones en los criterios mencionados:

a) La disminución de la correlación entre las versiones por debajo de 0.90.
b) Cuando no se produzca una coherencia entre las clasificaciones de ambas versiones entre las clasificaciones producidas entre los totales de ambas versiones (criterio 2 y 3).
c) Cuando se den coincidencias entre las clasificaciones estimadas por unidades muestrales, inferiores al 70% (criterio 5).
d) Cuando no se cumpla el criterio 5, habiendo eliminado ítems redundantes.

De esta forma, se replica el procedimiento con todas las variaciones posibles de elementos. La solución más válida será la que cumpla los siguientes criterios:

- *Eficacia:* maximizar la correlación de Pearson entre puntajes totales de la escala original y del micro-instrumento y maximizar la coherencia entre las clasificaciones derivadas de la escala original y la del micro-instrumento.
- *Eficiencia*: minimizar el coste en el número de ítems necesario para informar con un nivel suficiente de fiabilidad y validez acerca del constructo que se mide con la escala original.
- *Funcionalidad*: representar las cualidades sustantivas del constructo evaluable, tal cual fue definido teóricamente para el diseño de la escala original, es decir, manteniendo su validez de constructo.

Finalizado el proceso, obtenemos el micro-instrumento (MI) definitivo, la versión 4. Su número de ítems será el *N* de la escala original, menos los ítems anómalos (IA), menos los ítems con menos calidad métrica (IsinQ), menos los ítems que paulatinamente se han eliminado en la última reducción, por ser así clasificados según los criterios de la fase 5 como ítems que ofrecen menor información.

A modo de conclusión

La uniformidad y homogeneidad que se observa en los planes de evaluación de sistemas probablemente está en la base de la percepción generalizada de falta de utilidad.

Hay una dificultad constatada en la comunicación de resultados de las evaluaciones de sistemas: de la ingente cantidad de información que ofrecen, los medios de comunicación suelen destacar exclusivamente datos controvertidos como los *rankings* derivados de las pruebas de rendimiento, que se convierte en la única información que finalmente llega a la sociedad.

Incluso para los lectores más aventajados, las evaluaciones de sistemas adolecen de falta de utilidad. Y una de las razones más claras es la falta de calidad de los llamados cuestionarios de contexto. Es necesario que estos permitan extraer información explicativa de los resultados y, por tanto, indicaciones y argumentaciones para las políticas educativas en las distintas regiones. Buscar alternativas metodológicas para el diseño de cuestionarios de contexto de calidad ha sido el objetivo prioritario de los Proyectos AVACO y MAVACO, financiados por el Ministerio de Educación de España y sintetizar ese procedimiento ha sido el objetivo de este artículo. Confiamos haber iniciado un camino, desde el punto de vista de la medición educativa. El trabajo de diferentes equipos de investigación, en distintos territorios, para confirmar, rebatir y mejorar la utilidad del procedimiento que hemos presentado será imprescindible para contribuir desde el conjunto de la investigación educativa a que la evaluación de sistemas responda a las finalidades por las que se planteó.

Notas

[1] En ambos casos, ha actuado como investigador principal J. M. Jornet y la Universitat de València ha sido la coordinadora de la red de universidades (Universidad Complutense, Universidad del País Vasco, Universidad de Cádiz, Universidad de Málaga, Universidad de Navarra, Universidad de Castilla La Mancha, Universitat Jaume I y Universidad Autónoma de Barcelona), que han dado respuesta a ambos proyectos. La línea de investigación iniciada por estos proyectos además tiene continuidad en el Proyecto EVALEF, "Validación de un instrumento de evaluación de estilos educativos familiares y establecimiento de lineamientos para el diseño de programas de intervención con familias" (financiado en el Plan Nacional de I+D+i, Referencia EDU2011-29467, dirigido por M. J. Perales). Tomando como constructo central los estilos educativos parentales, el proyecto parte de la metodología desarrollada en Avaco y Mavaco para desarrollar la escala de estilos educativos familiares. Asimismo, los proyectos *Evaluación del clima social del aula en educación secundaria* (financiado por la Universitat de València) y *Diseño de instrumentos de valoración del clima de aprendizaje en estudiantes universitarios* (financiado por la Generalitat Valenciana), centrados en el clima social de aula, surgen de los proyectos Avaco y Mavaco y aplican su metodología de diseño de instrumentos de contexto.

[2] Probablemente con enfoques metodológicos alternativos pueda llegar a establecerse esa continuidad desde lo micro-analítico hasta lo macro-analítico, pero se requerirá otro tipo de diseños de los planes de evaluación de sistemas educativos, que también será necesario explorar.

[3] En el artículo de De la Orden y Jornet en este mismo número de Bordón se analiza en profundidad el concepto de calidad y su concreción en efectos que se van a considerar en la evaluación de sistemas educativos.

[4] Si bien este trabajo se presenta por los autores del artículo, la definición del procedimiento R-AVACO de optimización de la medida ha sido desarrollado contando además con las aportaciones de diversos investigadores: Emelina López González (UV), Javier Tourón (UNAV), Luis Lizasoain (EHU), Luis Joarísti (EHU) y Javier Tejedor (USAL), por lo que si se utilizara el procedimiento y se deseara citar, debería referirse al conjunto de investigadores (autores del artículo y participantes).

[5] Mediante SPSS, se pueden usar comandos tales como explorar, pedir gráficos de cajas, tallos y hojas y frecuencias, o bien, utilizar procedimientos más potentes, como los disponibles en *R*.

[6] Habitualmente trabajaremos con muestras grandes ($N \geqslant 25$), por lo que la *T* debe transformarse a *Z* para comprobar la probabilidad.

Referencias bibliográficas

ANDRICH, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

BACKHOFF, E. y GONZÁLEZ-MONTESINOS, M. (2012). Evidencias de validez del cuestionario para docentes del Estudio Internacional sobre Enseñanza y Aprendizaje (TALIS-2009). *Bordón*, 64 (2), 173-194.

BACKHOFF, E., BOUZAS, A., GONZÁLEZ-MONTESINOS, M., ANDRADE, E., HERNÁNDEZ, E. y CONTRERAS, C. (2008). *Factores asociados al aprendizaje de estudiantes de 3º de primaria en México.* México D. F.: Instituto Nacional para la Evaluación de la Educación (INEE).

CARDONA, L., PERALES, M. J. y GÓMEZ-COSTA, D. (2009). *Conferencia: «Familia y transformación social. Análisis del papel de las familias en los estudios de evaluación de sistemas educativos. Introducción al estudio de validación de un cuestionario»*, Huelva, XIV Congreso de AIDIPE: Educación, investigación y desarrollo social.

CHIVA, I. y MORAL, A. (2009). *Conferencia: «Diseño y revisión lógica de una escala para evaluar la metodología docente en primaria y secundaria»*, Huelva, XIV Congreso de AIDIPE: Educación, investigación y desarrollo social.

CRONBACH, L. J. (1951). Coefficient alpha and the internal structure of test. *Psychometrika*, 16, 297-334.

DE LA ORDEN, A. (2007). Evaluación de la calidad de la educación. Un modelo sistémico como base para la construcción de un sistema de indicadores. En Seminario Internacional de Indicadores Educativos (memoria): *Conceptos, metodologías y experiencias para la construcción de indicadores educativos.* México: Instituto Nacional para la Evaluación de la Educación (INEE), 6-21.

DE LA ORDEN, A. y JORNET, J. M. (2012). La utilidad de las evaluaciones de sistemas educativos: el valor de la consideración del contexto. *Bordón*, 64 (2), 69-88.

EMBRETSON, S. E. y REISE, S. P. (2000). *Item Response Theory for Psychologists*. London, Mohaw, N. J.: Lawrence Erlbaum Associate Publishers.

GÓMEZ-COSTA, D. y CARDONA, L. (2009). *Conferencia: «Diseño y validación de un instrumento para la evaluación del autoconcepto académico: Ensayo piloto con alumnas y alumnos de sexto de primaria de la provincia de Valencia dentro del marco de la evaluación de sistemas educativos»*, Huelva, XIV Congreso de AIDIPE: Educación, investigación y desarrollo social.

GONZÁLEZ-BARBERA, C., GARCÍA-GARCÍA, M., GARCÍA-CORONA, D. y BIENCINTO, CH. (2009). *Conferencia: «EVADIE. Cuestionario para la evaluación de la atención a la diversidad. Diseño y validación»,* Huelva, XIV Congreso de AIDIPE: Educación, investigación y desarrollo social.

GONZÁLEZ-MONTESINOS, M. (2012). *El modelo métrico de Rasch: Fundamentación, implementación, interpretación.* Madrid: La Muralla (en prensa).

JOARISTI, L., LIZASOAIN, L. y GAMBOA, E. (2012). Construcción y validación de un instrumento de medida del nivel socioeconómico y cultural (NSE) de estudiantes de educación primaria y secundaria. *Bordón*, 64 (2), 151-172.

LÓPEZ-GONZÁLEZ, E., PÉREZ-CARBONELL, A. y RAMOS-SANTANA, G. (2011). Modelos complementarios al análisis factorial en la construcción de escalas ordinales: un ejemplo aplicado a la medida del clima social aula, *Revista de Educación,* 354, 369-397.

LÓPEZ-GONZÁLEZ, E., TOURÓN. J. y TEJEDOR, F. J. (2012). Diseño de un micro-instrumento para medir el clima de aprendizaje en cuestionarios de contexto. *Bordón*, 64 (2), 111-126.

MASTERS, G. N. (1982). A Rasch model for Partial credit scoring. *Psychometrika,* 60, 523-547.

MUÑOZ, J. F. y ÁLVAREZ, E. (2009). Métodos de imputación para el tratamiento de datos faltantes. *Revista de métodos cuantitativos para la economía y la empresa,* 7, 3-30.

PÉREZ-CARBONELL, A., RAMOS-SANTANA, G. y LÓPEZ-GONZÁLEZ, E. (2009). Diseño y análisis de una escala para la valoración de la variable clima social aula en alumnos de educación primaria y secundaria, *Revista de Educación,* 350, 221-252.

RAMOS SANTANA, G. y PÉREZ CARBONELL, A. (2008). *Conferencia: «El clima social aula: un reto para la formación integral de alumnos»,* Zaragoza, XIV Congreso Nacional y III Iberoamericano de pedagogía, educación, ciudadanía y convivencia, diversidad y sentido social de la educación.

RAMOS, G. y PÉREZ CARBONELL, A. (2009). *Conferencia: «Utilidad del diseño de una escala de valoración de la percepción clima social aula en los niveles de primaria y secundaria»*, Huelva, XIV Congreso de AIDIPE: Educación, investigación y desarrollo social.

SÄRNDAL, C. E., SWENSSON, B. y WRETMAN, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SCHULTZ, W. y SIBBERNS, H. (2004). *IEA Civic Education Technical Report*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

WILLMS, J. D. (2006). *Learning Divides: Ten Policy Questions About the Performance and Equity of Schools and Schooling Systems*. Montreal: UNESCO.

Fuentes electrónicas

BIENCINTO-LÓPEZ, C., GONZÁLEZ-BARBERA, C., GARCÍA-GARCÍA, M., SÁNCHEZ-DELGADO, P. y MADRID-VIVAR, D. (2009). Diseño y propiedades psicométricas del AVACO-EVADIE. Cuestionario para la evaluación de la atención a la diversidad como dimensión educativa en las instituciones escolares. *Relieve*, 15, 1. <http://www.uv.es/RELIEVE/v15n1/RELIEVEv15n1_4.htm.> [Fecha de consulta: 15/diciembre/2011]

DE LA ORDEN, A. (dir.) (1997). Desarrollo y validación de un modelo de calidad universitaria como base para su evaluación. *Relieve*, 3, 1 y 2. <http://www.uv.es/RELIEVE/.> [Fecha de consulta: 13/diciembre/2011].

GONZÁLEZ-MONTESINOS, M. J. y BACKHOFF, E. (2010). Validación de un cuestionario de contexto para evaluar sistemas educativos con Modelos de ecuaciones estructurales. *Relieve*, 16, 2. <http://www.uv.es/RELIEVE/v16n2/RELIEVEv16n2_1.htm> [Fecha de consulta: 13/diciembre/2011].

LIZASOAIN, L. y JOARISTI, L. (2010). Estudio diferencial del rendimiento académico en lengua española de estudiantes de educación secundaria de Baja California (México). *Revista Iberoamericana de Evaluación Educativa*, 3 (3), 115-134. <http://www.rinace.net/riee/numeros/vol3-num3/art6.pdf.> [Fecha de consulta: 13/diciembre/2012].

LUKAS, J. F. y SANTIAGO, K. M. (2004). Evaluación de centros escolares de educación secundaria del País Vasco. *Revista Electrónica de Investigación Educativa*, 6 (2). <http://redie.uabc.mx/vol6no2/contenido-lukas.html.> [Fecha de consulta: 13/diciembre/2012].

MIRANDA, L. (2008). Factores asociados al rendimiento escolar y sus implicancias para la política educativa del Perú. En BENAVIDES, M. (ed.), *Análisis de programas, procesos y resultados educativos en el Perú. Contribuciones empíricas para el debate*. Lima: Grade. <http://www2.minedu.gob.pe/umc/admin/images/publicaciones/artiumc/3.pdf.> [Fecha de consulta: 13/diciembre/2012].

MURILLO, F. J. y HERNÁNDEZ-CASTILLA, R. (2011). Factores escolares asociados al desarrollo socio-afectivo en Iberoamérica. *Relieve*, 17, 2, art. 2. <http://www.uv.es/RELIEVE/v17n2/RELIEVEv17n2_2.htm>[Fecha de consulta: 5/diciembre/2012].

MURILLO, J. (2009). Hacia un modelo de eficacia escolar. Estudio multinivel sobre los factores de eficacia en las escuelas españolas. *Revista electrónica Iberoamericana sobre calidad, eficacia y cambio en educación*. 6 (1), 4-28. <http://www.rinace.net/arts/vol6num1/vol6num1.pdf.> [Fecha de consulta: 13/diciembre/2012].

SAMEJINA, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 17. Richmond, VA: Psychometric Society. <http://www.psychometrika.org/journal/online/MN17.pdf> [Fecha de consulta: 18/noviembre/2011].

Abstract

Context questionnaire design for the evaluation of educational systems: optimization of complex constructs measurement

This paper is a review of a possible model for design and development of context questionnaires for the educational systems evaluation. General stages are reviewed and identify the composition of possible indicators (simple and complex) of these instruments. In the design of questionnaires of context one must consider various aspects, some of which the most important are: a) the selection of variables and indicators to be considered as part of the model, and b) the way in which complex indicators, i.e., those that are the result of a scale or a combination of simple indicators (represented by a single item) are evaluated. To evaluate complex indicators (for example, the social climate of the classroom) there are good construct validity scales; however, they cannot be used completely, since they are made up of a large number of items. In the latter case, we include a procedure that we have developed to design micro-instruments that are composed of the best items of the scale (i.e., that best predict the total score, keeping the dimensional structure of the scale). These micro-instruments must be able to represent complex indicators so that they can be included in the questionnaires of context. It is a procedure of reduction of scales, aimed at identifying the best elements or items that best predict the overall score of a scale, while maintaining its theoretical structure.

Key words: *Educational systems evaluation, Context questionnaires, Instrument design and construction methodology, Educational indicators.*

Résumé

La conception de questionnaires de contexte pour l'évaluation des systèmes éducatifs: l'optimisation de la mesure des construits complexes

Cet article fait une révision d'un possible modèle pour concevoir et construire des Questionnaires de Contexte pour l'évaluation des systèmes éducatifs. Nous revissons les phases générales, toutefois que nous identifions la possible composition des indicateurs (simples et complexes) de ces instruments. Dans la conception des questionnaires de contexte il faut tenir compte de plusieurs aspects. Parmi eux nous identifions comme des éléments clés: a) la sélection des variables et des indicateurs qui doivent être considérés et qui font partie du modèle, et b) la façon d'évaluer les indicateurs complexes, c'est à dire, ceux résultants d'une échelle ou bien d'une combinaison d'indicateurs simples (représentés par un seul élément). Pour l'évaluation d'indicateurs complexes (para exemple, le climat sociale de la classe) ils existent des échelles avec une bonne validité de construit; néanmoins elles ne peuvent pas être complètement utilisées puisqu'elles se composent d'un grand nombre d'items. Dans le dernier cas, nous développons une procédure pour la conception de micro-instruments composés par les meilleurs items de l'échelle (c'est à dire, les items qui prédisent le mieux le score total, en gardant la structure dimensionnelle de l'échelle). Ces micro-instruments doivent être capables de représenter des indicateurs complexes pour qu'ils puissent être inclus dans les questionnaires de contexte. Il s'agit d'un processus de

réduction des échelles, visant à l'identification des éléments ou items qui prédisent le mieux le score total d'une échelle, en gardant sa structure théorique.

Mots clés: *Évaluation des systèmes éducatifs, Questionnaires de contexte, Méthode de conception et constructions d'instruments, Indicateurs éducatifs.*

Perfil profesional de los autores

Jesús M. Jornet Meliá

Catedrático en el Departamento MIDE-UVEG. Coordinador del grupo GEM (MIDE-UVEG; www.uv.es/gem). Su trabajo se orienta en el área de medición y evaluación educativas al diseño de instrumentos para la evaluación de competencias, cuestionarios de contexto para la evaluación de sistema educativos y la evaluación de la dimensión educativa de la cohesión social.
Correo electrónico de contacto: jornet@uv.es

José González Such

Profesor titular en el Departamento MIDE-UVEG. Coordinador de la Unidad InnovaMide del grupo GEM (MIDE-UVEG). Su trabajo se centra en la medición y evaluación educativas, siendo sus líneas preferentes de investigación: diseño de pruebas e instrumentos de medición educativa, evaluación de la docencia y diseño y evaluación de materiales de innovación docente sustentados sobre nuevas tecnologías.
Correo electrónico de contacto: gonzalej@uv.es

Mª Jesús Perales Montolío

Profesora titular en el Departamento MIDE-UVEG. Coordinadora de la Unidad de Evaluación Socioeducativa del grupo GEM (MIDE-UVEG). Sus líneas centrales de investigación se dirigen a la evaluación de programas socioeducativos, evaluación institucional, evaluación de la formación ocupacional y continua, y el diseño y desarrollo de instrumentos de evaluación de competencias.
Correo electrónico de contacto: perales@uv.es.