Revisión de los Puntos de Corte en el Método General de Validación Empírica de la Matriz-Q

Pablo Nájera Álvarez

Máster en Metodología de las Ciencias del Comportamiento y de la Salud

MÁSTERES DE LA UAM 2017 - 2018

Facultad de Psicología

Universidad Autónoma de Madrid

Revisión de los Puntos de Corte en el Método General de Validación Empírica de la Matriz-Q

Estudiante
Pablo Nájera Álvarez

Tutor
Francisco José Abad García

Revisión de los Puntos de Corte en el Método General de Validación Empírica de la Matriz-Q

Pablo Nájera Álvarez

Máster en Metodología de las Ciencias del Comportamiento y de la Salud

Facultad de Psicología

Universidad Autónoma de Madrid

Junio de 2018

Índice

Resumen

Los modelos de diagnóstico cognitivo (CDMs) son modelos estadísticos multidimensionales de clases latentes que permiten clasificar con precisión a las personas mediante variables latentes discretas (llamadas atributos). Estos modelos requieren de una matriz-Q, que determina los atributos relevantes para cada ítem. El proceso de creación de la matriz-Q es subjetivo, y la presencia de errores de especificación en ella puede dar lugar a clasificaciones incorrectas. Por ello, en los últimos años se han desarrollado varios métodos empíricos de validación de la matriz-Q. El método propuesto por de la Torre y Chiu (Psychometrika, 2016), basado en un índice de discriminación, es uno de los más extendidos. Sin embargo, el estudio presenta como limitaciones el reducido número de condiciones examinadas y el uso de un punto de corte arbitrario (*EPS*). El presente trabajo plantea dos estudios de simulación para examinar este método de validación bajo un número mayor de condiciones con el objetivo de dotarlo de mayor generalización y de determinar empíricamente el *EPS* más adecuado. Los resultados indican un adecuado funcionamiento global del método, la presencia de interacciones entre los diferentes factores estudiados y la inviabilidad de emplear un único *EPS* general. Se proponen recomendaciones.

Palabras clave: CDM, G-DINA, matriz-Q, validación, EPS.

Abstract

*Cognitive diagnosis models (CDMs) are latent class multidimensional statistical models that help to classify people accurately by using discrete latent variables (attributes). These models require a Q-matrix, which determines the relevant attributes for each item. The Q-matrix construction process is subjective, and the existence of misspecification in the Q-matrix can lead to inaccurate classifications. Thus, in recent years several empirical Q-matrix validation methods have been developed. De la Torre and Chiu (Psychometrika, 2016) proposed one of the most spread methods, based on a discrimination index. However, the study has two limitations: the restricted number of conditions examined, and the use of an arbitrary cutoff point (*EPS*). The present work considers two simulation studies to test this validation method under a wider range of conditions, with the purpose of providing it a higher generalization, and of empirically determining the most suitable* EPS. *Results show a good global performance of the method, the presence of interactions between the different factors, and that using a single general* EPS *is unworkable. Recommendations are proposed.*

Key words: *CDM, G-DINA, Q-matrix, validation, EPS.*

Introducción

En las últimas décadas, alentados por la creciente influencia de la psicología cognitiva, se han desarrollado modelos estadísticos cuyo objetivo principal consiste en poder realizar inferencias sobre los diversos procesos cognitivos, así como las interacciones entre ellos, que subyacen a la hora de generar respuestas ante ítems. Los modelos de diagnóstico cognitivo (CDMs, por sus siglas en inglés *Cognitive Diagnosis Models*) forman parte de este grupo y han adquirido un notable interés en los últimos años, sobre todo dentro del ámbito educativo. Los CDMs nacen con el objetivo de ayudar a determinar con precisión las fortalezas y debilidades de aprendizaje de los estudiantes, de modo que los educadores tengan una información más específica que les permita dirigir sus esfuerzos de enseñanza hacia las áreas de conocimiento más problemáticas para sus alumnos. Esta perspectiva choca con aquélla que subyace a los tradicionales métodos de evaluación, a menudo reducidos a un mero proceso de calificación, en los que una única puntuación suele resumir el conocimiento de los alumnos en dominios relativamente amplios, impidiendo determinar exactamente dónde fallan. Esto dificulta la puesta en marcha de una estrategia educativa enfocada a reforzar estos conceptos concretos que no son comprendidos (de la Torre y Minchen, 2014).

Si bien la mayoría de artículos que tratan el tema de los CDMs se centra en el ámbito de la educación (p.ej., Chen, 2017; Chen y de la Torre, 2013; Chen, de la Torre y Zhang, 2013; Chiu, 2013; de la Torre, 2008, 2011; de la Torre y Chiu, 2016; de la Torre y Douglas, 2004; Ma y de la Torre, 2016; Romero, Ordóñez, Ponsoda y Revuelta, 2014), estos modelos son lo suficientemente generales como para ser usados en otras áreas. De esta forma, en los últimos años se han empleado en el estudio de patologías mentales (p.ej., de la Torre, van der Ark y Rossi, 2015; Jaeger, Tatsuoka, Berns y Varadi, 2006; Templin y Henson, 2006) o en el campo organizacional de la selección de personal (p.ej., García, Olea y de la Torre, 2014; Sorrel et al., 2016).

Los CDMs se conceptualizan como modelos multidimensionales de rasgos latentes, los cuales, en vez de estar definidos de modo continuo –como ocurre en la teoría de respuesta al ítem (TRI) o en el análisis factorial confirmatorio (AFC)–, se definen como categóricos/discretos. Estas variables latentes categóricas reciben el nombre de *atributos*, y hacen referencia a las habilidades o procesos cognitivos que un examinado debe poseer o dominar para resolver un ítem. Éstos pueden ser tanto dicotómicos ("maestría" vs. "no maestría" del atributo) como politómicos ("mal desempeño",

"desempeño regular", "buen desempeño"). Los atributos son variables latentes definidas con un cierto nivel de concreción que se desprenden de un dominio de conocimiento más amplio y, por tanto, están interrelacionados a la vez que son separables entre sí. La concreción o grado de especificidad de los atributos es flexible y puede variar en función de los propósitos del examinador, aunque estará sujeto al ajuste que posteriormente tenga el modelo (para un análisis más detallado sobre la definición y validación de los atributos, ver Li y Suen, 2013). El objetivo principal de los CDMs es el de proporcionar información detallada sobre los atributos que posee cada uno de los examinados (de la Torre y Sorrel, 2017) o, más técnicamente, de clasificar a los examinados en clases latentes especificadas por vectores de atributos (de la Torre y Douglas, 2004). Debido a la naturaleza discreta de los atributos, existirá un número restringido de clases latentes.

El número de atributos se suele representar con K, de tal modo que el examinado i tendrá un *perfil de atributos* $\boldsymbol{\alpha}_i = \{\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK}\}$, el cual es un vector binario donde $\alpha_{ik} = 1$ o 0 en función de si el examinado i domina o no el atributo k, respectivamente. Esto es cierto en caso de que los atributos sean dicotómicos, que es lo más habitual en la literatura; a partir de ahora nos centraremos en este tipo de casos (para un análisis del tratamiento de atributos politómicos, ver Chen y de la Torre, 2013). Así, existen un total de 2^K vectores de atributos diferentes que constituyen el espectro de clases latentes. Las clases latentes se representan como $\boldsymbol{\alpha}_l$, siendo $1 \leq l \leq 2^K$.

Los CDMs quedan expresados por su definición de $P(X_j = 1|\boldsymbol{\alpha}_l)$, la probabilidad condicional de acertar el ítem *j* dada la clase latente l (Sorrel et al., 2016). En función de cómo se especifique la interacción de los atributos para producir la respuesta a los ítems estaremos ante uno u otro CDM.

Revisión de los distintos modelos

En los últimos años se han desarrollado varios CDMs con diferente grado de generalidad. En este contexto, los modelos más específicos se suelen denominar reducidos, porque se consideran anidados dentro de los más generales. Dentro de los modelos reducidos más conocidos se encuentran el modelo *deterministic input, noise* and *gate* (DINA; Haertel, 1984; Junker y Sijtsima, 2001), que asume un proceso *conjuntivo* en el que el examinado debe poseer todos los atributos requeridos por el ítem para contestar correctamente, y el modelo *deterministic, input, noisy* or *gate* (DINO; Templin y Henson, 2006), que asume un modelo *disyuntivo* en el que es suficiente poseer uno de los atributos requeridos por el ítem para contestarlo correctamente. El modelo *noisy input,*

deterministic output and *gate* (NIDA; Maris, 1999; Junker y Sijtsima, 2001), el modelo *compensatory and reduced reparameterized unified model* (C-RUM and R-RUM; Hartz y Roussos, 2008), el CDM *aditivo* (A-CDM; de la Torre, 2011) y el modelo lineal logístico (LLM; de la Torre y Douglas, 2004) también pertenecen al grupo de modelos reducidos. Aunque estos modelos son fácilmente interpretables, es complicado establecer comparaciones entre sus diferentes estructuras. Por esta razón se han desarrollado CDMs más generales y flexibles, como el modelo de diagnóstico general (GDM; von Davier, 2005), el CDM loglinear (LCDM; Henson, Templin y Willse, 2009), y el modelo DINA generalizado (G-DINA; de la Torre, 2011). Estos modelos tienen la ventaja de que permiten relacionar muchos de los modelos reducidos anteriormente nombrados, así como estimar los parámetros y comparar el ajuste de diferentes CDMs ítem a ítem (de la Torre, 2011; de la Torre, van der Ark y Rossi, 2015; Sorrel, de la Torre, Abad y Olea, 2017). De todos ellos, el modelo G-DINA es el que más impacto ha tenido y el que ha sido más explorado, por lo que será objeto del presente artículo.

La formulación original del modelo G-DINA puede descomponerse en la suma de los efectos debido a la presencia de atributos específicos y sus interacciones (de la Torre, 2011):

$$P(\boldsymbol{\alpha}_{lj}^{*}) = \delta_{j0} + \sum_{k=1}^{K_j^{*}} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^{*}} \sum_{k=1}^{K_j^{*}-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{12\ldots K_j^{*}} \prod_{k=1}^{K_j^{*}} \alpha_{lk} \, ,$$

donde $\boldsymbol{\alpha}_{lj}^{*}$ representa el vector reducido de atributos cuyos elementos son relevantes para el ítem j; δ_{j0} es la probabilidad basal del ítem j; δ_{jk} es el efecto principal debido a α_k; $\delta_{jkk'}$ es el efecto de interacción debido a α_k y $\alpha_{k'}$; y $\delta_{12\ldots K_j^{*}}$ es el efecto de interacción debido a $\alpha_1, \ldots, \alpha_{K_j^{*}}$. Estos efectos (las δ) suelen ser los parámetros del G-DINA, con los que se estiman las probabilidades de acierto de las diferentes clases latentes. Sin embargo, al ser un modelo saturado, existe una correspondencia entre las δ y las clases latentes, por lo que las probabilidades de acierto de éstas también pueden ser los parámetros del modelo. Concretamente, dado $\widehat{\mathbf{P}}_j = \{\hat{P}(\alpha_{lj}^{*})\}$, la estimación de $\boldsymbol{\delta}_j$ puede obtenerse mediante:

$$\boldsymbol{\delta}_j = (\mathbf{M}_j'\mathbf{M}_j)^{-1}\mathbf{M}_j'\widehat{\mathbf{P}}_j \, ,$$

donde $\mathbf{M}_j$ es una matriz con dimensiones $2^{K_j^{*}} \times p$, siendo p el número de parámetros del modelo a estimar. Para el modelo saturado, $p = 2^{K_j^{*}}$.

La matriz-Q

Los CDMs, independientemente de su formulación, tienen dos inputs principales: las respuestas a los ítems y una *matriz-Q* (Tatsuoka, 1983), generalmente binaria y de dimensiones J (número de ítems) $\times K$ (número de atributos), que determina los atributos que se requieren para responder correctamente a cada ítem, explicitando así la estructura interna del test y la naturaleza confirmatoria de los CDMs. Más concretamente, siendo $\boldsymbol{Q} = \{q_{jk}\}$, cada *entrada-q* ($q_{jk}$) de la matriz indica si es relevante o no el atributo k para responder correctamente al ítem j. Cada ítem, por tanto, contará con su propio *vector-q* ($\boldsymbol{q}_j$), que especifica los atributos relevantes para ser contestado correctamente. El número de atributos especificados en un vector-q se representa como K_j^*, de tal modo que un determinado ítem podrá discriminar entre $2^{K_j^*}$ clases latentes diferentes. Por su parte, el principal output de los CDMs es una *clasificación de los perfiles de atributos*, una matriz de dimensiones N (número de evaluados) $\times K$ en donde se establece la probabilidad posterior de que cada evaluado domine cada uno de los atributos; probabilidad que puede ser más tarde dicotomizada en torno a un punto de corte (normalmente $0{,}5$) para determinar maestría o no maestría. En ocasiones se establece una región de incertidumbre en torno a ese punto de corte, en la cual no se hace ninguna clasificación (de la Torre y Sorrel, 2017).

La especificación de la matriz-Q suele suponer el primer paso del proceso de estimación de un CDM; es un punto muy importante y a menudo complicado. En un primer momento es necesario determinar qué y cuántos atributos se desea medir (ver Li y Suen, 2013 para una explicación más detallada). Posteriormente, se realiza la especificación de la matriz-Q inicial, estableciendo qué atributos son relevantes para responder correctamente a cada uno de los ítems. Para este proceso existen varias estrategias. Las dos más empleadas y extendidas consisten, por un lado, en construir la matriz-Q en base a la información recogida de los propios examinados cuando realizan el test, que comentan en voz alta los pensamientos que van teniendo a medida que se enfrentan a los ítems; por otro lado, el uso de un panel de expertos en la materia que describa los procesos cognitivos subyacentes requeridos para contestar a cada ítem. Ambos procedimientos se basan en información subjetiva, por lo que es razonable pensar que dan lugar a errores de especificación en la matriz-Q (Chiu, 2013; de la Torre y Chiu, 2016; de la Torre y Sorrel, 2017; Li y Suen, 2013). Si bien estos métodos, al ser confirmatorios, asumen que la matriz-Q propuesta es próxima a la correcta, los errores

que se produzcan pueden ser problemáticos, pues pueden afectar drásticamente a la estimación de los parámetros del modelo y, por ende, a la precisión en la clasificación de los perfiles de atributos (Chiu, 2013; de la Torre, 2008; de la Torre y Chiu, 2016; de la Torre y Sorrel, 2017; Gao, Miller y Liu, 2017; Romero et al., 2014; Rupp y Templin, 2008; Sorrel et al., 2016). Por tanto, es imprescindible verificar que la matriz-Q está bien especificada.

Aunque la necesidad de realizar este proceso de verificación de la matriz-Q es ampliamente reconocida, actualmente no son muchos los métodos disponibles para detectar errores de especificación (de la Torre y Chiu, 2016), sobre todo en el campo de los CDMs generales. Autores como Barnes (2010), Liu, Xu y Ying (2012), y Chiu (2013, citados en de la Torre y Chiu, 2016) han desarrollado diferentes métodos de validación de la matriz-Q que, entre otras limitaciones, sólo pueden ser aplicados a un restringido número de modelos reducidos. Esta limitación es compartida por el estudio de Romero et al. (2014), quienes desarrollan un método de validación de la matriz-Q para el modelo *Least Squares Distance* (LSDM; Dimitrov, 2007, citado en Romero et al., 2014). Además, aunque el método parece adecuado a la hora de detectar errores de especificación en la matriz-Q, no sugiere otros vectores-q con los que sustituirlos. Recientemente, Chen (2017) ha desarrollado un método muy completo para detectar y modificar errores de especificación en la matriz-Q basándose en una serie de medidas de ajuste. El método muestra un buen funcionamiento, aunque presenta ciertas limitaciones, como su gran complejidad y dificultad de implementación o la gran influencia que el número de atributos especificados en el vector-q tiene en la potencia. Además, el número de condiciones examinadas en el artículo es reducido, siendo algunas de ellas demasiado favorables (p.ej., una discriminación alta de los ítems).

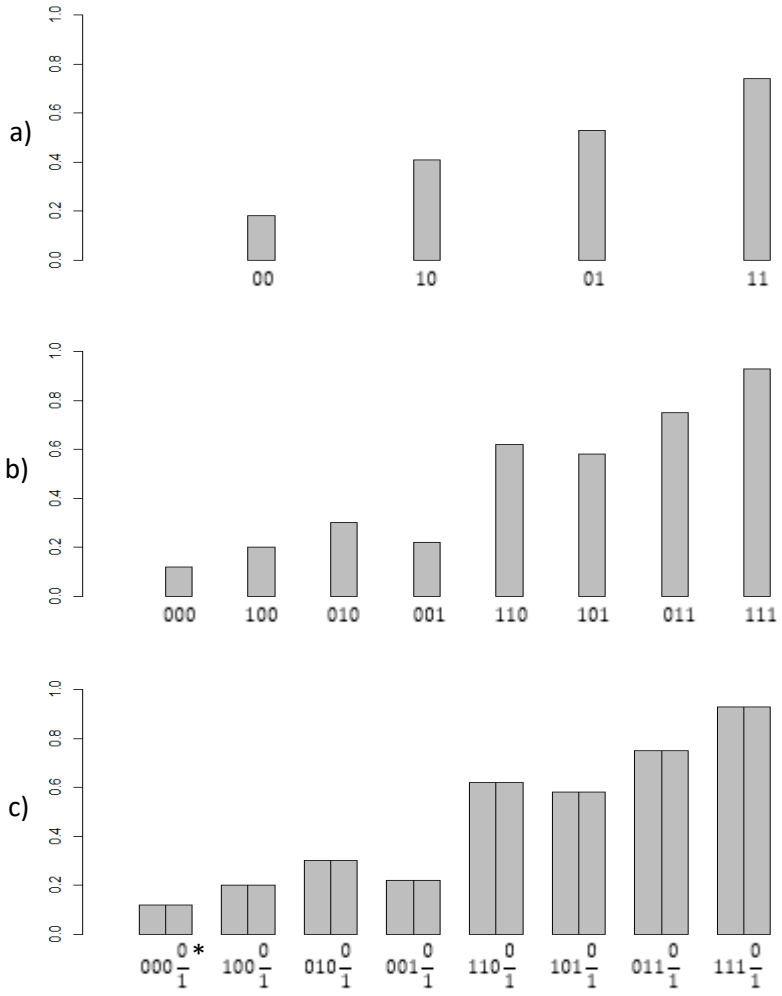
Uno de los métodos de validación de la matriz-Q que más impacto ha tenido es el basado en el *índice de discriminación*, desarrollado en un primer momento para el modelo DINA (de la Torre, 2008) y ampliado posteriormente para el modelo G-DINA (de la Torre y Chiu, 2016). El método se fundamenta en la idea de que un vector-q bien especificado será aquel que permita diferenciar con claridad las diferentes clases latentes existentes para ese ítem ($2^{K_j^*}$) según sus probabilidades de acierto. Por el contrario, un vector-q mal especificado dará lugar a probabilidades de acierto más homogéneas a través de las clases latentes especificadas. Siguiendo esta línea, el vector-q correcto para cada ítem será aquel que maximice la varianza de las probabilidades de acierto de las diferentes clases latentes

existentes para ese ítem. Esto se traducirá posteriormente en una clasificación más precisa de las personas.

Es importante destacar el compromiso entre el principio de ajuste y el principio de parsimonia. Un modelo que diferencie más clases latentes dará lugar a un mayor ajuste, ya que permite una mayor variabilidad en las probabilidades de acierto. En este sentido, los vectores-q más complejos resultarán en un mayor número de clases latentes. Por otro lado, el principio de parsimonia dicta que, a igualdad de variabilidad obtenida con dos vectores-q distintos, el más simple será el correcto. La Figura 1 ilustra esta lógica, representando las probabilidades de acierto para un ítem en función del número de atributos especificados en el vector-q (2, 3 o 4, respectivamente) y de las diferentes clases latentes posibles en cada una de estas situaciones. El vector-q con dos atributos

FIGURA 1.

Probabilidad de acierto del ítem en función de K^* y la clase latente (de la Torre y Sorrel, 2017)



Nota: La Figura 1c hace referencia a un vector-q con cuatro atributos especificados, el último de ellos irrelevante (para cada clase definida en relación a los tres primeros atributos, las dos barras representan las probabilidades de acierto en función del valor del cuarto atributo, que, para el caso representado, son iguales).

especificados (Figura 1a) no sería adecuado, puesto que la variabilidad entre las diferentes clases latentes es menor que la obtenida con el vector-q de tres atributos (Figura 1b); es decir, no cumple el principio de ajuste. El vector-q de cuatro atributos (Figura 1c), por su parte, muestra la misma variabilidad que el de tres atributos. Sin embargo, tampoco sería adecuado, debido al principio de parsimonia.

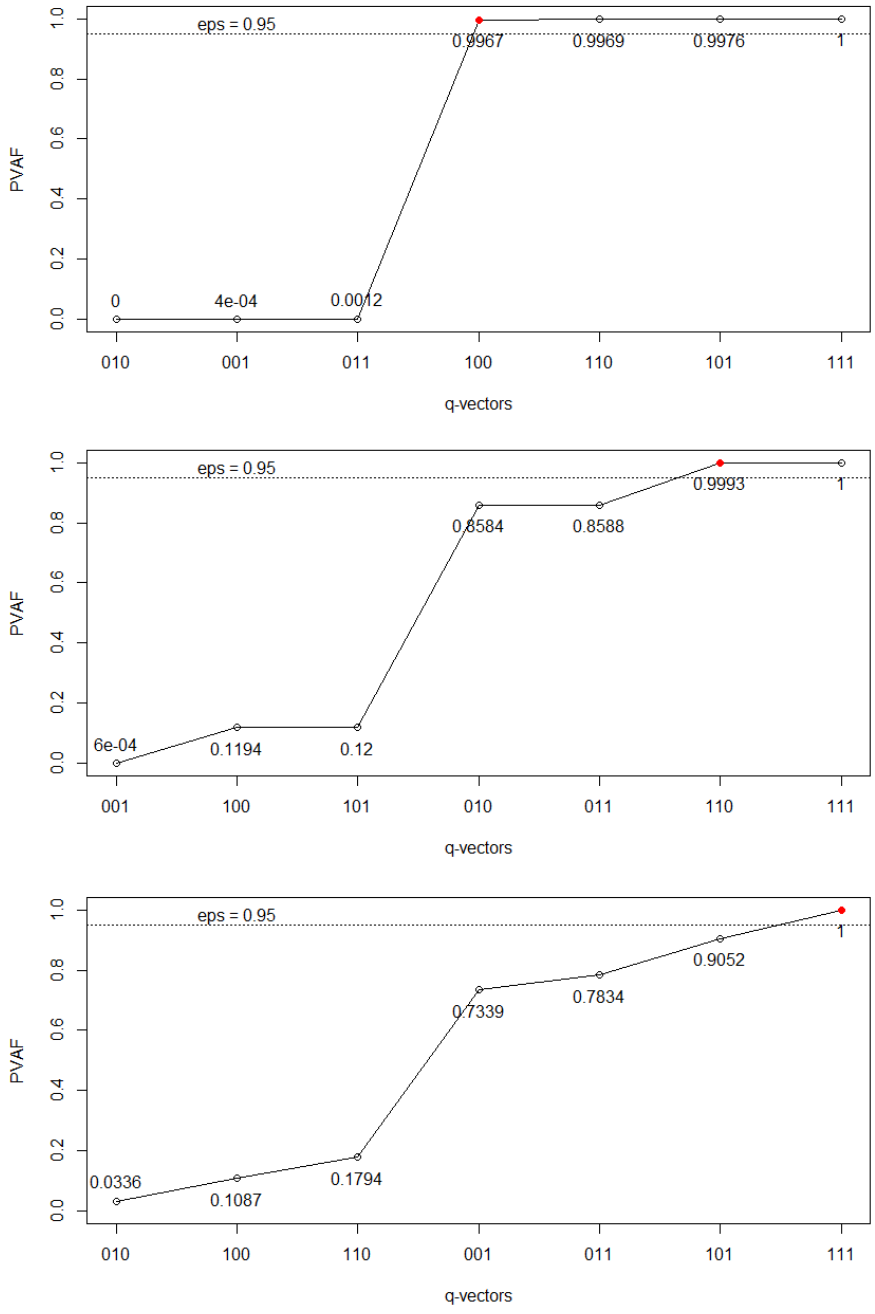
El caso más simple se correspondería con el modelo DINA. En este marco, el índice de discriminación para el ítem *j*, denominado φ_j, compara las probabilidades de acierto de dos grupos (o clases latentes) de examinados: el grupo $\eta_j = 1$, conformado por aquellos que poseen todos los atributos requeridos por el ítem j, y el grupo $\eta_j = 0$, conformado por todos los demás. El vector-q correcto será aquél que maximice las diferencias en la probabilidad de acierto entre ambos grupos. Desde un punto de vista formal, definiendo un parámetro *slip* (s_j) como la probabilidad de fallar el ítem para $\eta_j = 1$, o $P(X_j = 0|\eta_j = 1)$, y un parámetro *guessing* (g_j) como la probabilidad de acertar el ítem para $\eta_j = 0$, o $P(X_j = 1|\eta_j = 0)$, entonces el vector-q correcto será aquél que maximice la diferencia $1 - s_j - g_j$ (de la Torre, 2008).

Por otro lado, el caso más complejo se corresponde con el modelo G-DINA. En este marco, el índice de discriminación para el ítem *j* se representa como ς_j^2, el cual es la varianza de las probabilidades de éxito de los diferentes grupos latentes ponderada según la distribución muestral de esos grupos. En este caso, el vector-q con todos los atributos especificados (el que establece que todos los atributos son relevantes para contestar al ítem) será siempre el que presente el ς_j^2 máximo, puesto que la especificación de un atributo lleva a la diferenciación entre clases latentes y, por ende, a una mayor variabilidad en la probabilidad de acierto. Sin embargo, esta mayor variabilidad puede ser espuria. Por ello, tratando de encontrar un equilibrio entre el principio de ajuste y el principio de parsimonia, se escogerá como vector-q verdadero aquel que, teniendo menos atributos especificados (siendo más simple), consiga explicar una parte importante de la varianza máxima, es decir, un ς_j^2 aproximado al ς_j^2 máximo (de la Torre y Chiu, 2016). Para ello, se calcula la proporción de varianza explicada (PVAF, por sus siglas en inglés *Proportion of Variance Accounted For*) para cada vector-q, la cual se define como $\varsigma_j^2/\varsigma_{j_{1:K}}^2$, siendo $\varsigma_{j_{1:K}}^2$ el ς_j^2 máximo. El vector-q escogido será aquel que, cumpliendo $\varsigma_j^2/\varsigma_{j_{1:K}}^2 > \epsilon$, siendo ϵ (también denominado *EPS*, de *épsilon*) un valor de PVAF

previamente establecido como punto de corte, tenga un menor número de atributos especificados, es decir, sea más parsimonioso.

La Figura 2 ilustra esto para $K = 3$ y $EPS = 0{,}95$, mediante los llamados *mesaplots* (Ma y de la Torre, 2017). Los mesaplots son gráficos en los que se representan, en el eje X, los distintos vectores-q resultantes de la permutación de los K atributos y, en el eje Y, el PVAF. Los vectores-q del eje X están ordenados de menor a mayor PVAF, de tal manera que el gráfico es siempre creciente y el valor máximo, correspondiente al

FIGURA 2.
Mesaplots para tres ítems (vectores-q verdaderos, respectivamente: 100, 110, 111)



vector-q con todos los atributos especificados, será siempre de 1. La Figura 2a se corresponde con un ítem simulado cuyo vector-q verdadero es 100, es decir, solamente el primer atributo es relevante para ser contestado correctamente ($K_j^* = 1$). Se puede ver que únicamente hay un incremento abrupto de PVAF que divide el gráfico en dos "mesetas" (por esta razón se llaman mesaplots) y que coincide con el vector-q 100, el correcto. A la izquierda de éste, con unos PVAF muy reducidos, se encuentran los vectores-q que no cuentan con el atributo relevante; a la derecha se encuentran los vectores-q que, además de contar con el atributo relevante, presentan uno o más atributos no relevantes. Aunque estos últimos tienen un PVAF ligeramente mayor al vector-q verdadero, la ganancia es espuria. En este caso, y con un *EPS* de 0,95, el vector-q sugerido por el método de validación sería claramente el de 100, que es el vector más parsimonioso entre los que tienen un PVAF asociado superior a 0,95. La Figura 2b y la Figura 2c ilustran lo mismo, pero para ítems cuyos vectores-q correctos son 110 ($K_j^* = 2$) y 111 ($K_j^* = 3$), respectivamente. Cabe destacar que, en el caso de la Figura 2b, si el *EPS* escogido hubiera sido de 0,85, el vector-q sugerido por el método de validación habría sido el de 010, cometiendo una equivocación (también se estaría errando si en el caso de la Figura 2c se escogiera un *EPS* de 0,90, pues el método sugeriría el vector-q 101). Es importante remarcar que estos gráficos son meramente ilustrativos, y los ítems han sido simulados bajo unas condiciones muy favorables. El mesaplot no será siempre tan claro (no existirán siempre "mesetas" tan definidas), y el *EPS* de 0,95 no llevará siempre a sugerir el vector-q correcto.

En su artículo, de la Torre y Chiu (2016) exponen el buen funcionamiento de su método de validación tanto con un estudio de simulación como con un estudio con datos reales. Los autores escogen un *EPS* de 0,95 sin aportar ninguna justificación al respecto, y los resultados con el modelo G-DINA muestran que el método detecta un 80% de atributos mal especificados y mantiene un 98% de atributos bien especificados. Además del buen rendimiento, son tres las principales ventajas de este método: en primer lugar, su desarrollo dentro del modelo G-DINA permite una gran flexibilidad a su aplicación a otros modelos restringidos (DINA, DINO, NIDA, NIDO, etc.); por otro lado, el método no sólo identifica los vectores-q mal especificados, sino que sugiere el vector-q candidato; por último, el método está disponible en el paquete GDINA (Ma y de la Torre, 2017) del software estadístico R (R Core Team, 2016) y tiene un bajo coste computacional, lo cual lo convierte en uno de los más accesibles.

Sin embargo, los estudios de de la Torre (2008) y de la Torre y Chiu (2016) también presentan algunas limitaciones. En primer lugar, las condiciones de simulación exploradas en ambos son bastante restringidas y, sobre todo, favorables, alejadas de las condiciones habituales de los contextos aplicados. Por ejemplo, el tamaño muestral empleado es de 2000 evaluados, cuando en la mayoría de los estudios empíricos la muestra es menor de 1300 personas (ver el apartado de Diseño del Estudio 1). Además, en ningún momento se especifica un método de elección del punto de corte del PVAF (es decir, el *EPS*). Aunque el punto de corte por defecto está situado en 0,95, en parte con el objetivo de que el vector-q contemple al menos un 95% de la varianza máxima, éste es un valor arbitrario e injustificado (Chen, 2017; Wang et al., 2018).

El objetivo del presente trabajo es estudiar el funcionamiento del método de validación empírica de la matriz-Q basado en el índice de discriminación para el marco G-DINA en condiciones más variadas y representativas del contexto aplicado que las exploradas en el artículo original. En relación a este objetivo, se valorará si existe un *EPS* que pueda recomendarse de forma generalizada a través de las diferentes condiciones.

Se realizaron dos estudios de simulación para evaluar el funcionamiento del método de validación de la matriz-Q en ausencia de errores de especificación en la matriz-Q (Estudio 1) o en presencia de éstos (Estudio 2). La Tabla 1 resume las condiciones empleadas en cada uno de los estudios en comparación con las condiciones empleadas en el artículo original de de la Torre y Chiu (2016). En los apartados de método de cada estudio se ofrece una justificación para los niveles seleccionados y una explicación

TABLA 1.
Resumen de las condiciones de simulación

Factores	de la Torre y Chiu (2016)*	Estudio 1	Estudio 2
Estructura atributos	Higher-order	Uniforme	
J	30	15, 30, 60	
N	2000	500, 1000, 2000	
IQ	Media: $P(\mathbf{1}) = Unif(0{,}7; 0{,}9)$; $P(\mathbf{0}) = Unif(0{,}1; 0{,}3)$	Alta: $P(\mathbf{1}) = Unif(0{,}8; 1{,}0)$; $P(\mathbf{0}) = Unif(0{,}0; 0{,}2)$ Media: $P(\mathbf{1}) = Unif(0{,}7; 0{,}9)$; $P(\mathbf{0}) = Unif(0{,}1; 0{,}3)$ Baja: $P(\mathbf{1}) = Unif(0{,}6; 0{,}8)$; $P(\mathbf{0}) = Unif(0{,}2; 0{,}4)$	
EPS	0,95	0,60; 0,65; 0,70; 0,75; 0,80; 0,85; 0,90; 0,95; 0,99	
Modelo	G-DINA	G-DINA	
K	5	5	
% errores matriz-Q	5	0	10

* Estudio 2 del artículo

detallada para cada factor. El objetivo del Estudio 1 consiste en examinar en qué grado el método de validación de la matriz-Q introduce modificaciones incorrectas cuando el modelo es correcto, y en qué condiciones se dan en mayor medida. En el estudio 2 se examina el funcionamiento del método de validación introduciendo errores en la matriz-Q, cubriendo así el posiblemente típico escenario de los contextos aplicados, en los que los expertos especifican incorrectamente algunas de las entradas-q. En ambos estudios se examinará qué *EPS* presenta mejores resultados bajo las diferentes condiciones.

Estudio 1: matriz-Q sin errores de especificación

Método

Diseño. El número de atributos fue fijado a $K = 5$. Se empleó un diseño factorial mixto, con el *EPS* como factor intra-sujetos (con nueve niveles: de 0,60 a 0,95, aumentando 0,05 cada vez, y 0,99) y tres factores inter-sujetos: longitud del test, tamaño muestral y calidad de los ítems. Los niveles de los factores inter-sujetos fueron escogidos con el objetivo de que fueran representativos de los rangos de valores habitualmente encontrados en trabajos empíricos. Para cada uno de ellos se trató de seleccionar un nivel bajo, medio y alto, dando como resultado un diseño intra-sujetos $3 \times 3 \times 3$ ($J \times N \times IQ$), con un total de 27 combinaciones. A continuación, se muestra una breve descripción de la lógica empleada para la selección de los niveles de los diferentes factores.

1. *Longitud del test (J)*: con niveles de 15, 30 y 60 ítems. En la literatura revisada especializada es común el uso de tests que cuentan con entre 11 y 30 ítems (p.ej., Chen, 2017; Chen y de la Torre, 2013; Chen, de la Torre y Zhang, 2013; Chiu, 2013; de la Torre, 2008, 2011; de la Torre y Chiu, 2016; de la Torre y Douglas, 2004; Ma y de la Torre, 2016; Romero et al., 2014; Sorrel et al., 2016). También se emplean tests con más de 30 ítems (p.ej., de la Torre, van der Ark y Rossi, 2015; Templin y Henson, 2006), llegando incluso a los 90 (de la Torre, 2008).
2. *Tamaño muestral (N)*: con niveles de 500, 1000 y 2000 participantes. Si bien hay estudios que cuentan con muestras elevadas, de más de 2000 participantes (p.ej., de la Torre, 2008; de la Torre y Douglas, 2004; Romero et al., 2014), muchos emplean muestras de entre 700 y 1300 personas (p.ej., Chen, 2017; de la Torre, 2011; Ma y de la Torre, 2016) e incluso de menos de 600 (p.ej., Chen, 2017; Chen y de la Torre, 2013; Chen, de la Torre y Zhang, 2013; Chiu, 2013; de la Torre, 2011; de la Torre y Chiu, 2016; Sorrel et al., 2016; Templin y Henson, 2006).

3. *Calidad de los ítems (IQ)*: la calidad de los ítems se operativizó a través de la discriminación, calculada como la diferencia de probabilidades de acierto del ítem de la clase latente con los atributos relevantes, $P(\mathbf{1})$, y la que no los tiene, $P(\mathbf{0})$ (para el caso de $K_j^* = 3$, $P(\mathbf{1}) = \{111\}$ y $P(\mathbf{0}) = \{000\}$). Los niveles escogidos de calidad de los ítems fueron de 0,4, 0,6 y 0,8. Estos niveles se basan, por un lado, en los escogidos por otros autores en sus estudios de simulación (p.ej., Ma, Iaconangelo y de la Torre, 2016; Sorrel, Abad, Olea, de la Torre y Barrada, 2017) y, por otro, en las discriminaciones medias encontradas en estudios empíricos, en los que se encuentran desde ítems de calidad alta en el ámbito de la medición educativa (p.ej., Chen, 2017; de la Torre, 2008) hasta discriminaciones bajas, más similares a las encontradas en otros contextos de aplicación, como el ámbito clínico u organizacional (p.ej., Sorrel et al., 2016; Templin y Henson, 2006).

Para cada longitud del test, se estableció siempre una matriz-Q con igual número de ítems de 1, 2 y 3 atributos. La matriz-Q con $J = 30$ (*Q30*) se muestra en la Tabla 2, y fue la misma que la empleada por de la Torre y Chiu (2016). La matriz-Q con $J = 60$ (*Q60*) fue *Q30* duplicada, mientras que la matriz-Q con $J = 15$ (*Q15*) contenía el subconjunto de ítems marcado con asterisco en la Tabla 2.

TABLA 2.
Matriz-Q para los datos simulados (J = 30)

Ítem	α_1	α_2	α_3	α_4	α_5	Ítem	α_1	α_2	α_3	α_4	α_5
1*	1	0	0	0	0	16	0	1	0	1	0
2*	0	1	0	0	0	17	0	1	0	0	1
3*	0	0	1	0	0	18*	0	0	1	1	0
4*	0	0	0	1	0	19	0	0	1	0	1
5*	0	0	0	0	1	20*	0	0	0	1	1
6	1	0	0	0	0	21*	1	1	1	0	0
7	0	1	0	0	0	22	1	1	0	1	0
8	0	0	1	0	0	23*	1	1	0	0	1
9	0	0	0	1	0	24	1	0	1	1	0
10	0	0	0	0	1	25	1	0	1	0	1
11*	1	1	0	0	0	26*	1	0	0	1	1
12	1	0	1	0	0	27*	0	1	1	1	0
13	1	0	0	1	0	28	0	1	1	0	1
14*	1	0	0	0	1	29	0	1	0	1	1
15*	0	1	1	0	0	30*	0	0	1	1	1

* Ítems escogidos para *Q15*

Generación de datos. Para generar ítems de distinta calidad se manipularon las distribuciones de las probabilidades de acierto de la clase latente con todos los atributos relevantes, $P(\mathbf{1})$, y las probabilidades de acierto de la clase latente con ninguno, $P(\mathbf{0})$. Para cada uno de los niveles de la variable se simuló la misma distribución uniforme para todos los ítems. En concreto: calidad baja: $P(\mathbf{1}) = Unif(0{,}6;\ 0{,}8)$ y $P(\mathbf{0}) = Unif(0{,}2;\ 0{,}4)$; calidad media: $P(\mathbf{1}) = Unif(0{,}7;\ 0{,}9)$ y $P(\mathbf{0}) = Unif(0{,}1;\ 0{,}3)$; calidad alta: $P(\mathbf{1}) = Unif(0{,}8;\ 1)$ y $P(\mathbf{0}) = Unif(0;\ 0{,}2)$. Para el resto de las clases latentes (aquéllas que tienen algunos de los atributos relevantes y otros no) se simularon las probabilidades de acierto de forma que ésta se incrementara a medida que aumenta el número de atributos dominados (i.e., restricción de monotonicidad). Es decir, una clase latente que domine más atributos que otra tendrá siempre mayores probabilidades de acertar el ítem. La restricción de monotonicidad puede establecerse en el paquete G-DINA de R, y actúa estableciendo probabilidades de éxito aleatorias para cada clase latente, respetando dicha restricción.

La distribución de las clases latentes fue uniforme. De la Torre y Chiu (2016) emplean un modelo higher-order, aunque no justifican su uso. En la literatura es más común generar los datos desde una estructura uniforme (p.ej., Sorrel, Abad, Olea, de la Torre y Barrada, 2017; Sorrel, de la Torre, Abad y Olea, 2017), al no haber motivos para asumir una estructura concreta.

Para cada una de las condiciones resultantes de la combinación de los niveles de los factores se generaron 100 bases de datos. El código empleado en la generación de datos fue escrito íntegramente en R (R Core Team, 2016), empleando funciones de elaboración propia y del paquete GDINA (Ma y de la Torre, 2017).

Variables dependientes y análisis de datos. Para evaluar el funcionamiento del método de validación se utilizó la tasa de verdaderos positivos (*true positive rate*, o TPR), que indica la proporción de entradas-q correctamente especificadas que son retenidas (es decir, no modificadas).

Con el objetivo de comprobar el peso de cada uno de los factores en el funcionamiento del método de validación, se llevó a cabo un ANOVA de medidas repetidas con un factor intra-sujetos (*EPS*, con 9 niveles) y tres factores inter-sujetos (*J*, *N* e *IQ*). Debido al elevado tamaño muestral, todos los efectos resultaron significativos (se examinaron los estadísticos *F* con grados de libertad modificados y los estadísticos de la aproximación multivariada, debido al incumplimiento del supuesto de esfericidad). Por

esta razón, se empleó la medida de tamaño del efecto eta-cuadrado parcial (η_p^2) para establecer el impacto de las variables. Se consideraron los efectos con un valor superior a 0,1379 el cual suele ser tomado como umbral de tamaño del efecto grande (Cohen, 1988).

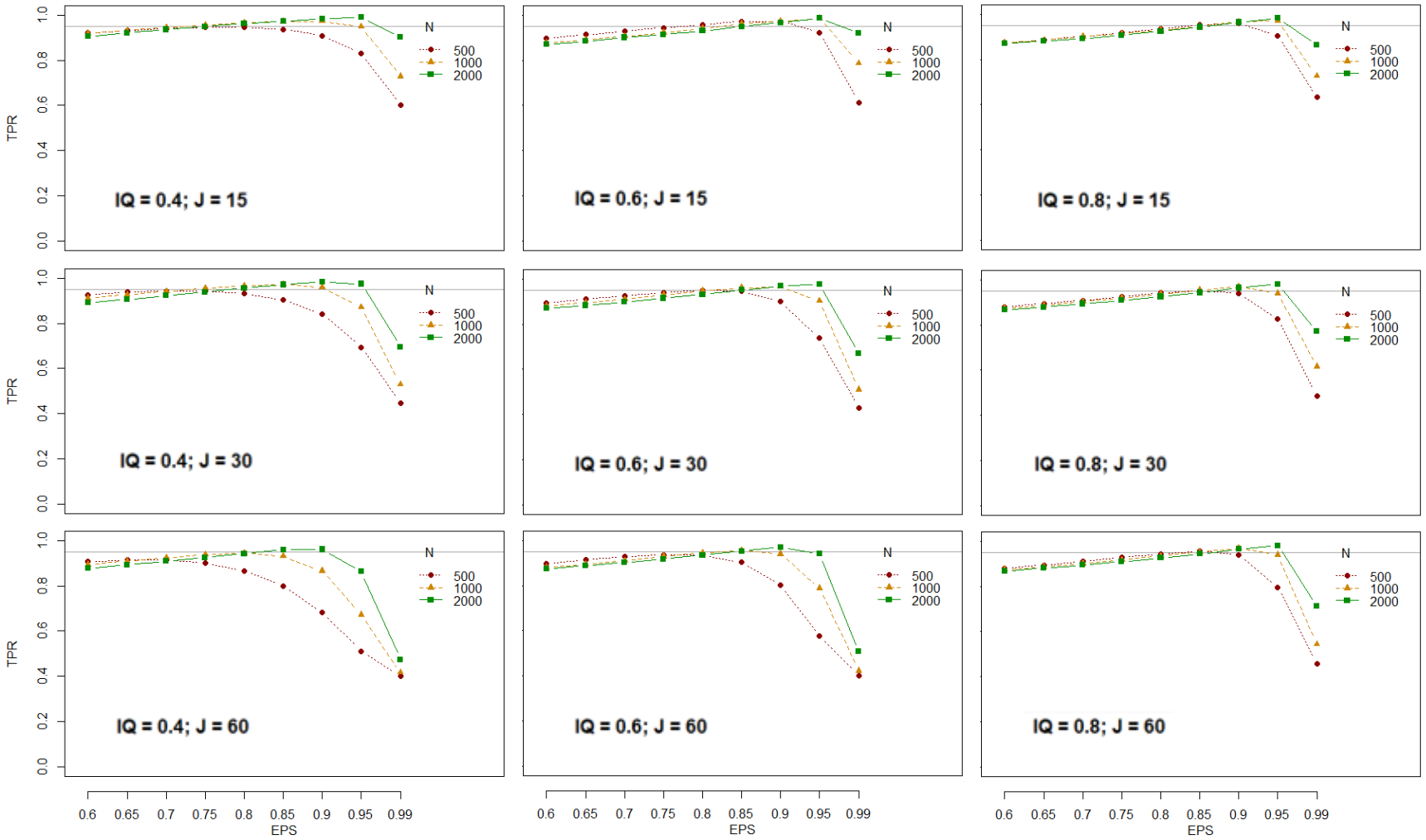
Resultados

En el ANOVA, todos los efectos obtuvieron una η_p^2 por encima del punto de corte seleccionado. Al ser todos ellos relevantes, se pasará a comentar el efecto de la interacción de orden mayor ($eps \times J \times N \times IQ$). La Figura 3 muestra la TPR para cada una de las combinaciones de condiciones. A grandes rasgos, la TPR tiende a ser mayor a medida que aumenta el tamaño muestral, aumenta la discriminación y se reduce la longitud del test. Por otro lado, el *EPS* de 0,99 muestra resultados claramente diferentes y peores que el resto de los valores de *EPS*. Obviando este valor se ve que, en general, un tamaño muestral elevado se asocia a TPR elevados en todas las condiciones, con la excepción de la condición de baja discriminación y larga longitud del test. El efecto del tamaño muestral se va incrementando a medida que aumenta la longitud del test y, sobre todo, se reduce la discriminación de los ítems. En este sentido, en condiciones de discriminación baja hay una mayor diferencia entre los diferentes niveles del tamaño muestral que en condiciones donde la discriminación es alta. La longitud del test también gana influencia en los resultados a medida que se reduce la discriminación, observándose unas mayores diferencias en la TPR entre sus diferentes niveles.

Todas las variables estudiadas tuvieron, por tanto, un efecto relevante en la TPR, tomadas tanto de forma independiente como en conjunto. Bajo condiciones de elevado tamaño muestral, alta discriminación y longitud del test corta, el método de validación de la matriz-Q demostró tener una alta TPR (i.e., no se modificaron entradas-q correctamente especificadas), especialmente con *EPS* elevados (0,85–0,95). Sin embargo, con estos *EPS*, para muestras pequeñas, tests largos y, especialmente, ítems poco discriminativos, el método presentó tasas inaceptables de TPR.

En cualquier caso, en todas las condiciones se obtuvo una TPR mayor de 0,90 con al menos un *EPS*. Con muestras grandes 0,95 parece el mejor en la mayoría de los casos, pero el valor de *EPS* óptimo tiende a reducirse a medida que se reduce el tamaño de la muestra, aumenta la longitud del test y disminuye la discriminación de los ítems. Esto señala la necesidad de escoger un *EPS* determinado en función de las condiciones concretas.

FIGURA 3.
TPR en función del *EPS*, *N*, *J* e *IQ*



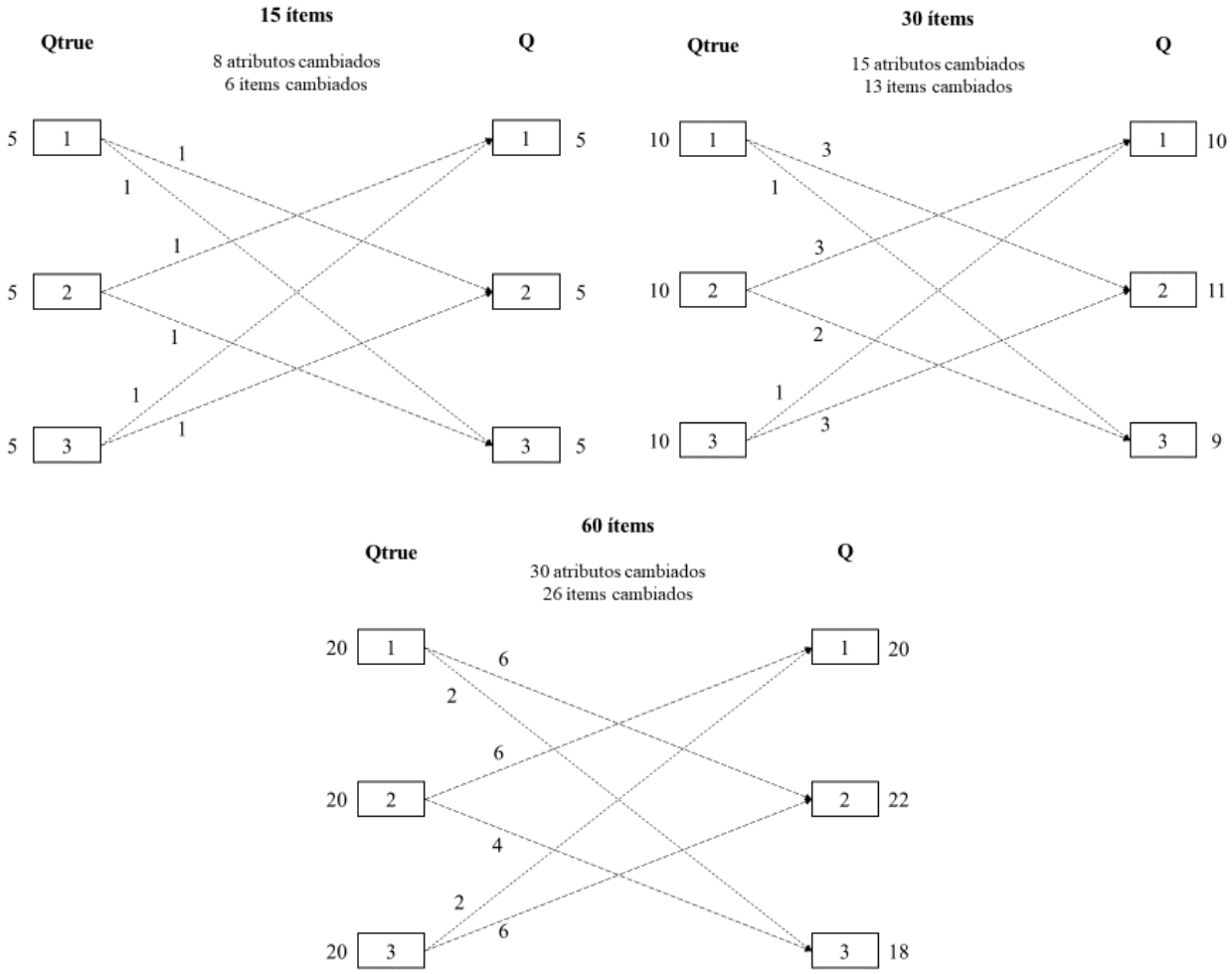
Estudio 2: matriz-Q con errores de especificación

Método

Diseño. Para el Estudio 2 se manipularon los mismos factores que en el primero (ver Tabla 1): tamaño muestral, calidad de los ítems y longitud del test. La diferencia entre ambos estudios es que en éste se incluyó un 10% errores de especificación. De la Torre y Chiu (2016), en su estudio, trabajan con un 5% de erroresde especificación. La razón de escoger un porcentaje de errores más elevado que el empleado por los autores, pero igualmente verosímil, radica en aportar una mayor robustez a los resultados, empleando una condición menos favorable.

Para cada una de las 100 réplicas por condición, se generó una matriz-Q distinta, siempre manteniendo el porcentaje de errores correspondiente. Partiendo de las matrices-Q (*Q15*, *Q30* y *Q60*) definidas en el primer estudio, las modificaciones se incluyeron de tal forma que la proporción de ítems midiendo un número de atributos concreto se mantuviera similar. La Figura 4 muestra un diagrama del modo en el que se introdujeron los cambios. En el diagrama, *Qtrue* representa la matriz-Q verdadera (i.e., la empleada para simular las respuestas de los participantes) y *Q* la matriz-Q con errores de especificación. Los números situados en los recuadros simbolizan el número de atributos especificados en el vector-q (K_j^*), y los números situados a su lado representan el número de ítems con dichos atributos en su vector-q. Los números correspondientes a cada flecha representan cuántos vectores-q de *Qtrue* con un determinado K_j^* pasaron a tener K_j^* en *Q*. Por ejemplo, para el caso de 15 ítems, *Qtrue* contaba con cinco ítems especificados con un único atributo ($K_j^* = 1$), cinco ítems especificados con dos atributos ($K_j^* = 2$) y cinco ítems especificados con tres atributos ($K_j^* = 3$). De los ítems con un único atributo en la *Qtrue*, uno pasó a tener dos atributos (se le añadió un atributo erróneamente) y otro pasó a tener tres atributos (se le añadieron dos atributos erróneamente). Los tres ítems restantes con un atributo no fueron modificados. Los cambios se cuentan a nivel de atributo; es decir, aunque sólo hubo 6 ítems modificados para la matriz-Q de 15 ítems, en dos de ellos se modificaron dos atributos: el ítem de un atributo que pasó a tener tres (es decir, se le añadieron dos atributos), y el ítem de tres atributos que pasó a tener uno (se le quitaron dos atributos). Por lo tanto, hubo un total de 8 atributos modificados.

FIGURA 4.
Errores introducidos en las diferentes matrices-Q



Variables dependientes y análisis de datos. Al incluir errores de especificación en la matriz-Q, el método de validación debería sugerir cambios correctos. Por ello, para evaluar el funcionamiento del método de validación se utilizaron la tasa de verdaderos positivos (*true positive rate* o TPR) y la tasa de verdaderos negativos (*true negative rate* o TNR). La definición de TPR coincide con la vista en el Estudio 1 (i.e., proporción de entradas-q correctamente especificadas que son retenidas). Por el contrario, la TNR indica la proporción de entradas-q erróneamente especificadas que son modificadas (corregidas). La Figura 5 representa la naturaleza de las dos medidas, siendo *Q** la matriz-Q sugerida por el método de validación. Con el objetivo de explorar el efecto de las distintas variables en la TPR y la TNR, se realizaron dos ANOVA de medidas repetidas como los descritos en el Estudio 1.

FIGURA 5.
Definición de la TPR y la TNR

Condición	Nº de casos que cumplen la condición
$q_{true} = q = q^*$	A
$q_{true} = q \neq q^*$	B
$q_{true} = q^* \neq q$	C
$q_{true} \neq q^* = q$	D

$$TPR = \frac{A}{A+B};\ TNR = \frac{C}{C+D}$$

Nota: q_{true}, q y q^* hacen referencia a una entrada-q de las matrices Qtrue, Q y Q*, respectivamente. Se ha representado de esta manera por motivos de claridad.

Es importante destacar que la TPR y la TNR son dos criterios complementarios de calidad del método de validación de la matriz-Q; es decir, para considerar que el método funciona bien bajo ciertas condiciones, se tienen que dar una alta TPR y una alta TNR de forma conjunta. Para interpretar estas medidas hay que tener en cuenta lo siguiente. De la definición de la TPR se deduce que, si no se emplease ningún método de validación de la matriz-Q, ésta sería igual a 1, puesto que ninguna de las entradas (las correctamente especificadas entre ellas) sería modificada. La línea base, es decir, el valor de comparación de la TPR debe ser, por tanto, 1. Cualquier valor por debajo de éste supondrá que el método está empeorando las cosas con respecto a "no hacer nada". La TNR, por su parte, representa la mejora o aportación que ofrece el hecho de aplicar el método de validación en comparación a no aplicarlo. Si no se aplicase ningún método de validación, se modificaría un 0% de las entradas incorrectamente especificadas. Por tanto, aunque cualquier TNR mayor de 0 implicará que el método está aportando valor, lo ideal será corregir la mayoría de entradas-q erróneas. Considerando los resultados de de la Torre y Chiu (2016) comentados anteriormente, en el presente estudio se considerarán como adecuados valores mayores de 0,95 para la TPR y de 0,80 para la TNR.

Aparte de estas medidas, se incluyó como medida de fiabilidad la proporción de atributos correctamente clasificados (PACC) a lo largo de la matriz $N \times K$ de clasificación de perfiles de atributos. Es decir, la PACC refleja, para el total de atributos estimados para los evaluados, la proporción de atributos bien clasificados. Para evaluar

el efecto del método de validación de la matriz-Q en las tasas de clasificación, se calculó un índice $PACC_D$, que consistía en la diferencia entre la PACC resultante tras estimar el modelo con una matriz-Q corregida mediante el método de validación (*Q**), con un *EPS* determinado ($PACC_{EPS}$), y la PACC resultante tras estimar el modelo con la matriz-Q con errores (*Q*), que es tomada como línea base ($PACC_B$). Un $PACC_D$ positivo implicará, por tanto, que *Q** dará lugar a una mejor clasificación que *Q*. También se calculó la máxima diferencia en PACC ($PACC_{D\text{-}MAX}$): la resta entre la PACC resultante de estimar el modelo con la matriz-Q verdadera (*Qtrue*), que será presumiblemente la condición más idónea y, por tanto, la que dé como resultado la PACC óptima ($PACC_O$), y la $PACC_B$. La $PACC_{D\text{-}MAX}$ reflejará la máxima ganancia posible en clasificación.

Por último, se empleó otra medida de incremento relativo en la precisión. Concretamente, teniendo en cuenta las dos líneas de referencia, tanto superior ($PACC_O$) como inferior ($PACC_B$), se calculó la ratio de atributos correctamente clasificados (RACC), que representa la ratio entre la mejora de fiabilidad obtenida con el método y la máxima mejora posible.

Las diferentes medidas explicadas obedecen a las siguientes fórmulas:

$$PACC_D = PACC_{EPS} - PACC_B$$

$$PACC_{D-MAX} = PACC_O - PACC_B$$

$$RACC = \frac{PACC_{EPS} - PACC_B}{PACC_O - PACC_B} = \frac{PACC_D}{PACC_{D-MAX}}$$

Por último, se llevó a cabo una regresión lineal múltiple con el objetivo de hallar una fórmula predictiva del *EPS* óptimo (aquél que mostró una mayor $PACC_D$) en función del tamaño muestral, la longitud del test y la discriminación de los ítems. La regresión se realizó con el método de introducción por pasos hacia delante. Debido a la naturaleza acotada del *EPS* (con un valor mínimo de 0 y un valor máximo de 1), la regresión lineal se realizó sobre el *logit* del *EPS* óptimo:

$$\text{logit}(EPS) = \log\left(\frac{EPS}{1 - EPS}\right).$$

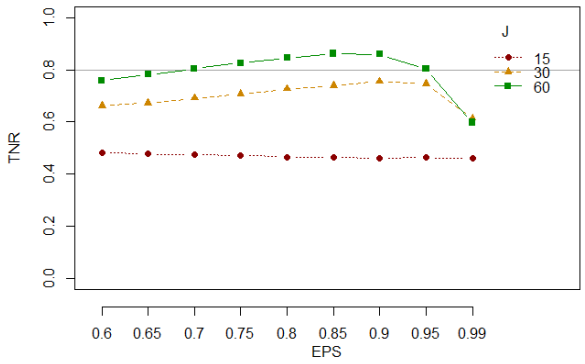
Resultados

Tasas de clasificación de los atributos de la matriz-Q. En lo referente a la TPR, los resultados fueron muy similares a los descritos en el Estudio 1. Este resultado

era esperable, pues las matrices-Q, con un 10% de errores de especificación, no difieren en exceso de las verdaderas (comparten con ellas el 90% restante de las entradas-q). Todos los tamaños del efecto fueron mayores a 0,1379.

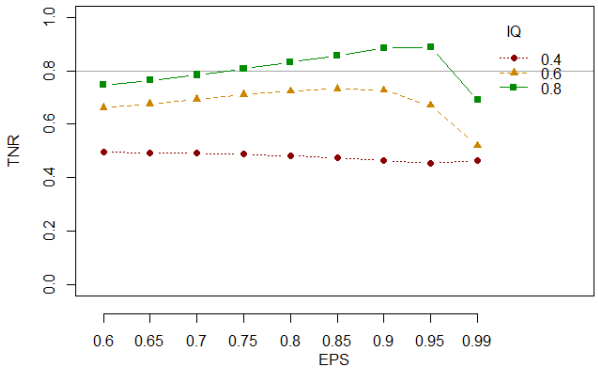
En cuanto a la TNR, el ANOVA mostró que únicamente las interacciones del *EPS* con la longitud del test, por un lado, y con la discriminación de los ítems, por otro, obtuvieron una η_p^2 mayor a 0,1379 La Figura 6 muestra el primero de estos efectos, en donde se ve que, a medida que aumenta la longitud del test, aumenta la TNR. Esto ocurre de forma consistente para todos los *EPS* (a excepción del *EPS* de 0,99, que viene presentando desde el Estudio 1 resultados mucho peores que el resto de los niveles). Mientras que con longitudes del test media y alta la TPR varía a medida que lo hace el *EPS*, con tests cortos la TPR mantiene el mismo valor con independencia del *EPS*.

FIGURA 6.
TNR en función del *EPS* y *J*



En la Figura 7 se muestra el efecto referido a la discriminación de los ítems; a medida que aumenta la discriminación, aumenta la TNR para todos los *EPS*. Además, mientras que con discriminaciones media y alta la TPR aumenta a medida que lo hace el *EPS*, con discriminación baja la TPR se mantiene constante a lo largo de los *EPS*. De aquí se desprende que la TNR será mayor en condiciones de alta discriminación y tests largos.

FIGURA 7.
TNR en función del *EPS* e *IQ*



La Figura 8 muestra un gráfico en el que se combina la TPR y la TNR. En cada cuadrante se presentan tres condiciones de tamaño muestral para una determinada combinación de calidad de los ítems y longitud del test. En la figura se ve que las tres condiciones de la esquina superior izquierda (i.e., baja discriminación, test corto) son inaceptables debido a la baja TNR. Las condiciones de la esquina inferior izquierda (i.e., baja discriminación, test largo) y la esquina superior derecha (i.e., alta discriminación, test corto) sólo presentarían resultados aceptables con un tamaño muestral elevado. El resto de las condiciones muestran un mejor funcionamiento general. Por tanto, el método de validación presenta un funcionamiento aceptable cuando, al menos, se cumple que la discriminación media de los ítems del test es igual o mayor a 0,6 y la longitud del test es igual o mayor a 30 ítems. Bajo estas condiciones, si el tamaño muestral es elevado, el *EPS* de 0,95 presenta un buen funcionamiento, muy similar al del *EPS* de 0,90. Si el tamaño muestral es mediano (en torno a 1000 participantes), el *EPS* más adecuado se situará también entre 0,90 y 0,95, este último sobre todo recomendado cuando la discriminación y la longitud del test son más elevados. Si el tamaño muestral es reducido (en torno a 500 participantes), el *EPS* más adecuado pasa a ser el de 0,85.

Tasas de clasificación de los atributos de los individuos. Las recomendaciones hechas hasta ahora están basadas en la TPR y la TNR de forma conjunta. Si bien estas medidas son relevantes, a nivel práctico parece más adecuado tomar decisiones en base a un criterio que mida de forma directa el impacto de los diferentes factores en la clasificación correcta de los diferentes sujetos, es decir, la fiabilidad. La Figura 9 muestra cómo varía la $PACC_D$ en función de los distintos factores y los distintos *EPS* (se ha eliminado de los gráficos el *EPS* de 0,99 al haberse comprobado a lo largo de los anteriores análisis que da lugar a muy malos resultados). Las líneas continuas representan la $PACC_D$ resultante en cada condición. Un valor cercano a cero indicará que, para esa condición y ese *EPS*, la $PACC_{EPS}$ es muy similar a la $PACC_B$, es decir, que la matriz-Q corregida con el método de validación y ese *EPS* (Q*) no proporciona una mejor clasificación que la matriz-Q errónea (Q). Si, por ejemplo, el valor es de 0,02, querrá decir que con Q* se ha clasificado correctamente un 2% de atributos más que con Q. Este 2% se corresponderá con un diferente número de atributos en función de la N (se calcula sobre la matriz $N \times K$, siendo, en este estudio, $K = 5$). Este valor también puede ser negativo, indicando que Q* está llevando a una peor clasificación que Q. Con líneas discontinuas se ha dibujado la $PACC_{D\text{-}MAX}$ para cada condición, con el objetivo de tener una línea de referencia superior.

FIGURA 8.
TPR y TNR en función del *EPS*, *N*, *J* e *IQ*

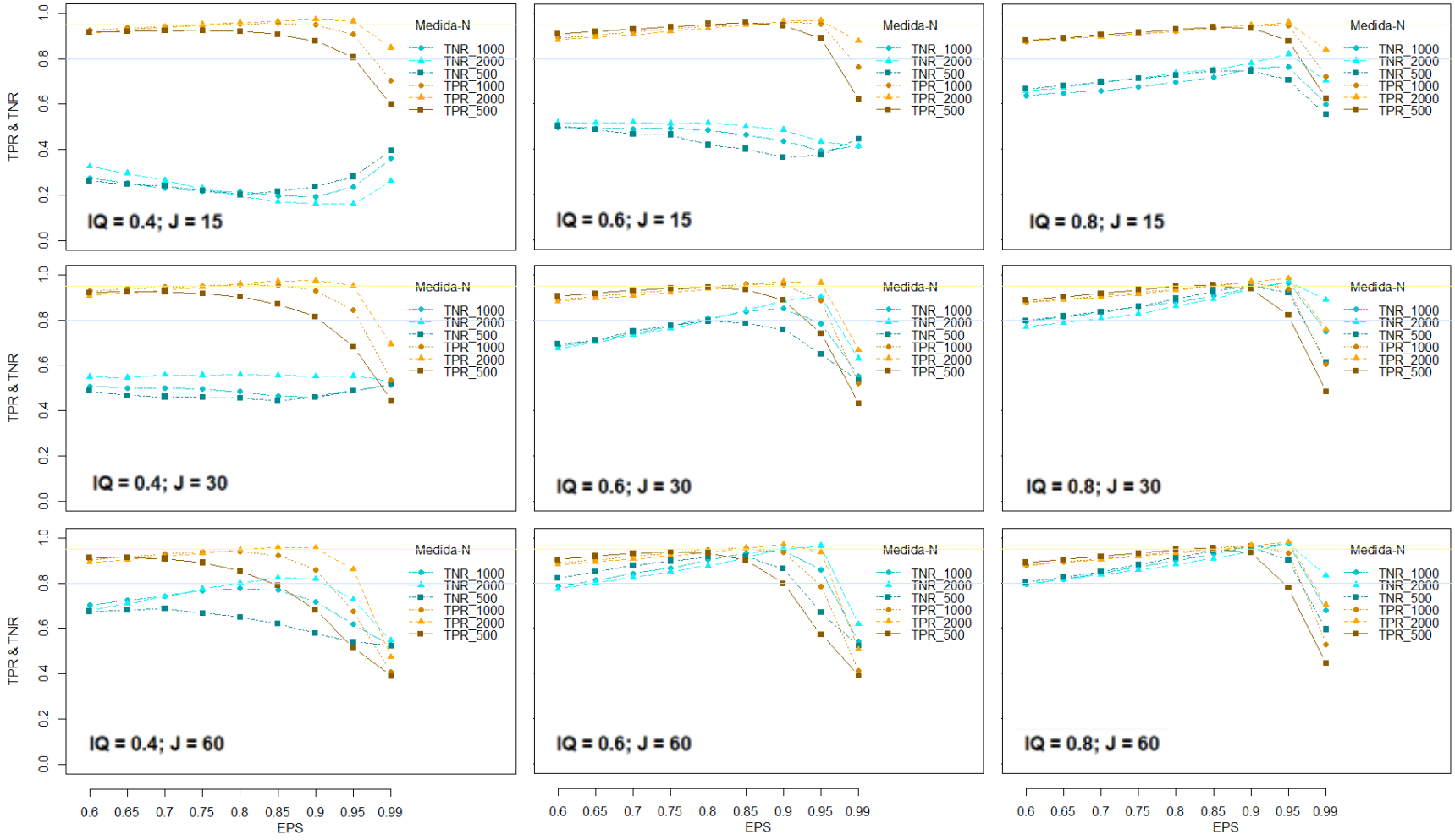
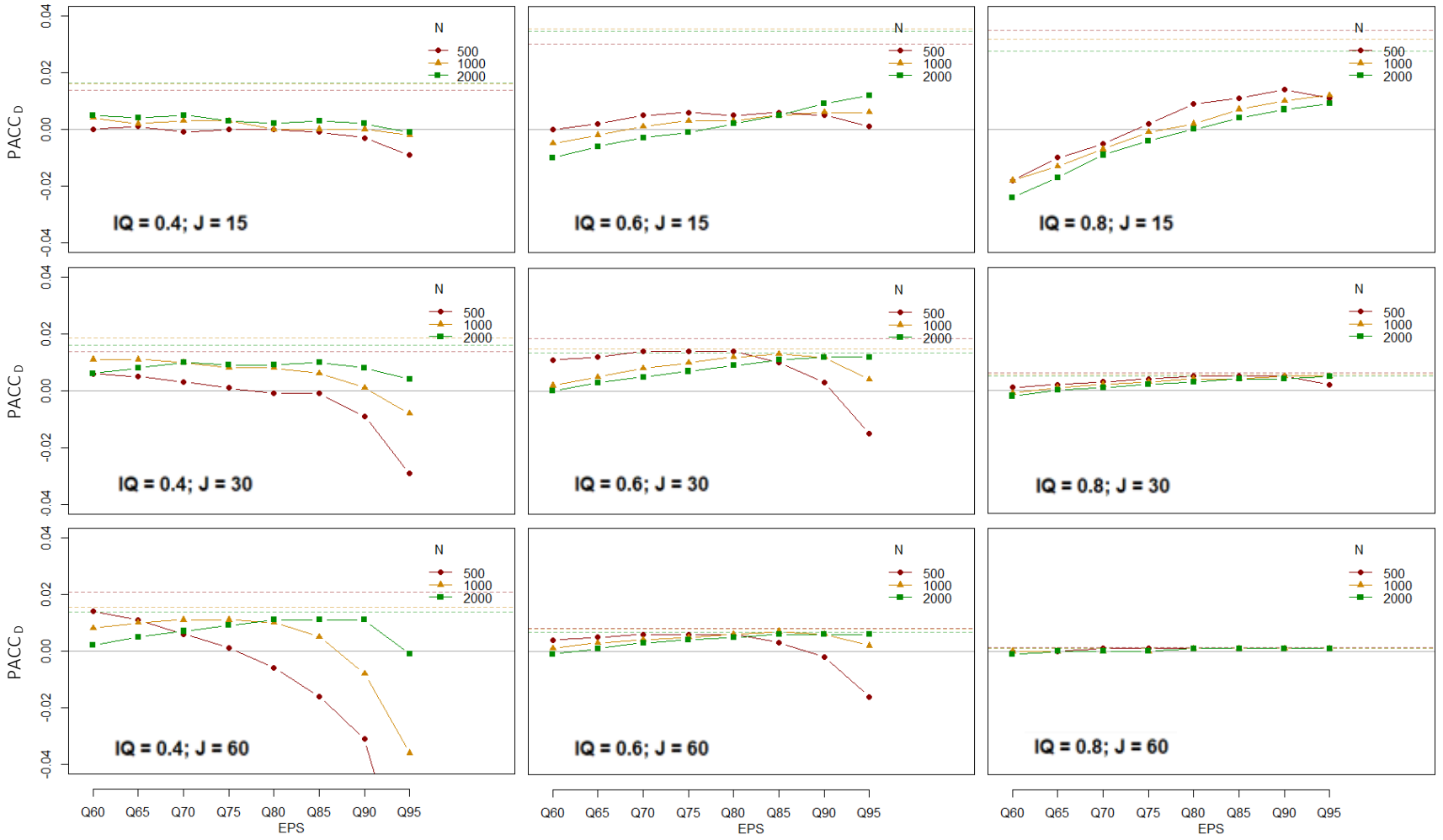


FIGURA 9.
$PACC_D$ en función del *EPS*, *N*, *J* e *IQ*



En primer lugar, se observa que, a medida que aumenta la discriminación de los ítems, la $PACC_D$ y la $PACC_{D\text{-}MAX}$ se reducen considerablemente, siempre y cuando el test no sea corto (15 ítems), donde ocurre lo contrario. Que se reduzcan la $PACC_D$ y la $PACC_{D\text{-}MAX}$ quiere decir que el efecto de los errores que haya en la matriz-Q tiene menos relevancia, pues matrices-Q ligeramente distintas van a dar resultados muy parecidos. Por tanto, parece que, a medida que la discriminación y la longitud del test van aumentando, el efecto de los posibles errores de la matriz-Q se ve mitigado. Es decir, en condiciones donde hay más información disponible, matrices-Q ligeramente diferentes darán lugar a resultados muy similares. Cuando las condiciones son menos idóneas (una discriminación o una longitud del test baja), los errores de especificación presentes en la matriz-Q ganan relevancia, haciendo que la $PACC_B$ disminuya y, por tanto, aumenten la $PACC_D$ y la $PACC_{D\text{-}MAX}$. Es decir, en condiciones desfavorables va a ser más relevante, en términos de clasificación, que haya o no errores en la matriz-Q, puesto que éstos van a tener una mayor influencia en los resultados.

Estos hallazgos presentan un ligero contraste con los resultados de la TPR y la TNR, donde se vio que, a medida que las condiciones se hacían más favorables, la especificación de la matriz-Q era más precisa. Es decir, condiciones favorables darán lugar a más aciertos en la matriz-Q, pero a una menor ganancia en la clasificación de las personas. Por el contrario, en condiciones desfavorables, aunque se cometan errores a la hora de especificar la matriz-Q y sólo se consigan corregir algunos de los atributos mal especificados, estas pequeñas mejoras pueden producir una mayor ganancia en términos de clasificación.

Otro resultado que se observa de forma consistente en el gráfico es que, a medida que mejoran las condiciones, los resultados son mejores con *EPS* más elevados. Sin embargo, si las condiciones no son favorables (p.ej., tamaño muestral reducido), los *EPS* más bajos tenderán a proporcionar mejores soluciones.

En este punto es importante destacar que, en los gráficos relativos a la TPR y la TNR, siempre y cuando las condiciones no fueran muy desfavorables (i.e., baja discriminación y test corto), los *EPS* bajos (0,60–0,70) mostraban de forma consistente unos resultados relativamente aceptables. Esto puede llevar a pensar que su uso está justificado de forma generalizada, puesto que muestran unos resultados que, sin ser óptimos, son adecuados. Sin embargo, bajo esas mismas condiciones, si se observa la medida de la $PACC_D$, puede verse que estos *EPS* realmente producen malos resultados.

En definitiva, no hay ningún *EPS* que funcione bien de forma consistente, por lo que es imprescindible tener en cuenta las condiciones del test y del tamaño muestral.

La Figura 10 muestra cómo varía la RACC en función de los distintos factores y los distintos *EPS*. De forma similar a la $PACC_D$, la RACC muestra que en condiciones favorables (e.g., discriminación alta, test largo, muestra grande) son los *EPS* más elevados los que generan una ganancia relativa en fiabilidad mayor, y viceversa: en condiciones desfavorables (e.g., discriminación baja, test corto, muestra pequeña), son los *EPS* bajos los que muestran mejores resultados, mientras que los *EPS* elevados pueden llegar a generar pérdidas de fiabilidad. Se puede ver también que, en general, a medida que se reduce la longitud del test, la RACC tiende a mostrar resultados más próximos a cero; es decir, la ganancia en fiabilidad debida a la matriz-Q corregida ($PACC_D$) es bastante menor que la máxima posible ($PACC_{D\text{-}MAX}$).

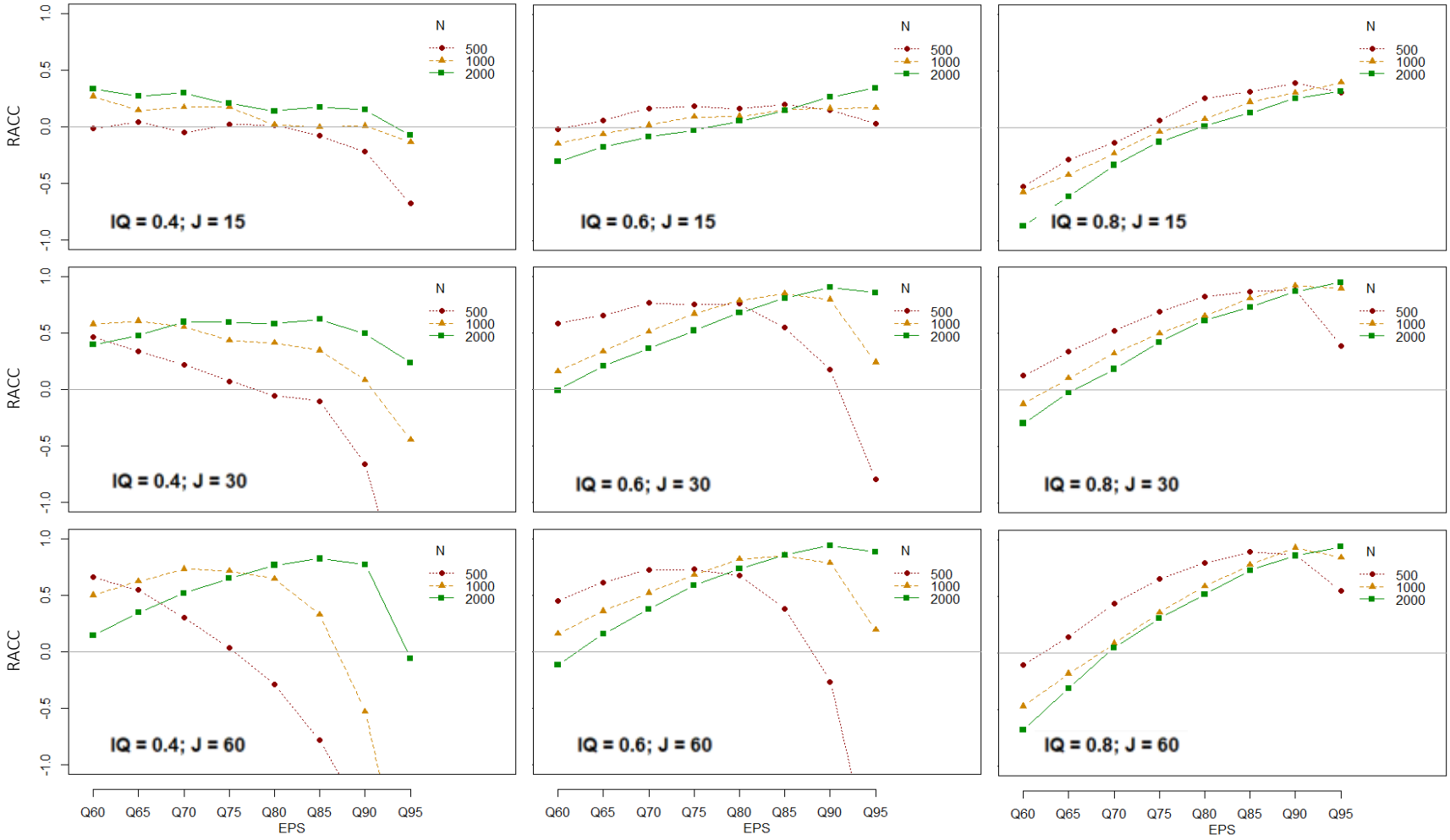
Predicción del punto de corte óptimo. En la regresión lineal múltiple las tres variables independientes fueron introducidas en el modelo, y sus coeficientes fueron significativos ($p < ,001$). La variable más relevante del modelo, tanto por su coeficiente de regresión tipificado como por su contribución al ajuste global del modelo, fue la discriminación de los ítems, seguida por el tamaño muestral y, por último, la longitud del test. La Tabla 3 muestra estos resultados.

TABLA 3.
Principales resultados de la regresión lineal del *logit* del *EPS* óptimo sobre *N*, *J* e *IQ*

	Coeficientes no estandarizados	Coeficientes estandarizados	Sig.	Cambio en R^2 al ser introducidos en el modelo
Constante	– 0,405	–	0,000	–
IQ	2,867	0,568	0,000	0,324
N	$4{,}840 \cdot 10^{-4}$	0,366	0,000	0,134
J	$-3{,}316 \cdot 10^{-3}$	– 0,075	0,000	0,006

En primer lugar, se examinó el cumplimiento de los supuestos de linealidad, no colinealidad, normalidad y homocedasticidad (la independencia de las observaciones estaba garantizada por el diseño). En el diagrama de dispersión de los pronósticos y los residuos se observó que la nube de puntos no estaba distribuida de forma totalmente homogénea, lo que indica falta de homocedasticidad (Pardo y San Martín, 2010). Aunque la falta de homocedasticidad conlleva una menor eficiencia de los coeficientes de

FIGURA 10.
RACC en función del *EPS*, *N*, *J* e *IQ*



regresión (Pardo y San Martín, 2010), no es esperable que la ligera heterocedasticidad observada tenga repercusión en las conclusiones, puesto que todos los coeficientes fueron significativos, con errores típicos reducidos. No se dieron casos influyentes: la distancia de Cook más alta fue de 0,007 (< 1), y el valor de influencia más alto fue de 0,002 ($< 0{,}50$) (Pardo y San Martín, 2010). La R_c^2 del modelo fue de 0,462.

Discusión, conclusiones y recomendaciones

La matriz-Q es uno de los inputs esenciales de cualquier CDM. Mediante ella se establece la relación entre los ítems y cada uno de los atributos a medir, definiendo así la naturaleza confirmatoria de estos modelos. Una adecuada correspondencia ítems-atributos, es decir, una correcta especificación de la matriz-Q, es indispensable para garantizar la mejor clasificación posible de las personas en función de los atributos que poseen. Errores de especificación en la matriz-Q pueden producir errores de clasificación, con las graves consecuencias que ello puede suponer. Por estas razones, en los últimos años se ha hecho patente la necesidad de añadir un paso posterior a la especificación (subjetiva) de la matriz-Q: su validación.

De la Torre y Chiu (2016) desarrollaron un método de validación particularmente ventajoso. Sus principales ventajas consisten en su flexibilidad para ser implantado en una multitud de CDMs, tanto reducidos como generales; su doble capacidad para detectar y corregir los vectores-q mal especificados; y su alta accesibilidad debido al paquete G-DINA de R, con un bajo coste computacional. Sin embargo, en su estudio, los autores exploraron su funcionamiento bajo unas condiciones particularmente limitadas y favorables. Además, el uso de un punto de corte arbitrario (*EPS*) plantea una serie de dudas. El objetivo del presente trabajo era, por tanto, estudiar el funcionamiento de este método de validación bajo un abanico más grande de condiciones de forma que los resultados fueran más generalizables, y estudiar el comportamiento de distintos *EPS* a lo largo de dichas condiciones con el objetivo de que este punto de corte se escoja en función de criterios empíricos, y no de forma arbitraria.

En el Estudio 1 se examinó el funcionamiento del método de validación asumiendo la matriz-Q como verdadera. En primer lugar, quedó patente que el uso de un único *EPS* de forma indiscriminada no está justificado. La TPR sufrió grandes variaciones para los distintos *EPS* a través de las diferentes condiciones examinadas. Los tres factores estudiados demostraron tener un efecto relevante, tanto de forma separada como en su

interacción. De esta forma, el *EPS* óptimo (el que mostraba una mayor TPR) se incrementaba a medida que se reducía la longitud del test y aumentaba el tamaño muestral y la discriminación de los ítems. A nivel general, el método mostró, para cada cruce de condiciones, al menos un *EPS* en el que la TPR alcanzó valores adecuados. Esto refleja un buen funcionamiento del método bajo el modelo correcto.

En el Estudio 2 se introdujeron un 10% de errores en la matriz-Q para estudiar el funcionamiento del método en condiciones más realistas. A la hora de detectar entradas-q incorrectamente especificadas, sólo la discriminación de los ítems y el tamaño muestral resultaron tener un efecto relevante, aumentando la TNR a medida que lo hacían dichos factores.

Tomando en consideración la TPR y la TNR de forma conjunta, el *EPS* óptimo adoptó valores más altos a medida que aumentaron los tres factores estudiados. Además, se dio la tendencia de que, a mayor *EPS* óptimo, mejores resultados globales. Esto tiene sentido en la medida en, al aumentar el tamaño muestral, la discriminación de los ítems y la longitud del test, aumenta la información disponible, ayudando al método de validación a llevar a cabo una especificación más correcta de la matriz-Q.

De la Torre y Chiu (2016) obtuvieron en sus resultados una TPR de 0,980 y una TNR de 0,804, empleando como condiciones un tamaño muestral alto (2000), una longitud del test media (30) y una discriminación de los ítems media (0,6). El *EPS* que usaron fue de 0,95. Bajo estas mismas condiciones, la TPR obtenida en el Estudio 1 del presente artículo (no hay TNR por la inexistencia de errores en la matriz-Q) fue de 0,975; en el Estudio 2 la TPR fue de 0,963 y la TNR de 0,902. Los resultados son muy similares, si bien el Estudio 2 muestra ligeras diferencias con respecto al trabajo original. Esta diferencia puede deberse al mayor porcentaje de errores de especificación empleado (10% vs. 5%), que favoreció la detección de más entradas-q mal especificadas, al tiempo que empeoró ligeramente la capacidad del método de no modificar las entradas-q bien especificadas. Cabe destacar que, si bien la TPR máxima del Estudio 1 y la TNR máxima del Estudio 2 se consiguieron con un *EPS* de 0,95, la TPR máxima del Estudio 2, que resultó ser 0,969, se consiguió con un *EPS* de 0,90.

Por su parte, el estudio de la fiabilidad volvió a revelar la gran influencia que tienen los diferentes factores estudiados en los resultados. Es particularmente interesante el hallazgo de que, si bien con condiciones desfavorables el método tiene un peor funcionamiento especificando los atributos en la matriz-Q (menor TPR y TNR), luego presenta unas mayores ganancias en términos de clasificación de atributos (mayor

$PACC_D$). La clave de esto puede estar en la cantidad de información disponible. Las condiciones favorables son las que aportan una mayor información, y una mayor cantidad de información disponible puede ayudar a mitigar el efecto de los (pocos) errores que haya en la matriz-Q. Esto hará que las ganancias en especificación de atributos no tengan una gran repercusión en términos de mejora en la clasificación (se estará dando un efecto techo). Por el contrario, la fiabilidad resultante bajo condiciones desfavorables dependerá en mayor medida de cómo esté especificada la matriz-Q; al no disponer de tanta información por parte del resto de condiciones, la matriz-Q adquiere una influencia mucho mayor. Por supuesto, esto también estará influenciado por el porcentaje de errores que tenga la matriz-Q. En la Figura 9 puede verse que la $PACC_{D\text{-}MAX}$ (líneas discontinuas), la máxima ganancia esperable, es muy baja, alcanzando como máximo valores en torno a 0,04. Es esperable que tasas de error más altas den lugar a efectos mayores. Aun así, es muy importante destacar que, para cada una de las condiciones, hay al menos un *EPS* con el que se produce una ganancia en la clasificación.

Por tanto, los resultados globales muestran que el método puede usarse bajo cada una de las condiciones estudiadas. Cuando las condiciones son muy favorables, el método detectará y modificará correctamente los errores de la matriz-Q, mientras que dejará sin modificar las entradas-q que estén correctamente especificadas; sin embargo, la mejora que esto producirá en la clasificación de los atributos no será dramática. Por el contrario, cuando las condiciones no son favorables, el método tenderá a modificar erróneamente algunas entradas-q que estaban correctamente especificadas, y detectará y modificará sólo unas pocas entradas-q mal especificadas; sin embargo, las pequeñas mejoras en el cómputo global marcarán una diferencia en cuanto a la clasificación de los atributos, justificando el uso del método de validación en estos casos. Una posible excepción a esto puede darse cuando el tamaño muestral es reducido. Las ganancias que se pueden conseguir con esta condición, tanto en términos de especificación de atributos en la matriz-Q como de clasificación de atributos, no son muy elevadas, independientemente de la discriminación y la longitud del test; sin embargo, sí se pueden empeorar mucho los resultados a nivel global si la elección del *EPS* es inadecuada. Por tanto, una recomendación general es intentar que el tamaño muestral sea lo más grande posible, al menos de 1000 participantes.

Por otro lado, para que los resultados sean favorables habrá que escoger un *EPS* determinado en función de las condiciones, puesto que si se emplea un *EPS* de 0,95 (el establecido por de la Torre y Chiu, 2016) de forma indiscriminada, se pueden llegar a

cometer errores muy graves, generando una clasificación deficiente. Las recomendaciones de elección del *EPS* en función de las diferentes condiciones se muestran en la Tabla 4. Los valores de la columna *Predicción EPS* son los resultantes de aplicar la función inversa del *logit* (*expit*) al resultado obtenido en la regresión lineal múltiple del *logit* del *EPS* óptimo.

TABLA 4.
EPS recomendado en función de *N*, *IQ* y *J*

N	*IQ*	*J*	*EPS* recomendado	Predicción *EPS*
2000	0,8 0,6	–	0,90 – 0,95	0,89 – 0,94
	0,4	60 30	0,85	0,82 – 0,83
		15	0,60 – 0,70	0,84
1000	0,8	–	0,90 – 0,95	0,90 – 0,91
	0,6	–	0,85	0,83 – 0,85
	0,4	60	0,70	0,74
		30 15	0,60	0,76
500	0,8	–	0,85 – 0,90	0,87 – 0,89
	0,6	–	0,70 – 0,80	0,79 – 0,82
	0,4	–	0,60	0,69 – 0,72

Nota: Las líneas en las casillas del factor *J* representan que el *EPS* óptimo, bajo las condiciones especificadas de *N* e *IQ*, es independiente de la longitud del test.

En esta línea, algunos autores han propuesto fórmulas para establecer el *EPS* óptimo en función de ciertos factores. Liu (citado en de la Torre y Chiu, 2017) propone la siguiente:

$$EPS = 1 - (\log N)^{-1}$$

De la Torre y Chiu (2017) critican que esta fórmula dará lugar a *EPS* óptimos entre 0,81 y 0,87 cuando se usen tamaños muestrales habituales (entre 200 y 2000 participantes), los cuales son valores restringidos, sabiendo que con muestras altas los *EPS* óptimos tienden a estar en torno a 0,95. Además, teniendo en cuenta los resultados del presente estudio, a la hora de predecir el *EPS* óptimo es un error ignorar otros factores que tienen una gran influencia, como la longitud del test o la discriminación de los ítems. En este sentido, el modelo de regresión lineal que se contrasta en este trabajo, que explica casi un 50% de la varianza, parece más adecuado. Las predicciones del modelo fueron

satisfactorias bajo las diferentes condiciones examinadas, si bien tiende a sobrestimar el *EPS* cuando la discriminación es baja. Por tanto, la recomendación general es la de usar el *EPS recomendado* (Tabla 4) cuando las condiciones del estudio sean similares a las examinadas en este trabajo; cuando no sea así, se puede emplear la fórmula de la regresión lineal (Tabla 3) para determinar un *EPS* que sirva como guía, teniendo en cuenta que, si la discriminación de los ítems es baja, probablemente esté sobrestimado.

Con todo esto, y contestando a los objetivos de investigación presentados en la Introducción: 1) el método de validación de la matriz-Q basado en el índice de discriminación propuesto por de la Torre y Chiu (2016) presenta unos resultados adecuados bajo las condiciones estudiadas, mejorando tanto el número de atributos bien especificados en la matriz-Q como la clasificación de los perfiles de atributos; 2) esto es cierto siempre y cuando se emplee un *EPS* ajustado a las condiciones particulares del estudio; es decir, el uso de un único *EPS* de forma indiscriminada no está justificado, y puede llevar a cometer errores en la clasificación de los atributos, con graves consecuencias.

Este estudio presenta una serie de limitaciones. En primer lugar, sólo se ha empleado un CDM: el modelo G-DINA. Aunque se escogió por ser un modelo general que engloba a la mayoría de los modelos reducidos, su uso no siempre está justificado, y es preferible usar, siempre que sea posible, un modelo reducido por su mayor eficiencia, facilidad de convergencia, menor coste computacional, menor tamaño muestral requerido y mayor parsimonia (Ma, Iaconangelo y de la Torre, 2016; Rojas, de la Torre y Olea, 2012; Sorrel, Abad, Olea, de la Torre y Barrada, 2017). Por tanto, aunque los resultados encontrados bajo el modelo G-DINA generan una panorámica global del funcionamiento del método de validación, sería adecuado comprobar las idiosincrasias con otros modelos reducidos que en la práctica pueden llegar a ser más usados debido a sus facilidades técnicas.

Por otro lado, aunque se ha profundizado en el estudio del funcionamiento del método de validación con la inclusión de nuevos factores relevantes, quedan otros tantos que pueden ser muy influyentes (p.ej., la estructura de la matriz-Q, el número de atributos especificados en los vectores-q, porcentaje de errores de especificación en la matriz-Q, etc.). Especialmente relevante puede ser el estudio del efecto de la distribución de las clases latentes. Además, en la regresión lineal no se han tenido en cuenta posibles relaciones no lineales, y sólo se han tenido en cuenta los efectos principales de las variables, con el objetivo de que el modelo resultante fuese sencillo de interpretar.

Por último, como continuación a este estudio, sería interesante comparar el método de validación aquí examinado con otros métodos más nuevos y menos conocidos. El método de Chen (2017) es un buen candidato para esta comparación, debido al buen funcionamiento que presenta y a que las limitaciones del estudio son similares a las del trabajo de de la Torre y Chiu (2016).

Para terminar el trabajo, es importante destacar que incrementar el conocimiento en este ámbito y conseguir unos métodos que validen lo mejor posible los atributos definidos en la matriz-Q contribuirán a mejorar el rendimiento global de los CDMs, consiguiendo una clasificación precisa de los atributos de los evaluados. Esto favorecerá que esta clase de modelos empiecen a ser usados con más frecuencia en el ámbito aplicado. De este modo, y con las garantías de una buena clasificación, los CDMs podrán contribuir a la mejora de los métodos de selección de personal, la evaluación de los estudiantes y la clasificación diagnóstica de pacientes con dolencias físicas o enfermedades mentales. En definitiva, podrán ayudar a crear un mejor y más justo entorno laboral, educativo y sanitario.

Referencias

Chen, J. (2017). A residual-based approach to validate Q-Matrix specifications. *Applied Psychological Measurement, 41*(4), 277–293.

Chen, J., y de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419–437.

Chen, J., de la Torre, J., y Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140.

Chiu, C.-Y. (2013). Statistical refinement of the Q-Matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2ª ed.). Hillsdale, NJ: Erlbaum.

de la Torre, J. (2008). An empirically based method of Q-Matrix validation for the DINA model: Development and Applications. *Journal of Educational Measurement, 45*(4), 343–362.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., y Chiu, C.-Y. (2016). A general method of empirical Q-Matrix validation. *Psychometrika, 81*(2), 253–273.

de la Torre, J., y Chiu, C.-Y. (2017). On the consistency of Q-Matrix estimation: a rejoinder. *Psychometrika, 82*(2), 528–529.

de la Torre, J., y Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

de la Torre, J., y Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*, 89–97.

de la Torre, J., y Sorrel, M. A. (2017). *Cognitive diagnosis modeling: a general framework approach and its implemmentation in R*. Presentación realizada en la Universidad Autónoma de Madrid los días 7 y 8 de junio.

de la Torre, J., van der Ark, L. A. y Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, 1–16. Publicación en línea. doi: 10.1177/0748175615569110

Gao, M., Miller, M. D., y Liu, R. (2017). The impact of Q-Matrix misspecification and model misuse on classification accuracy in the generalized DINA model. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 391–403.

García, P., Olea, J., y de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema, 26*(3), 372–377.

Haertel, E. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8*(3), 333–346.

Hartz, S., y Roussos, L. (2008). The fusion model for skills diagnosis: blending theory with practicality. *Educational Testing Service, Research Report, RR-08-71*.

Henson, R., Templin, J., y Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210.

Jaeger, J., Tatsuoka, C., Berns, S., y Varadi, F. (2006). Distinguishing neurocognitive functions in schizofrenia using partially ordered classification models. *Schizofrenia Bulletin, 32*(4), 679–691.

Junker, B., y Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric IRT. *Applied Psychological Measurement, 25*, 258–272.

Li, H., y Suen, H. K. (2013). Constructing and validating a Q-Matrix for cognitive diagnosis analyses of a reading test. *Educational Assessment, 18*, 1–25.

Ma, W., y de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69,* 253–275.

Ma, W., y de la Torre, J. (2017). GDINA: The generalized DINA model framework. R Package Version 1.4.2. Tomado de https://cran.r-project.org/package=GDINA

Ma, W., Iaconangelo, C., y de la Torre, J. (2016). Model similarity, model selection and attribute classification. *Applied Psychological Measurement, 40*(3), 200–217.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187–212.

Pardo, A., y San Martín, R. (2010). *Análisis de datos en ciencias sociales y de la salud II*. Madrid, España: Editorial Síntesis.

R Core Team (2016). R (Version 3.3) [Computer Software]. Viena, Austria: R Foundation for Statistical Computing.

Romero, S., Ordóñez, X., Ponsoda, V., y Revuelta, J. (2014). Detection of Q-Matrix misspecifications using two criteria for validation of cognitive diagnosis structures under the least squares distance model. *Psicológica, 35*, 149–169.

Rojas, G., de la Torre, J., y Olea, J. (2012, abril). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Trabajo presentado en el congreso del National Council of Measurement in Education, Vancouver, Canadá.

Rupp, A., y Templin, J. (2008). The effects of Q-Matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78–96.

Sorrel, M. A., Abad, F., Olea, J., de la Torre, J., y Barrada, J.R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement, 41*(8), 614–631.

Sorrel, M. A., de la Torre, J., Abad, F. J., y Olea, J. (2017). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology: European Jorunal of Research Methods for the Behavioral and Social Sciences*, *13*(S1), 39.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., y Lievens, F. (2016). Validity and reliability of situational judgment test scores: a new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*(3), 506–532.

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconception based on item response theory. *Journal of Education Statistic, 20*, 345–354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. En N. Frederiksen, R. Glaser, A. Lesgold y M. Shafto (Editores). *Diagnostic Monitoring of Skill and Knowledge Acquisition*, 453–488. Hillsdale, NJ: Erlbaum.

Templin, J., y Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *Educational Testing Service, Research Report, RR-05-16*.

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., y Jie, Y. (2018). An EM-based method for Q-Matrix validation. *Applied Psychological Measurement*, 1–14.