REIRE
Revista d'Innovació i Recerca en Educació
www.ub.edu/ice/reire.htm

<Artículo>

¿Pueden las escalas Likert aumentar en sensibilidad?

Rafael Bisquerra y Núria Pérez-Escoda

Fecha de presentación: 03/03/2015
Fecha de publicación: 07/07/2015

//Resumen
En las investigaciones que utilizan escalas tipo Likert se aplican principalmente escalas de 5 puntos, sin una fundamentación metodológica que lo justifique. La revisión de 3 conocidas revistas permite llegar a esta conclusión. Parece ser que se hace así por tradición y porque es difícil poner nombre a más de 5 opciones de respuesta. En este artículo cuestiona esta tradición y se aportan argumentos para proponer otras alternativas. Se analiza la importancia de mejorar la sensibilidad de las escalas aumentando las opciones de respuesta; se recomienda evitar el uso de denominaciones categóricas porque stricto sensu impide su uso como escala de intervalo, ya que la convierte en nominal (categórica); se analiza el rechazo a valores extremos, etc. Como consecuencia, se recomienda la propuesta a favor de escalas de 11 puntos (de 0 a 10).

// Palabras clave
Escalas de medición, confiabilidad, instrumentos de medición, encuestas, evaluación.

// Referencia recomendada
Bisquerra, R. y Pérez-Escoda, N. (2015). ¿Pueden las escalas Likert aumentar en sensibilidad? *REIRE, Revista d'Innovació i Recerca en Educació*, 8 (2), 129-147. DOI: 10.1344/reire2015.8.2828

// Datos de los autores
Rafael Bisquerra. Departamento de Métodos de Investigación y Diagnóstico en Educación
Universitat de Barcelona, España, rbisquerra@ub.edu

Núria Pérez-Escoda. Departamento de Métodos de Investigación y Diagnóstico en Educación
Universitat de Barcelona, España, nperezescoda@ub.edu

1. Introducción

Pasados ya más de 80 años desde la publicación del artículo que dio lugar a las denominadas escalas tipo Likert (1932), es una buena ocasión para revisar cómo se están utilizando en la actualidad. Es fácil observar que la mayoría de revistas del campo de la psicología y educación suelen incluir artículos con estudios basados en escalas tipo Likert, lo cual es un indicador de su vigencia. El uso actual de este tipo de escalas plantea una serie de cuestiones que se presentan en este artículo con el objetivo de contribuir a mejorar su práctica.

Conviene distinguir entre *escala* y *elemento*. La *escala* es la suma de las respuestas de los elementos del cuestionario. Los *elementos* son los ítems o afirmaciones que suelen ir acompañados por las opciones de respuesta. Un elemento de tipo Likert es una afirmación a la que hay que responder, generalmente indicando el nivel de acuerdo o desacuerdo. Una escala tipo Likert puede tener, por ejemplo, 30 ítems o elementos con varias opciones de respuesta asociadas a cada uno de ellos. A veces, un elemento se puede representar mediante una línea horizontal en la que se marca la respuesta. En resumen, *escala tipo Likert* es la suma de todos los elementos o afirmaciones.

Likert (1932) no fue el primero en obtener valoraciones subjetivas basadas en autoinformes. Previamente, Freyd (1923) discutió las diversas formas de escalas disponibles en aquellos momentos y observó que, en general, se basaban en elementos de 10 o de 100 puntos. Sin duda, el sistema centesimal es el más intuitivo y fácil de conceptualizar. Posteriormente, Ferguson (1941) también aportó evidencias de las ventajas de los elementos de 10 puntos.

Sin embargo, fue Likert el que tuvo un mayor impacto social con sus escalas, que en general tendían a ser de elementos de 5 puntos. Puede ser muy instructivo considerar si la formulación original de las escalas Likert se ha mantenido por razones de carácter epistemológico, metodológico o psicométrico, o si, por el contrario, son simplemente razones prácticas o de tradición.

Cummins (1997) revisó más de 400 instrumentos basados en escalas tipo Likert. La conclusión de su estudio fue una ausencia de reglas según las cuales los investigadores diseñaban sus escalas. Observó que el número de opciones de respuesta variaba desde 2 (sí-no, acuerdo-desacuerdo) hasta 100. Algunas incluyen el punto medio (escalas impares) y otras no (pares). Lo más habitual es utilizar alrededor de 5-7 opciones.

Un argumento psicométrico para justificar escalas de 5-7 puntos es que diversos estudios observaron que la presencia de más puntos en los ítems no aumenta la fiabilidad de la prueba. Por lo tanto, si con 5-7 es suficiente, ¿por qué se tienen que utilizar más? Esto conlleva que el número de puntos no se haya considerado un aspecto importante. Lo que interesa es la validez y la fiabilidad de la escala, que puede ser prácticamente la misma, independientemente del número de opciones de respuesta. Sin embargo, investigaciones más recientes han puesto en entredicho este aspecto (Cummins y Gullone, 2000). Diversos autores han aportado evidencias de que, al aumentar el número de opciones, no disminuye la fiabilidad (Diefenbach, Weinstein y O'Reilly, 1993; Lozano, García-Cueto y Muñiz, 2008; Matell y Jacoby, 1971; Muñiz, García-Cueto y Lozano, 2005), sino que, por el contrario, puede aumentarla. En un meta-análisis de 131

estudios, Churchill y Peter (1984) hallaron una correlación positiva entre la fiabilidad y el número de puntos de los ítems de la escala. Es decir, aumentar el número de opciones no reduce la fiabilidad de la escala, sino que la aumenta.

Por otra parte, existen otras razones que no se han tomado en consideración como criterio a la hora de decidir el número de opciones de los ítems. Un aspecto importante es la sensibilidad de la prueba. A mayor amplitud de las opciones, el instrumento es más sensible para detectar cambios a lo largo del tiempo.

En conclusión, hay elementos suficientes para replantear cuál debe ser el número de puntos apropiado en los ítems de las escalas Likert. Esta opinión ha sido argumentada por otros autores (Cummins, 1997; Cumins y Gullone, 2000; Finstad, 2010; Shaftel, Nash y Gillmor, 2012), lo cual nos lleva a plantearlo en el ámbito psicoeducativo con la intención de sensibilizar a los investigadores para que se planteen el tema en el momento de tomar decisiones sobre el número de puntos de los ítems y aporten criterios que justifiquen su elección.

2. Escalas Likert más habituales

Con la intención de conocer las características de las escalas Likert utilizadas más frecuentemente, se ha realizado una revisión de los artículos publicados durante los años 2010-2012 en tres reconocidas revistas: *Revista de Investigación Educativa* (RIE), *Revista Electrónica de Investigación Psicoeducativa* (EJREP) y *Revista Electrónica de Investigación y Evaluación Educativa* (RELIEVE). La Tabla 1 presenta el número de artículos que contienen escalas Likert en la citada muestra de revistas.

Tabla 1.

Artículos con escalas Likert.

Año	RIE	JEREP	RELIEVE	Total
2010	2	28	6	36
2011	2	31	7	40
2012	9	27	4	40
Total	13	86	17	116

En la Tabla 2 hay una clasificación de las escalas Likert utilizadas en los artículos de las revistas de la muestra en función del número de puntos o categorías de los ítems. Conviene aclarar que se recogen 116 artículos que utilizan escalas Likert, pero algunos de ellos utilizan más de una escala en el mismo artículo. Por esto, en total se contabilizan 152 escalas en la Tabla 2.

Tabla 2.

Número de escalas tipo Likert, según categorías o puntos, utilizadas en los artículos analizados

	Categorías o puntos de la escala											
Escalas	2	3	4	5	6	7	8	9	10	11	20	TOTAL
	4	8	36	72	11	11	2	1	5	1	1	152

Como se puede observar, el abanico de puntos va de 2 a 20, siendo la frecuencia más alta 5 puntos, en 72 escalas. Le siguen las escalas de 4 puntos, utilizadas en 36 ocasiones. Es lógico que haya 5 escalas con 10 puntos y, en cambio, sorprende que solamente haya 1 con 11. Fijémonos que un ítem de 11 puntos posibles representa incluir el 0 y el 10. La mayoría de investigaciones utilizan entre 3 y 7 puntos en las escalas.

La Tabla 3 presenta las denominaciones de las categorías o puntos de los ítems. Son ejemplos representativos; no incluyen todos los casos. Es interesante observar que a partir de más de 5 categorías se tiende a no denominarlas, sino que se considera como un *continuum* con dos extremos a los que sí se les da nombre. Ahora bien, incluso en tal caso, la utilización de 0 está totalmente ausente, cuando expresiones como "nada", "totalmente en desacuerdo" o "muy insatisfecho" pueden encajar perfectamente en una valoración de 0.

Tabla 3.

Denominación de las categorías.

Núm. de categorías	Denominación de las categorías
2	Sí, No
3	Nunca, Algunas veces, Muy frecuentemente
4	Nada, Algo, Bastante, Mucho.
	Nunca, Apenas, Pocas veces, Bastante, Mucho
	Ninguna experiencia, Poca experiencia, Bastante experiencia, Mucha experiencia
5	Nada, Poco, Suficiente, Bastante, Mucho
	Completamente en desacuerdo, Algo en desacuerdo, Indiferente, Algo de acuerdo, Totalmente de acuerdo.
	Totalmente de acuerdo, De acuerdo, Ni de acuerdo ni en desacuerdo, En desacuerdo, Totalmente en desacuerdo.
	Siempre, A menudo, Alguna vez, Rara vez, Nunca
6	1 = Totalmente en desacuerdo - 6 = Totalmente de acuerdo
7	1 = Muy insatisfecho - 7 = Muy satisfecho
8	1 = Nada cierto – 8 = Muy cierto
9	1 = Bajo - 9 = Alto
10	1 = Nada - 10 Máxima intensidad
11	0 = muy mal - 10 = excelente

Al analizar los datos de las Tablas 2 y 3 se constata lo siguiente: a) las escalas tipo Likert son un instrumento muy utilizado en las investigaciones de psicología y educación; b) lo más habitual es utilizar escalas de 5 puntos, aunque pueden tener un amplio rango de variación que va de 2 a 20; c) la denominación de las categorías varía de una escala a otra; d) para más de 5 puntos se utiliza como una escala de intervalo (o de razón) y no se proporciona la denominación de las opciones de respuesta.

Al examinar estos trabajos se observa que los autores se limitan a expresar que utilizan un determinado número de puntos en los ítems. La mayoría de autores se limita a decir algo así como "se ha aplicado una escala tipo Likert de 5 puntos" sin hacer mención, sin embargo, a los criterios por los cuales se han elegido 5 y no 4 o 6, por ejemplo. No se exponen los argumentos por los cuales se toma una determinada decisión a este respecto. Se puede interpretar, pues, que la tradición así lo ha establecido.

Ante estos datos cabe preguntarse: ¿Cuántos puntos son más apropiados: 4, 5, 7, 9, 11? ¿Hay evidencias empíricas que permitan tomar decisiones al respecto con conocimiento de causa? ¿Podemos mejorar la precisión del instrumento variando el número de puntos? ¿Es conveniente establecer denominaciones para las diversas categorías? ¿Estas escalas, que llevan más de ochenta años utilizándose, son susceptibles de mejora? A continuación se aportan argumentos y evidencias que permitan responder mejor a estas preguntas.

3. ¿Por qué son tan populares las escalas de 5 opciones?

Después de 80 años desde la propuesta de Likert (1932) la tradición ha mantenido los ítems de 5 puntos, sin platearse si realmente es lo más conveniente. ¿Por qué la formulación original se ha mantenido de forma tan aceptada y popular a lo largo de los años? Una de las razones es porque es difícil encontrar denominaciones para más de 5 opciones. Esto queda claro en la revisión realizada ante la evidencia de que no se proporciona la denominación para más de 5 categorías.

Conviene tener en cuenta que este criterio no es de carácter epistemológico, metodológico o psicométrico, sino que responde simplemente a razones prácticas. Paralelamente, es interesante observar la inercia en seguir esta tradición, sin replantearse la conveniencia de revisarla.

Un argumento complementario es que a partir de 5 puntos, el hecho de aumentar más puntos no aumentaba prácticamente la fiabilidad. Sin embargo, investigaciones recientes han demostrado lo contrario (Cummins, 1997 y Cummins y Gullone, 2000). Este criterio, que ahora se ha visto cuestionado, ha evitado tomar en consideración la *sensibilidad* como otro criterio a tener en cuenta.

En resumen, los argumentos que justifican las escalas de 5 opciones se basan en la dificultad en dar nombre a más de 5 opciones. De esta forma, se ha seguido una tradición popular sin análisis posteriores que aporten criterios de mayor fundamento y que tomen en consideración la *sensibilidad* de la escala (Cummins, 1997).

4. Importancia de la sensibilidad de las escalas

La sensibilidad de una escala de medida (*measurement sensitivity*) es la capacidad para detectar cambios a través del tiempo. La sensibilidad de un instrumento de medida a pequeños cambios es muy importante si se quiere utilizar para analizar la evolución en periodos de tiempo, análisis de tendencias, diseños pretest-postest, etc. Se dispone de instrumentos de alta validez y fiabilidad, pero que no son sensibles a la hora de detectar los efectos de una intervención. Al aplicar las habituales pruebas de contraste (t de Student, análisis de la varianza) las diferencias observadas pueden no ser estadísticamente significativas debido a la falta de sensibilidad del instrumento, y no al hecho de que no haya habido cambios reales. En educación, más allá del diagnóstico puntual, interesa evaluar cambios como consecuencia de procesos educativos o de intervención.

La sensibilidad de los instrumentos es particularmente importante en las ciencias de la salud, donde ha recibido especial atención (Guyatt y Jaeschke, 1990). Pero es realmente curioso que este parámetro crucial de la medición haya sido prácticamente ignorado sistemáticamente en psicología y educación. En la medición psicométrica y educativa se ha insistido mucho en la importancia de la validez y de la fiabilidad. En cambio, sorprende que prácticamente no se haya tratado la sensibilidad. A menudo no se ha previsto este aspecto y, como consecuencia, cuando posteriormente interesaría poder analizar la evolución de la variable objeto de estudio, el instrumento no resulta suficientemente sensible. Un análisis de los artículos de las tres revistas objeto de análisis de este estudio permite concluir que este es un tema prácticamente ignorado. Solamente en una revisión más amplia sobre la temática se ha encontrado un artículo en cuyo título aparece el término "sensibilidad", referido a los instrumentos de medida; se trata del artículo publicado por Fuentes *et al.* (2003).

Diefenbach *et al.* (1993) hallaron que una escala Likert de 7 puntos tenía mayor sensibilidad que una de 5. Otros autores como Russell y Bobko (1992) observaron la ventaja de una escala más amplia, en concreto de 15 puntos, al utilizar estadística paramétrica y en particular la regresión. Guyatt y Jaeschke (1990) y Cummins y Gullone (2000) pusieron de manifiesto que una escala de 5 puntos no es suficientemente sensible para detectar pequeñas diferencias entre pretest y postest.

Cummins (1997) aporta evidencias de que las escalas con 5-7 opciones no explotan toda la capacidad discriminativa de la mayoría de las personas en términos de su percepción sobre el fenómeno que se está midiendo. Es decir, las personas, en general, tienen una capacidad discriminativa superior de lo que permiten las escalas de 5 puntos, lo cual supone despreciar una potencialidad que podría utilizarse en estos instrumentos de medida. En este mismo sentido, Alwin (1997) afirma que una escala que presente muchas categorías permite refinar mucho la medida, ya que el encuestado es capaz de marcar aquella opción que más se ajusta a su situación de una manera más precisa. Dado que la gente tiene una capacidad discriminativa que va más allá de los 5 puntos, restringir las respuestas a estas opciones comporta una pérdida de datos potencialmente más discriminativos. Este autor concluye que este tipo de escalas representan instrumentos muy rudos que no permiten medir cambios a través del tiempo. Es decir, las escalas no tienen la sensibilidad que podrían tener ampliando las opciones de respuesta a cada ítem.

Aumentar el número de opciones de una escala aumenta su sensibilidad. Supongamos una escala con 5 opciones: Nada, Poco, Suficiente, Bastante, Mucho. Para que una persona cambie de una categoría a otra (por ejemplo, de "Poco" a "Suficiente") entre un prestest y un postest tiene que darse un cambio muy importante, que se podría cifrar en un 20% si realmente fuese una escala de razón. En cambio, si las opciones fuesen de 0 (Nada) a 10 (Mucho), es más fácil que un sujeto cambie de un dígito a otro (por ejemplo de 6 a 7). En este caso es suficiente un 9% de cambio para ser detectado. Es tan evidente que al aumentar el rango de opciones de respuesta aumenta la sensibilidad del instrumento que no haría falta tener que insistir. Entonces, ¿cuáles son los impedimentos para utilizar escalas más amplias? Uno de ellos es la dificultad de encontrar denominaciones distintas para más de 5 opciones, tal como se ha señalado previamente. Este aspecto se comenta en el apartado siguiente.

5. El uso de denominaciones categóricas

Desde las primeras escalas de Likert hay una tradición en denominar a las categorías, pero se ha observado que para más de 5 opciones hay una dificultad en encontrar expresiones adecuadas para denominarlas. No es fácil encontrar matices, lo cual no significa que una persona no pueda percibir matices con más detalle, aunque no sepa como denominarlos verbalmente.

Cummins (1997) argumenta y aporta evidencias de que usar escalas con denominación verbal para cada categoría significa pasar de una variable continua a una variable categórica (nominal). Además, en este tipo de escalas nominales el sujeto a veces no halla exactamente la opción a su respuesta y se ve obligado a responder dentro de las opciones que se le presentan. De esta forma se puede encontrar que la opción central de "No sé" o "Indiferente" no le convence y no encuentra ninguna opción que verbalmente se ajuste a su percepción. Cummins (1997) aporta, asimismo, argumentos para demostrar que las denominaciones pueden confundir la respuesta. Por ejemplo, lo que para unos es "frecuentemente" para otros es "algunas veces" y viceversa. Por todo ello concluye que utilizar una denominación verbal para las opciones de respuesta es una desventaja.

La teoría psicométrica de las escalas Likert asume que son escalas de intervalo o de razón. Por lo tanto, debería haber una cierta equidistancia entre las opciones de respuesta, lo cual es un requisito para el uso de estadística paramétrica. Este principio queda violado cuando se denominan las categorías. No se puede afirmar que se dé la misma distancia, por ejemplo, entre "Totalmente en desacuerdo" y "En desacuerdo", que entre esta última y "Ni de acuerdo ni en desacuerdo".

En una investigación de la Roy Morgan Research (1993) se presenta una escala de 9 opciones: Encantado, Muy satisfecho, Satisfecho, Bastante satisfecho, Sentimientos ambiguos, Más bien insatisfecho, Infeliz, Muy infeliz, Terrible (*Delighted, Very pleased, Pleased, Mostly satisfied, Mixed feelings, Mostly dissatisfied, Unhappy, Very unhappy, Terrible*). Este trabajo asume que la distancia entre las diversas opciones de la escala es siempre la misma. Esto es muy difícil de argumentar, ya que no hay evidencias que puedan apoyar esta argumentación (Cummins y Gullone, 2000).

Las personas tienen una interpretación distinta del valor que representan las denominaciones categoriales: "A menudo" se puede interpretar de formas muy distintas; para algunos puede ser "Una vez a la semana", mientras que para otros es "Cada día". Esto significa que la denominación puede confundir el valor de la respuesta. Por ejemplo, ¿es lo mismo "En desacuerdo" que "Algo en desacuerdo"? Solomon y Kopelman (1984) apoyan la idea de que las denominaciones pueden confundir al sujeto que tiene que responder. Cummins (1997) aporta evidencias de la falta de equidistancia entre las opciones de las diferentes respuestas. Se basa en ejemplos como "Rara vez", "A menudo", "Ocasionalmente", "Usualmente", "A veces", etc., para demostrar la falta de equidistancia entre las opciones.

Autores como Matell y Jacoby (1971), Wyatt y Meyers (1987), Dixon, Bobo y Stevick, (1984), Cummins (1997) y otros, han aportado evidencias de las ventajas de no utilizar denominaciones, sino valores numéricos entre dos extremos (Todo-Nada, Satisfecho-Insatisfecho, Siempre-Nunca).

En la Tabla 3 se puede observar que para más de 6 categorías las escalas se limitan a poner las denominaciones de los dos extremos, lo cual significa estar de acuerdo con los argumentos que aquí se aportan.

Cañadas y Sánchez Bruno (1998) analizan la siguiente gradación de denominaciones: *Siempre*, Muchísimas veces, *Casi siempre*, Muchas veces, Muy a menudo, Generalmente, Con frecuencia, A menudo, *Normalmente*, Ordinariamente, *A veces*, Algunas veces, Ocasionalmente, De vez en cuando, Alguna vez, Raras veces, Raramente, Muy raramente, *Casi nunca*, Nunca. De este conjunto proponen las categorías que aparecen en cursiva para una escala de 5 puntos. Ahora bien, si nos fijamos en los matices entre todas las denominaciones, se comprueba que la confusión es relativamente fácil.

El ejemplo de los colores puede ilustrar lo que pretendemos argumentar. Las personas normales conocen el nombre de un número reducido de colores, pero pueden distinguir entre multitud de matices aunque no sepan el nombre. Pongamos por caso alguno de los matices para el color amarillo: amarillo limón, amarillo primavera, amarillo sólido, amarillo oscuro, amarillo albaricoque, amarillo capuchino, etc. Probablemente muchas personas desconocen los nombres de estos matices y tendrían dificultad en ordenarlos por intensidad solamente con el nombre. En cambio, nadie dudaría en ordenar de mayor a menor, numerosos matices del amarillo a simple vista, sin ninguna denominación categórica. Lo que queremos demostrar es que las personas poseen una capacidad discriminativa muy superior a lo que permite el lenguaje. De hecho, el lenguaje puede confundir la capacidad discriminativa, de lo cual se deriva la conveniencia de suprimir muchas veces la denominación categórica, si bien se puede mantener en los dos extremos.

En el estudio desarrollado por Darbyshire y McDonald (2004) se utilizaron en la misma encuesta dos escalas paralelas: una numérica y otra semántica. La escala numérica constaba de 9 categorías numéricas (del 1 al 9), a cuyos extremos se añadían dos valores semánticos: "Muy mal" y "Muy bien". La escala semántica se componía de 5 categorías ordinales: "Muy mal", "Mal", "Indiferente", "Bien" y "Muy bien". La comparación de los datos obtenidos con estas dos escalas les sirvió para llegar a la conclusión de que en la escala semántica los encuestados psicológicamente interpretaban que los significados de cada categoría no eran expresables en

términos de intervalos. La distancia entre los significados "Muy mal" y "Mal" es mucho más alta que la distancia existente entre los dos extremos del primer intervalo de la escala numérica. En conclusión, hay razones para desaconsejar el uso de las denominaciones en las opciones de respuesta cuando se habla de variables latentes continuas. El uso de denominaciones en las categorías suprime la naturaleza de escala de intervalo y la convierte en una escala cualitativa nominal (categórica). También dificulta aumentar el número de opciones de respuesta. Estas argumentaciones no han sido suficientemente difundidas y merece la pena que lo sean para que los investigadores que utilicen escalas tipo Likert lo tengan claramente presente y lo puedan aplicar en su práctica investigadora.

6. Rechazo de valores extremos

A los argumentos anteriores conviene señalar que las personas tienden a rechazar los valores extremos al responder. Consideremos, por ejemplo, una escala de 5 puntos: Siempre, A menudo, Alguna vez, Rara vez, Nunca. Lo más probable es que la mayoría de personas rechace responder a los dos valores extremos (Siempre, Nunca). Es lógico y evidente un cierto rechazo a responder los valores extremos (Totalmente de acuerdo, Totalmente en desacuerdo, Nunca, Nada, Todo, Siempre), ya que siempre hay excepciones que hacen que casi nunca se puedan dar respuestas extremas absolutas (Siempre, Todo, Nada, Nunca). En muchas ocasiones, los valores extremos no son escogidos por nadie o por muy pocas personas. Cañadas y Sánchez Bruno (1998) aportan evidencias de que la opción "Nunca" es elegida muy pocas veces, lo cual significa que en la práctica, una escala de 5 valores se puede convertir en una de 3 (En desacuerdo, Ni de acuerdo ni en desacuerdo, De acuerdo). Este hecho limita las opciones de cambio en medidas sucesivas a través del tiempo. Así, para pasar de una categoría a otra, se debería producir un cambio de aproximadamente el 30 %, y esto es muy difícil.

La tendencia a responder los valores centrales es una razón por la cual, a veces, se proponen escalas pares para forzar que el sujeto se incline por uno de los dos lados. Esto supone otro punto de discusión permanente que también puede inducir a confusión, tal como se comenta en el apartado siguiente.

7. La alternativa "indiferente"

Es motivo de controversia si debe haber un número par o impar de opciones de respuesta. En el caso de que sean pares, la persona encuestada se ve obligada a inclinarse hacia un lado u otro. En el caso de que haya un número impar, se presenta una categoría intermedia.

Hay opiniones para defender ambas posturas. Hernández, Espejo, González y Gómez (2001) aportan evidencias de que la alternativa "Indiferente" es muy poco elegida, hasta tal punto que se podría eliminar. Ahora bien, se puede interpretar que ser poco elegida, no siempre significa que la persona que responde no se pueda sentir identificada con un valor central, sino que la

denominación verbal no se ajusta exactamente a lo que piensa. Esto queda bastante claro con las denominaciones "Indiferente" o "No lo sé". Una persona puede pensar que "sí lo sabe" o que "no le es indiferente"; que se inclina por valores medios (centrales) pero que, sin embargo, no se identifica con la denominación. Presentar la escala solamente con las denominaciones de los dos extremos elimina la dificultad a la hora de denominar los valores centrales y facilita el posicionamiento de la respuesta.

8. Se puntúa más alto con pocas alternativas de respuesta

Cummins (1997) y Dawes (2008) observaron que al responder una escala de 5 o 7 puntos o alternativas de respuesta se tiende a puntuar más alto de lo que se hace en una escala de 0 a 10. Cummins (1997) señala que la mayoría de las respuestas tienden a estar en el lado positivo o alto de la escala, lo cual, con ítems de pocas opciones, puede llegar a aumentar exageradamente la percepción de la realidad. Por otra parte, esto produce una asimetría negativa, que se escapa de la normalidad, y por lo tanto contribuye a distorsionar los datos de cara a la utilización de estadística paramétrica.

9. A más categorías más sensibilidad para detectar cambios: Justificación de las escalas de 0 a 10

Como consecuencia de lo que se ha expuesto, podemos imaginar una escala de 5 puntos, cuyas respuestas pueden, en la práctica, reducirse a los 3 valores centrales. Cuando se aplica esta escala en dos momentos diferentes a los mismos sujetos (por ejemplo, en un diseño pretest-postest), es muy difícil detectar cambios, ya que las personas que han respondido, por ejemplo a la categoría 2 "En desacuerdo", tendrían que experimentar un cambio realmente espectacular para aumentar o disminuir de categoría en sus respuestas (pasar a la categoría 1 o 3). En cambio, si se dispone de una escala de 0 a 10, es más probable que una persona pueda pasar del 6 al 7, por ejemplo. Nos referimos al caso en que entre las dos medidas puede haber una acción formativa o simplemente comparar dos grupos distintos. En otras palabras, a más categorías en la escala, más sensibilidad tiene el instrumento y, por lo tanto, es más probable que detecte diferencias entre mediciones distintas.

A partir de estos argumentos, consideramos oportuno proponer el uso de escalas Likert cuyas opciones de respuesta en cada ítem puedan oscilar de 0 a 10. Ello significa 11 opciones de respuesta, ya que se incluye el 0 y el 10. Un ejemplo de elemento de 0-10 puede plantearse de la siguiente forma:

Valore de 0 a 10 las siguientes afirmaciones (0 = Ausencia total de competencia; 10 = Dominio total; el 5 representa un dominio mediano).

Cuadro1.
Ejemplo de ítem o elemento en una escala de Likert.

Sabe expresar sus emociones de forma apropiada	

Las escalas de 0 a 10 son intuitivas, ya que representan el sistema decimal al que estamos tan acostumbrados. Por otra parte, esto representa aumentar considerablemente la sensibilidad del instrumento. A lo largo de este artículo se aportan argumentos para justificar esta propuesta. Un número creciente de autores proponen el uso de escalas de 0 a 10 como, por ejemplo, Hooker y Siegler (1993); Cummins (1997); Watkins *et al.* (1998); Batista-Foguet, Saris, Boyatzis, Guillén y Serlavós (2009); etc. Una variante de ello son las escalas centesimales (0 a 100) con alta sensibilidad, que ya fueron descritas por Freyd (1923). Sin embargo, los valores tienen un rango tan grande que dejan de ser tan intuitivos y prácticos como los de cero a diez.

Para las escalas de 0 a 10 se sugiere incluir denominaciones atribuidas solamente a las puntuaciones extremas que orienten la tendencia de los valores. En este sentido, se sugiere la utilización de pares como los siguientes: "Totalmente en desacuerdo (0) / Totalmente de acuerdo (10)"; "Nunca (0) / Siempre (10)"; "Nada (0) / Todo (10)"; "Ausencia de competencia (0) / Dominio total (10)"; "Nada importante (0) / Muy importante (10) "; "En contra (0) / A favor (10) ".

A continuación se presentan ejemplos en los que se ponen de manifiesto las ventajas de utilizar escalas de 0 a 10, respecto a otras opciones más comunes, como las de 5 puntos.

10. Aplicación a casos prácticos: Comparación entre escalas de 5 puntos versus 0-10

A continuación se presentan 3 ejemplos de comparación de análisis estadísticos con elementos de 5 puntos (de 1 a 5) y de 11 puntos (de 0-10). Se comparan ítems en concreto y también toda la escala. En todos los casos se ha utilizado un instrumento para la evaluación de las competencias emocionales diseñada *ad hoc*. El propósito es aportar argumentos para tomar decisiones acerca del número de opciones más conveniente (5 o 10) a la hora de elaborar ítems de una escala que sea lo más sensible posible. Para ello la escala se ha aplicado en dos versiones, con distinto número de categorías en cada ítem. Se ha contado con una muestra de estudiantes de secundaria. En las Experiencias 1 y 3 se han utilizado grupos independientes: grupo experimental (con formación en educación emocional) y grupo control (sin formación). En la Experiencia 2 se ha utilizado el diseño intragrupo de medidas repetidas (pretest y postest).

11. Experiencia 1: Un elemento con datos independientes

Se quiere comparar dos grupos respecto a un elemento tipo Likert. La misma afirmación se ha formulado primero con una escala de 1 a 5 y, en otro lugar del cuestionario, con una escala de 0 a 10. Ambas afirmaciones se han aplicado a los dos grupos.

Para asegurar que los dos ítems miden lo mismo se ha calculado la correlación entre ambos elementos y el resultado ha sido de r = 0,965. Es decir, miden exactamente lo mismo. Al comparar los dos grupos se obtienen los siguientes estadísticos (Tabla 4).

Tabla 4.

Estadísticos de un elemento con datos independientes.

	Grupo	N	Media	DT
Escala 1-5	GC	53	2,72	1,01
	GE	55	3,04	1,03
Escala 0-10	GC	53	4,19	2,41
	GE	55	5,15	2,49

Al aplicar la prueba t de Student, los resultados son los siguientes (Tabla 5).

Tabla 5.

Comparación de las respuestas a un elemento con datos independientes en función del número de puntos del ítem.

	Prueba de Levene		Prueba T para la igualdad de medias		
	F	Sig.	t	gl	Sig.(bil)
Escala 1-5	0,166	0,685	-1,624	106	0,107
Escala 0-10	0,089	0,766	-2,025	106	0,045

Como se puede observar, con dos ítems que miden exactamente lo mismo, cuando se dispone de 5 opciones de respuesta no se observan diferencias estadísticamente significativas entre los grupos (p = 0,107). En cambio, cuando se utiliza una escala de 0 a 10 se detectan diferencias entre los dos grupos (p = 0,045). Estos resultados evidencian una mayor sensibilidad del elemento para detectar diferencias cuando se puede optar entre 0 y 10.

12. Experiencia 2: Un elemento con datos apareados

En una acción formativa se ha comparado entre la situación inicial (pretest) y la situación después de la intervención (postest). En ambos casos, se ha formulado la misma afirmación, pero en dos formatos distintos: primero con una escala de 1 a 5 y en otro lugar del cuestionario se mide de 0 a 10. Ambas afirmaciones se han aplicado en el pretest y en el postest.

Para asegurar que los dos ítems miden lo mismo se ha calculado la correlación entre ambos elementos. Los resultados aparecen en la siguiente matriz de correlaciones (Tabla 6).

Tabla 6.

Correlación entre los elementos en función de las opciones de respuesta.

		1	2	3	4
PRETEST	1. Escala 0-10a	--	,966**	,318*	,339*
	2. Escala 1-5a	,966**	--	,325*	,383**
POSTEST	3. Escala 0-10b	,318*	,325*	--	964**
	4. Escala 1-5b	,339*	,383**	,964**	--

Nota: * p < .05; ** p < .01

Como se puede observar, en el pretest se ha obtenido una r = 0,966, y en el postest r = 0,964. Por lo tanto, se puede afirmar que en ambos casos se puede afirmar que se mide lo mismo. Los estadísticos descriptivos son los siguientes.

Tabla 7.

Estadísticos de un elemento con datos apareados.

	N	Media	DT
Escala 1-5a	55	2,75	1,01
Escala 1-5b	55	3,04	1,04
Escala 0-10a	55	4,25	2,40
Escala 0-10b	55	5,15	2,50

Al aplicar la prueba t de Student para datos apareados se puede observar (Tabla 8) que con una escala de 5 puntos no se detectan diferencias estadísticamente significativas (p = 0,062), en cambio al medir lo mismo con una escala de 0 a 10, la mayor sensibilidad del instrumento permite detectar diferencias con p = 0,025.

Tabla 8.

Comparación de las respuestas a un elemento con datos apareados en función del número de opciones.

	M	DT	t	Gl	Sig.
Pair1—Escala 1-5a Escala 1-5b	-,291	1,133	-1,904	54	,062
Pair2—Escala 0-10a Escala 0-10b	-,891	2,859	-2,311	54	,025

13. Experiencia 3: Con todos los elementos de la escala y datos independientes

Se aplica una escala Likert con 10 elementos a una muestra de 79 sujetos divididos en dos grupos de 40 y 39 sujetos respectivamente. Cada uno de los 10 elementos se presenta con el doble formato de 1-5 y de 0-10. Se consideran como dos escalas diferentes y se calcula la suma para cada una de ellas. Después se calcula la correlación entre ambas, cuyo resultado es r = 0,951. Es decir, las dos escalas miden prácticamente lo mismo. Los estadísticos de cada una de ellas aparecen en la Tabla 9. Obsérvese que hay dos escalas (Categoría5 y Categoría10) y que cada una de ellas se ha aplicado a los dos grupos independientes de estudio.

Tabla 9.

Estadísticos de una escala con datos independientes.

	Grupo	N	Media	DT
Categoría5	1	40	43,80	4,59
	2	39	45,08	4,00
Categoría10	1	40	76,40	10,15
	2	39	82,31	9,35

Cuando se aplica la prueba t de Student para comparar los dos grupos se obtienen los resultados siguientes (Tabla 10):

Tabla 10.

Comparación de dos escalas con datos independientes en función del número de puntos.

	Prueba de Levene		Prueba T para la igualdad de medias		
	F	Sig.	t	gl	Sig.(bil)
Categoría5	1,206	,276	-1,315	77	,192
Categoría10	0,573	,451	-2,687	77	,009

Con la escala de 5 puntos no se observan diferencias entre los dos grupos (p = 0,192), mientras que con la escala de 0-10, sí hay significación estadística (p = 0,009).

En conclusión, una escala de 0 a 10 permite detectar diferencias estadísticamente significativas donde una escala de 5 puntos que mide exactamente lo mismo no las detecta. Conviene, no obstante, dejar claro que estos ejemplos no ilustran lo que "acostumbra a pasar". Son ejemplos seleccionados para evidenciar que esto "puede pasar".

14. ¿Unificación de criterios?

El hecho de proponer escalas de 0-10 se podría interpretar, erróneamente, como una imposición para que todos los investigadores lo hagan igual. Conviene dejar claro que se pueden dar múltiples situaciones que justifiquen un determinado número de opciones de respuesta. Por lo tanto, en este sentido, las personas que se dedican a la investigación deben sentirse libres para hacerlo como consideren más conveniente.

Sin embargo, a veces conviene insistir y recordar que en la medida en que se disponga de unos instrumentos con criterios unificados se facilita el progreso científico. Recordemos que un instrumento de medida es el pie. Se conoce el pie romano, inglés, castellano, de Burgos, de agrimensura, maderero, etc. Cada uno de ellos tiene una longitud distinta, lo cual dificulta muchísimo su uso cuando se tienen datos de diversos tipos de pie. El sistema métrico decimal fue un gran paso de cara a la unificación de las medidas para facilitar las comparaciones. Algo parecido sucede con las escalas Likert. ¿Convendría orientar el uso de las escalas Likert hacia una unificación de criterios? ¿O es mejor defender la libertad de cada cual? Vamos a respetar las opiniones al respecto, entre otros factores porque hay que reconocer que, por diversos motivos, no siempre va a ser posible la unificación. Sin embargo, por las razones antes expuestas, si tendiéramos hacia un sistema unificado, por ejemplo de 0-10, se facilitaría la comparación entre diversas investigaciones sobre un mismo tema, que actualmente miden las mismas variables, pero con criterios distintos de valoración, hecho este último que dificulta la comparación. Este puede ser un argumento más para tomar decisiones sobre el número de opciones de los ítems de las escalas Likert en el sentido que estamos proponiendo. Esta tendencia es ya un hecho en las investigaciones del área de salud donde se observa el uso de escalas de 0-10 puntos en trabajos recientes como los de Antiel (2014), Neyens *et al.* (2014) y Greenstein, Greenstein, Senderovich y Mabjeesh (2014), entre otros. Asimismo, en la guía de la OECD (2013) sobre medición de Bienestar Subjetivo se recomienda explícitamente el uso de escalas de 0-10 puntos.

15. Conclusión: propuesta de escalas de 0-10

Las escalas tipo Likert se utilizan con profusión en psicología, educación y ciencias sociales. Son de los instrumentos más utilizados en las investigaciones publicadas en revistas científicas. En cada cuestionario se adopta un número de opciones de respuesta sin que, en general, quede justificada la elección. En este artículo se han aportado argumentos para justificar la decisión sobre el número de opciones más conveniente.

En muchas situaciones de investigación educativa y psicológica interesa detectar posibles diferencias a lo largo del tiempo. Es decir, interesa disponer de instrumentos con sensibilidad a los cambios. Como se ha demostrado, a mayor número de opciones de respuesta, el instrumento tiene mayor sensibilidad para detectar cambios. Esto se ha observado tanto en grupos independientes como en diseños intrasujeto, como es el caso de evaluar los efectos de una acción formativa.

Las escalas de 5 puntos se asumen habitualmente a partir de la idea de que con más opciones es difícil asignar a cada una de ellas una expresión verbal apropiada. Sin embargo, se han expuesto argumentos suficientes que permiten recomendar la no utilización de denominaciones verbales en cada opción, siempre que sea posible.

Además, las personas no suelen utilizar los valores extremos. Por lo tanto, en la práctica, una escala de 5 puntos puede quedar reducida a 4 o 3. Esto afecta mucho menos a los resultados cuando se dispone de una escala con más opciones de respuesta.

A partir de estas consideraciones, la propuesta radica en construir escalas que vayan de 0 a 10, con un total de once opciones de respuesta, siempre que sea posible. Si los investigadores toman en consideración estas propuestas, que se han argumentado de cara a la mejora en el uso de escalas Likert, se pueden facilitar comparaciones posteriores, por ejemplo, entre pretest y postest. En todo caso, sería de desear que los investigadores, en el momento de elegir un determinado número de opciones de respuesta, argumentaran las decisiones a partir de fundamentos epistemológicos y metodológicos. Ello puede ser un paso importante con vistas a pasar de la simple tradición a la decisión con fundamento, propia de la investigación científica.

<Referencias Bibliográficas>

Alwin, D. F. (1997). Feeling thermometers vs 7-point scales. *Sociological Methods and Research, 25* (3), 318-351. DOI: 10.1177/0049124197025003003

Antiel, R. M. (2014). *Professional Burnout, Career Satisfaction, and Wellness Practices. A National Survey of Pediatric Surgeons*. AAP National Conference and Exhibition. San Diego: American Academy of Pediatrics.

Batista-Foguet, J. M., Saris, W., Boyatzis, R., Guillén, L., y Serlavós, R. (2009). Effect of response scale on assessment of emotional intelligence competencies. *Personality and Individual Differences, 46*(5-6), 575–580. DOI:10.1016/j.paid.2008.12.011

Cañadas, I., y Sánchez Bruno, A. (1998). Categorías de respuesta en escalas tipo Likert. *Psicothema, 10*(3), 623-631.

Cummins, R.A. (1997). *The Directory of Instruments to measure quality of life and cognate areas of study. (4th Ed.)*. Melbourne: Deakin University.

Cummins, R.A., y Gullone, E. (2000). Why we should not use 5-point Likert scales: The case for subjective quality of life measurement. *Proceedings, Second International Conference on Quality of Life in Cities* (pp. 74-93). Singapore: National University of Singapore.

Churchill, G., y Peter, J. P. (1984). Research design effects on the reliability of rating scales: A Meta-Analysis. *Journal of Marketing Research*, *21*(4), 360-375. DOI: 10.2307/3151463

Darbyshire, P., y McDonald, H. (2004). Choosing Response Scale Labels and Length: Guidance for Researchers and Clients. *Australasian Journal of Market Research, 12*(2), 17-26.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5 point, 7 point and 10 point scales. *International Journal of Market Research, 50*(1), 61-104.

Diefenbach, M. A., Weinstein, N. D., y O'Reilly, J. (1993). Scales for assessing perceptions of health hazard susceptibility. *Health Education Research, 8*(2), 181-192. DOI:10.1093/her/8.2.181

Dixon, P. N., Bobo, M., y Stevick, R.A. (1984). Response differences and preferences for all-category-defined and end-defined Likert formats. *Educational and Psychological Measurement, 44*(1), 61-66. DOI: 10.1177/0013164484441006

Ferguson, L. W. (1941). A study of the Likert technique of attitude scale construction. *Journal of Social Psychology, 13*(1), 51-57. DOI:10.1080/00224545.1941.9714060

Finstad, K. (2010). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies, 5*(3), 104–110.

Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology, 14*(2), 83-102. http://dx.doi.org/10.1037/h0074329

Fuentes, L. J., González, C., Estévez, A. E., Carranza, J. A., Daza, M., Galián, M. D., y Álvarez, D. (2003). Sensibilidad al funcionamiento de la atención ejecutiva de algunos tests estandarizados en niños de siete años. *Electronic Journal of Research in Educational Psychology, 1* (2), 24-36.

Greenstein, A., Greenstein, I., Senderovich, S., y Mabjeesh, N. J. (2014). Is Diagnostic Cystoscopy Painful? Analysis of 1,320 Consecutive Procedures. *International Brazilian Journal of Urology (IBJU), 40(*4), 533-538. DOI:10.1590/S1677-5538.IBJU.2014.04.13

Guyatt, G.H., y Jaeschke, R. (1990). Measurement in clinical trials: Choosing the appropriate approach. En B. Spilker (Ed.), *Quality of life assessment in clinical trials* (pp.37-46). New York: Raven Press.

Hernández Baeza, A., Espejo, B., González Romá, V., y Gómez Benito, J. (2001). Escalas de respuesta tipo Likert: ¿es relevante la alternativa "indiferente"? *Metodología de encuestas, 3*(2), 135-150.

Hooker, K., y Siegler, I.C. (1993). Life goals, satisfaction, and self-rated health: Preliminary findings. *Experimental aging research, 19*(1), 97-110. DOI: 10.1080/03610739308253925

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 140, 1–55.

Lozano, L. M., García-Cueto, E., y Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*(2), 73-79. DOI:10.1027/1614-2241.4.2.73

Matell, M.S., y Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study 1: Reliability and validity. *Educational and Psychological Measurement, 31*(3), 657-674. DOI: 10.1177/001316447103100307

Muñiz, J., García-Cueto, E., y Lozano, L.M. (2005). Item format and the psychometric properties of the Eysenck Personality Questionnaire. *Personality and Individual Differences, 38*(1), 61–69. doi:10.1016/j.paid.2004.03.021

Neyens, I., Vermeulen, B., Bijlhout, D., y Van Audenhove, C. (2014). *Perceptions on distress of men with prostate cancer and their partner*. Poster presentado en EACH - 12th International Conference on Communication in Healthcare (ICCH). Amsterdam.

OECD (2013). *OECD Guidelines on Measuring Subjective Well-being, OECD Publishing*. http://dx.doi.org/10.1787/9789264191655-en

Roy Morgan Research (1993). *International values audit*, 22/23 May. Melbourne: Roy Morgan Research Centre.

Russell, C., y Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology, 77*(3), 336-342. http://dx.doi.org/10.1037/0021-9010.77.3.336

Shaftel, J.; Nash, B.L. y Gillmor, S. C. (2012) *Effects of the Number of Response* Categories on Rating Scales Roundtable. En *The annual conference of the American Educational Research Association*, Vancouver, British Columbia. En: https://cete.ku.edu/sites/cete.drupal.ku.edu/files/docs/Presentations/2012_04_Shaftel%20et%20al.,%20Number%20of%20Response%20Categories,%204-9-12.pdf

Solomon, E., y Kopelman, R. E. (1984). Questionnaire format and scale reliability: An examination of three modes of item presentation. *Psychological Reports, 54*(2), 447-452. DOI: 10.2466/pr0.1984.54.2.447

Watkins, D., Akande, A., Fleming, J., Ismail, M., Lefner k., Regmi M., Watson S., Yu, j., Adair, J., Cheng, C., Gerong, A., McInerney, D., Mpofu, E., Singh-Sengupta, S., y Wondimu, H. (1998). Cultural dimensions, gender, and the nature of self-concept: A fourteen-country study. *International Journal of Psychology, 33*(1), 17-31. DOI :10.1080/002075998400583

Wyatt, R. C., y Meyers, L. S. (1987). Psychometric properties of four 5-point Likert-type response scales. *Educational and Psychological Measurement, 47*(1), 27-35. DOI: 10.1177/0013164487471003