doi:10.5477/cis/reis.168.129

¿Influye en la respuesta el orden de administración de escalas valorativas 0-10? Una aplicación en encuestas telefónicas

Does the Order of Presentation of 0-10 Rating Scales Affect Responses? An Application to Telephone Surveys

Vidal Díaz de Rada

Palabras clave

Dirección de la escala
- Encuestas telefónicas
- Escalas de medida
- Orden de respuesta

Resumen

Pese a la gran tradición de las investigaciones sobre «efectos de respuesta» en encuestas, hay escasa literatura sobre la influencia de la dirección de administración de escalas ordinales. Investigaciones recientes han demostrado que variar la dirección de una escala de 11 puntos influye en las respuestas obtenidas/logra distribuciones diferentes. El presente trabajo busca analizar hasta que punto estas conclusiones, localizadas en la sociedad norteamericana, se producen también en la sociedad española. Una encuesta telefónica aplicada a dos muestras equivalentes de una comunidad autónoma proporciona una gran similitud en las respuestas con independencia de la dirección de administración de escalas ordinales. De las 14 escalas empleadas, tan solo 3 presentan distribuciones diferentes, influyendo más la edad y el nivel de estudios.

Key words

Scale Direction
- Telephone Surveys
- Measurement Scales
- Response Order

Abstract

Despite the long tradition of research on «response effects» in surveys, there is little literature on the impact of the direction of administration of ordinal rating scales. Recent studies have shown that varying the direction of an 11-point scale produces different distributions. This study seeks to analyse to what extent these conclusions, found among North American society, are also applicable to Spanish society. A telephone survey administered to two equivalent samples of an autonomous region provided very similar responses, regardless of the direction of ordinal rating scales.

Cómo citar

Díaz de Rada, Vidal (2019). «¿Influye en la respuesta el orden de administración de escalas valorativas 0-10? Una aplicación en encuestas telefónicas». *Revista Española de Investigaciones Sociológicas,* 168: 129-140. (http://dx.doi.org/10.5477/cis/reis.168.129)

La versión en inglés de esta nota de investigación puede consultarse en http://reis.cis.es

Vidal Díaz de Rada: Universidad Pública de Navarra | vidal@unavarra.es

Introducción

El estudio de los «efectos de respuesta» constituye un ámbito de estudio con gran tradición en la investigación con encuestas. Con más o menos intensidad, esta temática ha centrado la atención de numerosos investigadores desde mediados de la década de los años cuarenta hasta la actualidad (entre otros, Kamoen *et al*., 2011). La influencia del orden en que son administradas las categorías se ha explicado considerando numerosos factores. Una de las primeras interpretaciones explica este proceso aludiendo a los distintos tipos de *memoria*, considerando que la memoria a *largo plazo* es utilizada para retener las categorías presentadas en primer lugar, desplazando a la memoria a *corto plazo* a las últimas categorías (entre otros, Bruce y Papay, 1970).

Otros expertos (entre otros, Simon, 1957) aluden al deseo de terminar la respuesta al cuestionario cuanto antes, generando así una «situación de prisa» que dificulta notablemente tanto la comprensión de la pregunta como la elección de respuestas adecuadas (Tourangeau y Rasinski, 1988).

Una tercera vía interpretativa, desarrollada por Krosnick y Alwin (1987), considera que —pese a la existencia de ideas asentadas en la mente del entrevistado— la rapidez en seleccionar la información genera que las primeras alternativas provoquen la creación de un marco cognitivo por el que van a ser juzgadas las siguientes.

Sin desdeñar la importancia de estos aspectos, en los últimos años están apareciendo numerosos trabajos que demuestran que la dirección de administración de las escalas valorativas influye considerablemente en las respuestas de los entrevistados (entre otros, Bassili y Krosnick, 2000; Tourangeau, Couper y Conrad, 2013; Yan y Keusch, 2015). Estos expertos localizan un efecto «dirección de la escala», esto es, detectan que administrar una escala comenzando por los valores más bajos o por los más altos produce diferencias en las respuestas de los entrevistados. Este efecto, localizado en otros países, es el origen de la presente investigación. Se busca demostrar hasta qué punto estos hallazgos, localizados inicialmente en una encuesta de temas económicos (Yan y Keusch, 2015) y constatados después en estudios sobre temas políticos (Liu y Keusch, 2017; Yan, Keusch y He, 2018), se producen también en España.

En cuanto a la estructura del trabajo, comienza con un epígrafe donde se realiza una breve exposición de los trabajos más relevantes sobre la influencia de las categorías de respuesta, finalizando con los efectos producidos por la dirección de la escala. En el segundo se presenta la investigación utilizada para comprobar la presencia de este efecto en nuestro entorno, especificando las preguntas utilizadas y las técnicas de análisis de datos empleadas. En el tercer epígrafe se analizan los efectos en cada tipo de pregunta, precediendo a las conclusiones.

Influencia del orden de administración de las respuestas de un cuestionario

De todos los posibles efectos producidos por el orden de administración de las respuestas, este trabajo centrará la atención en las escalas valorativas. El uso de este tipo de escalas es muy usual en investigaciones sociales y políticas, siendo innumerables los ámbitos en los que se han utilizado: valoración de políticos y líderes, satisfacción de productos o servicios, probabilidades de participación política, ideología izquierda-derecha, etc.

Decidida la utilización de una pregunta de este tipo, posteriormente el investigador debe determinar el número de categorías de respuesta, existencia/ausencia de opción intermedia, empleo de términos en los extremos o en cada una de las categorías, uso de una formulación acuerdo/desacuerdo o de una escala específica, y el orden de administración de las alternativas de respuesta

(Krosnick y Presser, 2010). El investigador, lógicamente, utilizará cada uno de estos elementos en función de las necesidades de análisis, en la medida que las decisiones tomadas en este momento determinarán las técnicas de análisis de los datos a utilizar.

Cada una de estas decisiones requerirá de más o menos esfuerzo por parte del entrevistado. Así, por ejemplo, el formato acuerdo/desacuerdo con 11 categorías (0 a 10) con etiquetas únicamente en los extremos, muy utilizado habitualmente, es desaconsejado por numerosos expertos (entre otros, Dillman, Smyth y Christian, 2014; Revilla, Saris y Krosnick, 2014) por el mayor esfuerzo cognitivo que debe hacer el encuestado para procesar una información de este tipo. A estas implicaciones, comprobadas en numerosas investigaciones (entre otros, Krosnick y Presser, 2010), se añade que muchas personas tienen dificultad para expresarse en términos numéricos; puesto que implica un doble proceso: formular una opinión, que deberá posteriormente ser «convertida» a un criterio numérico.

Algunos experimentos (entre otros, Tourangeau, Couper y Conrad, 2013) localizan que las escalas numéricas con etiquetas en los extremos son respondidas con más rapidez cuando comienzan con las respuestas más positivas-favorables (el 10 en este caso), aunque advierten también que esta forma de colocación produce una mayor elección de las primeras opciones de la escala. Investigaciones más recientes (Dillman, Smyth y Christian, 2014) no detectan diferencias en la distribución cuando las escalas comienzan con el valor más positivo-favorable o el más negativo-desfavorable, confirmando que se responde más rápido cuando las escalas comienzan por los valores positivos-favorables.

Ante la disparidad de resultados de estas y otras investigaciones, la Association for Public Opinion Research-AAPOR realizó —en 2013 y 2014— dos encuentros sobre el tema que concluyeron que una misma amplitud de escala e igualdad en las etiquetas numéricas y/o verbales no garantiza la validez, en la medida que la dirección de la escala influye en la percepción —y en las respuestas— de los entrevistados. Conscientes de estos hallazgos, y considerando las diferencias entre la aplicación de las escalas en Norteamérica (del «totalmente de acuerdo» al «totalmente en desacuerdo») y en los Países Bajos (de forma inversa), Yan y Keush (2015) utilizaron una escala sobre «valoración del desarrollo de los países» que fue aplicada en ambas direcciones: una de más desarrollado a menos (10-0), y otra de menos a más (0-10).

Los resultados obtenidos desvelan, por un lado, unas puntuaciones más altas para todos los países (mayor desarrollo) cuando la escala comienza por el número más alto, concluyendo que «[...] las calificaciones presentan más variación en los países desarrollados cuando la escala comienza con 0, y son más estables en los países subdesarrollados cuando la escala comienza con 10» (Yan y Keush, 2015).

Buscando constatar la generalidad de sus hallazgos a otras temáticas se procedió de forma similar en un estudio con dos encuestas sobre temas políticos (Liu y Keusch, 2017; Yan, Keusch y He, 2018). Los resultados desvelan que la dirección de administración de la escala presenta más influencia en preguntas no actitudinales, cuando están colocadas en la segunda mitad del cuestionario, y cuando la escala presenta una gran amplitud (Yan, Keusch y He, 2018).

Pese a que las primeras investigaciones señalaron una influencia escasa del nivel educativo en los efectos de respuesta (entre otros, Schuman y Presser, 1981), investigaciones posteriores —utilizando la técnica del metaanálisis— han destacado una gran influencia del nivel educativo y la edad en la influencia del orden de respuestas (Krosnick, Narayan y Smith, 1996): el nivel educativo está relacionado con el uso de habilidades

cognitivas. Respecto a la edad, a partir de los 65 años se produce una pérdida de las capacidades cognitivas, principalmente un descenso de la memoria.

Diseño metodológico: aplicación de los hallazgos internacionales al estudio en una comunidad autónoma

Siguiendo la lógica del primer experimento de Yan y Keusch (2015) se utilizó una encuesta telefónica a teléfonos fijos en hogares. Considerando el Padrón como marco muestral se elaboraron dos muestras equivalentes de 448 personas estratificadas según zona de residencia y hábitat del municipio. Los municipios fueros seleccionados aleatoriamente, y los entrevistados, utilizando cuotas de sexo y edad. La realización de 30 rellamadas en los hogares no contactados y el empleo de estrategias de conversión de rechazos «suaves» proporcionaron una tasa de respuesta (TR4) del 64%, muy similar a la investigación inicial de Yan y Keusch (2015). La relevancia de la institución que realiza el estudio y la brevedad del cuestionario son aspectos que explican esta elevada cooperación.

La decisión de optar por una muestra a teléfonos fijos se fundamenta en el intento de replicar al máximo la investigación de Yan y Keusch, limitada a telefonía fija. Aunque es evidente que una muestra a estos teléfonos presenta algunos problemas de representatividad, dado que el Instituto Nacional de Estadística (2017) estima que un 17,4% de la población navarra no dispone de este equipamiento, el objetivo del trabajo no es tanto la generalización a una población, sino el análisis de las diferencias entre los tratamientos experimentales.

Es importante aludir a dos especificidades de la encuesta telefónica, como son la declaración de respuestas más extremas (Ye *et al*., 2011) y que el empleo de un canal oral puede plantear problemas al receptor para procesar adecuadamente las primeras alternativas de respuesta. La disposición de otras alternativas genera que la primera se procese a una velocidad mayor, aumentando así el número de elecciones de la última categoría (Gwartney, 2007; Díaz de Rada, 2010).

El cuestionario disponía de dos versiones (denominadas A y B) que se diferencian únicamente en el orden de aplicación de las opciones de respuesta, y que fueron administradas a dos muestras equivalentes. Las muestras son similares considerando sexo, edad, nivel de estudios, relación con la actividad, tamaño del hogar y situación de convivencia (valores V de Cramer inferiores a 0,08; con significaciones notablemente superiores a 0,10). Al tratarse de un cuestionario sobre actitudes políticas se ha comprobado también la similitud en ideología del entrevistado, intención de voto, simpatía hacia partidos y recuerdo de voto en las últimas elecciones autonómicas (valores V de Cramer 0,096, 0,138, 0,211 y 0,108; con significaciones 0,759, 0,145, 0,086 y 0,580 respectivamente).

El trabajo de campo se llevó a cabo en septiembre de 2017 con 10 entrevistadores. Estos no tenían conocimiento del objeto del estudio y, con el fin de eliminar su posible influencia, realizaron alternativamente cada cuestionario, de modo que todos aplicaron el mismo número de cuestionarios A y B. Con el fin de eliminar el posible impacto del horario, cada entrevistador debía realizar seguidos un cuestionario A y otro B, no terminando la sesión hasta que realizara ambos[1].

En cuanto a la temática del estudio, se utilizaron preguntas «usuales» en investigación política: la «habitual» escala de valoración (0-10) de líderes políticos, autodefinición de la ideología política del entrevistado, escala de valoración de cuatro instituciones relevantes en Navarra, valoración del gobier-

[1] Es decir, no se permitía terminar la sesión con el cuestionario A y dejar el B para el día siguiente, sino que ambos debían realizarse seguidos en la misma sesión de trabajo.

CUADRO 1. *Preguntas utilizadas*

Escala de valoración (0-10) de portavoces parlamentarios

¿Cómo valora la actuación política de ____ [NOMBRE DE CADA PORTAVOZ PARLAMENTARIO], utilizando una escala…

A. … de 0 a 10, donde 0 significa «muy mal» y 10 «muy bien»?

B. … de 10 a 0, donde 10 significa «muy bien» y 0 «muy mal»?

Autodefinición ideológica del entrevistado

Cuando se habla de política normalmente se utilizan las expresiones izquierda y derecha. ¿Me podría decir dónde se ubicaría usted en una escala…

A. … de 0 a 10, donde 0 significa extrema izquierda y 10 extrema derecha?

B. … de 10 a 0, donde 10 significa extrema derecha y 0 extrema izquierda?

Valoración de cuatro instituciones relevantes en Navarra

Cómo valora la actividad del Parlamento de Navarra empleando una escala…

A. … de 0 a 10, donde 0 es «valoración muy mala» y 10 es «valoración muy buena».

B. … de 10 a 0, donde 10 es «valoración muy buena» y 0 es «valoración muy mala».

Misma escala para el Defensor del Pueblo de Navarra, la Cámara de Comptos y el Delegado del Gobierno de Navarra.

Valoración de la gestión del gobierno autonómico y de su presidente

¿Cómo calificaría la actividad desarrollada por la presidenta del Gobierno de Navarra, Uxue Barkos, utilizando una escala…

A. … de 0 a 10, en la que 0 significa que funciona «muy mal» y 10 que funciona «muy bien»?

B. … de 10 a 0, en la que 10 significa que funciona «muy bien» y 0 que funciona «muy mal»?

Misma escala para el gobierno autonómico.

Fuente: Elaboración propia.

no autonómico y de su presidente (cuadro 1), repartidas a lo largo del cuestionario para evitar sesgos en las respuestas (Alvira, 2011). Son preguntas habituales en los «barómetros políticos» del Centro de Investigaciones Sociológicas (en adelante, CIS), y han sido frecuentemente utilizadas también en encuestas telefónicas (entre otros, CIS, 2017). Se ha seguido una recomendación de Alvira cuando señala que «al valorar, por ejemplo, a los líderes políticos, puede utilizarse como ayuda… una escala de valoración entre el 0 y el 10…» (2011: 36). En el cuestionario «A» esta escala fue colocada en dirección ascendente de 0 a 10, siendo la primera «valoración peor» y 10 la «mejor» (cuadro 1), invirtiendo la dirección de lectura en el cuestionario B de 10 a 0, de modo similar a como procedieron Schuman y Presser (1981) en su investigación clásica sobre aquiescencia en preguntas de acuerdo y desacuerdo.

Es importante dar cuenta que numerosas investigaciones han localizado que los entrevistados prestan más atención al número (de la escala) que al texto colocado en los extremos (entre otros, Sudman, Bradburn y Schwarz, 1996; Schwarz *et al*., 1991), influencia que es mayor en la encuesta telefónica por la mayor dificultad de recordar todas las opciones.

Se planteaba como hipótesis, siguiendo los trabajos previos sobre el tema (entre otros, Tourangeau, Couper y Conrad, 2013), unas valoraciones medias más altas en la escala 10 a 0 debido, entre otras razones, a la presencia de un efecto primacía. La segunda hipótesis considera que estos hallazgos, localizados en escalas de 0 a 10, disminuyen cuando la amplitud de la pregunta es menor. La tercera hipótesis, basada en los trabajos de Dillman, Smyth y Christian (2014), postula un tiempo de respuesta menor cuando las escalas comienzan por valores positivos.

Respecto a las técnicas de análisis de datos, para localizar el efecto producido por la diferente administración se utilizará el test de significación de la diferencia de medias, tal y como han procedido investigaciones similares realizadas en otros contextos (entre otros, Chang y Krosnick, 2010).

Al final del primer epígrafe se ha dado cuenta de la gran influencia del nivel de estudios y la edad en los efectos de respuesta, variables que han sido codificadas en tres y cuatro categorías en el presente estudio: básicos, secundarios y superiores en el nivel de estudios, y 16-29, 30-49, 50-64 y más de 65 años. En variables multicategóricas no es posible emplear la diferencia de medias, por lo que se ha optado por el análisis de varianza de un factor, utilizando la prueba de Brown-Forsythe en las distribuciones no homocedástidas. Con el fin de detectar la posible influencia conjunta del tipo de cuestionario, sexo y nivel de estudios se empleará el análisis de varianza de dos factores.

Resultados

Los resultados están estructurados siguiendo la lógica del cuadro 1, con tres epígrafes donde se analiza la escala de valoración de políticos y la autodefinición de la ideología política del entrevistado, en el segundo la valoración de varias instituciones relevantes en la comunidad, finalizando con la valoración de la gestión del gobierno autonómico y del presidente.

Escala de valoración (0-10) de portavoces parlamentarios y autodefinición de la ideología política del entrevistado

Los portavoces parlamentarios son, en su mayoría, los «cabezas de lista» o líderes de los partidos políticos que concurrieron en las últimas elecciones autonómicas. Esto es así en todos los casos excepto en el partido que gobierna, cuyo portavoz fue propuesto como diputado para las elecciones generales de

TABLA 1. *Escala de valoración de portavoces parlamentarios y autodefinición ideológica*

	No respuesta	Nivel de conocimiento	Nº de casos	Media total	Media 0-10	Orden 10-0	Diferencia
EH-Bildu	16/14	33,6%	257	4,3	4,4	4,2	0,21
PP	17/14	36,6%	271	3,5	3,4	3,6	-0,28
PSOE	17/14	53,0%	395	4,7	4,6	4,8	-0,18
UPN	16/14	51,3%	388	3,9	3,6	4,2	-0,54*
Geroa Bai	15/14	34,1%	260	5,2	5,3	5,0	0,27
Izquierda Esquerra-IE	16/15	13,0%	100	4,6	4,6	4,6	-0,03
Podemos	16/15	10,0%	67	4,5	4,7	4,3	0,32
Ideología	29/26		792	4,5	4,5	4,5	0

* <0,05.

Nota: Se presentan los líderes en el orden en que fueron preguntados.

Fuente: Elaboración propia.

diciembre de 2015. Es preciso considerar también que el portavoz de Podemos es el que menos tiempo lleva desarrollando esta función, desde el 3 de julio de 2017.

Antes de solicitar la valoración de cada líder se pidió a los entrevistados si eran capaces de identificar cada uno con el partido político al que pertenecen, obteniéndose los resultados que se muestran en la segunda columna de la tabla 1[2]. Los portavoces del PSOE y de UPN son los más reconocidos, situándose el portavoz de Podemos en la situación opuesta, sin duda por las razones apuntadas en el párrafo anterior. Sorprende el elevado conocimiento del portavoz de Geroa Bai, que no se presentó a las elecciones al Parlamento autonómico, aunque puede explicarse por su concurrencia a las elecciones generales. Los entrevistados que identificaron correctamente los candidatos con su partido fueron posteriormente preguntados por la valoración de cada líder; lo que implica una reducción del tamaño muestral, notable en el caso de los portavoces de IE y Podemos. El bajo tamaño muestral de este último, que implica 38 casos en la muestra A y 29 en la B, recomienda considerarlo con suma prudencia.

La aplicación 0-10, la habitual en la mayor parte de encuestas, sitúa al portavoz de Geroa Bai como el mejor valorado, seguido de los portavoces de Podemos y PSOE-UE. Cuando se considera la valoración en la administración 10-0 se localizan descensos en los portavoces de tres de los cuatro partidos que gobernaban en Navarra cuando se realizó la investigación: Geroa Bai, EH-Bildu y Podemos. Solo IE no cambia su valoración. Obsérvese que se trata de diferencias pequeñas, excepto en el caso del portavoz de UPN que se acerca al medio punto, la única diferencia significativa. La comparativa entre ambos supone también una alteración del orden de mejor valoración de los portavoces, al obtener los portavoces del PSOE e IE la segunda y tercera mejor valoración.

Constatada la ausencia del efecto dirección de escala, se procederá con el resto de variables susceptibles de influencia. Así, el nivel de estudios presenta una diferencia significativa en el caso de los portavoces del PSOE (*F* 10,5, *p* 0,000), UPN (*F* 5,9, *p* 0,003)

[2] En la primera columna se muestra el número de entrevistados que rechazaron responder la pregunta en cada muestra. Son 14-16 entrevistados en cada muestra, que suponen apenas un 3,8% del total, un escaso tamaño sin influencia en los resultados de la investigación. Un análisis detallado de este colectivo desvela que 30 no valoran a ningún político.

TABLA 2. *Valoración de la actividad desarrollada por cuatro instituciones navarras*

	No respuesta	Nivel de conocimiento*	Nº de casos	Media total	Media 0-10	Orden 10-0	Diferencia
Parlamento	7/6	41,9	345	5,2	5,3	5,2	0,05
Defensor del pueblo	2/2	40,7	322	5,6	5,4	5,7	-0,25
Cámara de Comptos	3/2	38,8	291	6,1	6,2	6,0	0,21
Delegado del Gobierno	2/2	35,2	291	4,3	4,0	4,6	-0,57*

* <0,05.

Fuente: Elaboración propia.

y —menor— del PP (*F* 3,79, *p* 0,049). Los tres portavoces logran altas valoraciones en los entrevistados con bajos niveles de estudios (5,5, 4,7 y 4,4 respectivamente), puntuación que desciende en los entrevistados con niveles de estudios elevados (4,3, 3,5 y 3,2). Los mismos líderes presentan también diferencia cuando se considera la edad de los entrevistados, desvelando que a medida que aumenta la edad desciende la valoración de cada líder. Esto sucede hasta los 64 años, cambiando la tendencia en los entrevistados de más edad, que asignan las puntuaciones más superiores. Ante la sospecha de que la variabilidad dentro de cada submuestra pudiera ser la causa de las diferencias entre muestras se analizó cada una por separado, localizando la misma tendencia. Ahora bien, no hay diferencia significativa cuando se consideran conjuntamente las tres variables (orden del cuestionario, edad y estudios).

En la parte inferior de la tabla se aprecia la nula diferencia de la ideología, con promedios idénticos en ambas muestras.

Valoración de la actividad desarrollada por cuatro instituciones relevantes

El segundo aspecto, tal y como se señaló en el cuadro 1, es valorar la actividad desempeñada por el Parlamento de Navarra, el Defensor del Pueblo autonómico, la Cámara de Comptos y el Delegado del Gobierno. Las respuestas se han administrado exactamente igual que en el caso anterior, con una falta de respuesta de 4-5 entrevistados, que llega al 13 en la pregunta sobre el Parlamento.

En cuanto a los resultados, destaca un nivel de conocimiento promedio ligeramente superior al 40% en el caso del Parlamento y del Defensor del Pueblo, y algo menor en la Cámara de Comptos y el Delegado del Gobierno (tabla 2). Respecto a las valoraciones, que fueron expresadas únicamente por los que conocían cada institución, el Delegado del Gobierno es el que presenta menores magnitudes, peor valoración. Ahora bien, es también el que obtiene mayores diferencias, medio punto entre una escala y otra; logrando mejores valoraciones en la escala 10-0, tal y como se planteó en la hipótesis. Las diferencias son menores en el caso del Defensor del Pueblo y la Cámara de Comptos, con diferencias ligeramente superiores a los 0,20 puntos, que no llegan a ser significativas. Resulta sorprendente, en el caso de esta última, que la valoración 0-10 consiga una valoración media superior a la escala inversa.

El estudio de las diferencias según nivel de estudios y edad únicamente aporta diferencias significativas en la edad en la Cámara de Comptos y el Delegado del Gobierno. En la primera, la valoración mejora a medida que aumenta la edad de los entrevistados, mientras que en el caso del Delegado del Gobierno las valoraciones más altas se producen en los grupos extremos: menores de 30 y mayores de 65 años. No hay diferencia

TABLA 3. *Valoración de la gestión del Gobierno autónomo y de su presidente*

	No respuesta	Nº de casos	Media total	Media 0-10	Orden 10-0	Diferencia
Valoración Gobierno	9/7	824	4,8	4,9	4,8	0,07
Valoración presidente	13/10	801	4,9	4,9	4,8	0,09

Fuente: Elaboración propia.

en el análisis conjunto de estudios, edad y dirección de las respuestas.

Valoración de la gestión del Gobierno autonómico y del presidente

El tercero de los aspectos considerados recoge información de casi toda la muestra[3], en la medida que son suficientemente conocidos por la mayor parte de los entrevistados. La diferencia es la menor de todas las consideradas (tabla 3). Únicamente la edad desvela una relación significativa en ambas preguntas, donde los mayores realizan mejoras valoraciones.

Otros factores de posible influencia

Localizada una escasa variabilidad, considerando la dirección de administración de las escalas 0-10 y logrando resultados diferentes a lo detectado en las investigaciones de Keusch, es el momento de constatar hasta qué punto estas escalas numéricas con etiquetas en los extremos son respondidas con más rapidez cuando comienzan con las respuestas más positivas, resultados constatados también por el equipo de Dillman (Dillman, Smyth y Christian, 2014).

Antes de proceder, debe tenerse en cuenta que las escalas empleadas en las secciones anteriores suponen, tan solo, 14 variables de las 80 comprendidas en el cuestionario, por lo que se decidió comparar las preguntas ordinales del cuestionario tipo A (que comienzan por concepciones positivas-favorables: «muy bueno», «muy favorable», etc.) con el cuestionario B, cuyas preguntas comienzan por concepciones negativas-desfavorables: «muy malo», «muy desfavorable», etc. Según los planteamientos recogidos en el párrafo anterior, el cuestionario A debiera tener menor duración que el cuestionario B. La segunda hipótesis, que daba cuenta que las diferencias disminuyen cuando la amplitud de la pregunta es menor, precisa considerar por separado estas preguntas, la mayor parte utilizando escalas de 4-5 categorías.

El cuestionario A fue respondido en un promedio de 11,9 minutos y el B precisó de 12,01, diferencia no significativa, que vuelve a constatar la escasa influencia de la dirección de administración de las respuestas.

Conclusiones

Apenas existe diferencia en las escalas de 0 a 10 según sean administradas en una dirección o en otra. Tan solo dos ítems, de los 14 utilizados, presentan diferencias relacionadas con la dirección de administración, siendo mayores los efectos de la edad y del nivel de estudios. Los efectos de estas variables podrían estar encubriendo otros factores, como el efecto *recencia* propia de las encuestas telefónicas, los diferentes niveles de cultura política y la *intensidad* de las *actitudes* políticas (Bassili y Krosnick, 2000). La primera interpretación implica una mayor elección de las últimas categorías de respuesta en los entrevistados con bajos niveles educativos, algo que no se aprecia cuando se analiza cada submuestra por separado. Tampoco se

[3] A excepción de 16 y 23 entrevistados que no responden ambas preguntas, 9 y 13 en la muestra A, 7 y 10 en la muestra B.

localiza influencia de la mayor o menor cultura política. La comprobación de la tercera vía interpretativa precisa analizar el interés por la política, en la medida en que los más interesados presentarán mayores diferencias en sus valoraciones políticas, algo que no llega a suceder. La influencia del nivel de estudios y la edad es menor en cada submuestra que cuando se comparan ambas, aunque las diferencias no son significativas.

Las diferencias son menores en el resto de preguntas del cuestionario, preguntas ordinales de 4 y 5 categorías. Por último, y respecto al tiempo de administración, la duración es similar en los cuestionarios A y B.

¿Resultados desalentadores? Es una de las sensaciones que pueden surgir en el lector que ha llegado hasta aquí. Mucho más cuando en la primera parte se han mostrado un gran número de trabajos que apuntaban en otra dirección. Todo lo contrario, es la sensación que surge en el autor en el momento que escribe estas líneas. La estabilidad de los hallazgos constata la adecuación de la herramienta de la encuesta en nuestro contexto, «valida» los hallazgos de investigaciones que —quizá por costumbre— siempre utilizan las escalas en sentido ascendente.

La ausencia del efecto señalado en el primer epígrafe puede explicarse considerando las diferencias culturales entre el país donde se detectaron tales efectos y el lugar donde se ha realizado la presente investigación.

Toda investigación tiene sus limitaciones, y en este caso la principal versa sobre la generalización de los hallazgos, al haber sido localizados en una comunidad autónoma. Es una limitación habitual de los «experimentos con encuestas» que utilizan muestras equivalentes, muchas de las cuales se realizan en muestras cautivas como clientes, estudiantes universitarios, etc. Aunque diversos expertos han comprobado la capacidad de generalización de estas muestras[4], los hallazgos aquí localizados precisan de una mayor investigación sobre el tema. Otro factor que puede ayudar a explicar estos resultados es la mayor formación del personal que realizó el trabajo de campo, la exhaustiva monitorización y la insistencia —y comprobación— en que el cuestionario fuera leído exactamente como está redactado, algo que no siempre tiene lugar. Todo ello precisa de una mayor investigación sobre el tema.

Bibliografía

Alvira, Francisco (2011). *La encuesta: una perspectiva general metodológica*. Madrid: CIS.

Bassili, John N. y Krosnick, Jon A. (2000). «Do Strength-related Attitude Properties Determine Susceptibility to Response Effects?». *Political Psychology*, 21(1): 107-132.

Bruce, Darly y Papay, James P. (1970). «Primacy Effects in Single Trial Free Recall». *Journal of Verbal Learning and Verbal Behaviour*, 9: 473-486.

Centro de Investigaciones Sociológicas (2017). *Preelectoral de Cataluña. Elecciones autonómicas 2017*, estudio 3198. Disponible en: http://www.analisis.cis.es/cisdb.jsp

Chang, Linchat y Krosnick, Jon A. (2010): «Comparing Oral Interviewing with Self-administered Computerized Questionnaires: An Experiment». *Public Opinion Quarterly*, 74: 154-167.

Díaz de Rada, Vidal (2010). *Comparación entre los resultados obtenidos por encuestas personales y telefónicas*. Madrid: CIS.

Dillman, Don; Smyth, Jolene D.; Melani Christian, Leah; y Stern, Michael J. (2005). *Comparing check-all and forced-choice question formats in web survey.* Pullman, Washington: SERSC.

Dillman, Don A.; Smyth, Jolene D. y Christian, Leah Melani (2014). *Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method (4ª ed.)*. New York: Wiley.

[4] Véanse, por ejemplo, los estudios de Dillman *et al*. (2005), entre otros, con estudiantes de la Universidad de Washington, cuyos hallazgos han sido generalizables en muestras nacionales.

Gwartney, Patricia (2007). *The Telephone Interviewer's Handbook: How to Conduct Standardized Conversations*. San Francisco: Jossey-Bass.

Instituto Nacional de Estadística (2017). *Encuesta de Tecnologías de la Información en los Hogares*. Madrid: INE.

Kamoen, Naomi; Holleman, Bregje; Mak, Pim y Sanders, Ted J.M. (2011). «Agree or Disagree? Cognitive Processes». *Discourse Processes*, 48(5): 355-385.

Krosnick, Jon A. y Alwin, Duane F. (1987). «An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement». *Public Opinion Quarterly*, 51: 201-219.

Krosnick, Jon A. y Presser, Stanley (2010). «Question and Questionnaire Design». En: Marsden, P. V. y Wright, J. D. (eds.). *Handbook of Survey Research*. Bingley: Emerald.

Krosnick, Jon A.; Narayan, Sowmya y Smith, Wendy R. (1996). «Satisficing in Surveys: Initial Evidence». *New Directions for Evaluation*, 70: 29-44.

Liu, Mingnan y Keusch, Florian (2017). «Effects of Scale Direction on Response Style of Ordinal Rating Scales». *Journal of Official Statistics*, 33: 137-154.

Revilla, Melanie A.; Saris, Willem E. y Krosnick, Jon A. (2014). «Choosing the Number of Categories in Agree-Disagree Scales». *Sociological Methods and Research*, 43: 73-97.

Schuman, Howard y Presser, Stanley (1981). *Questions and Answers in Attitude Surveys*. New York: Academic.

Schwarz, Norbert; Knäuper,Bärbel; Hippler, Hans-J.; Noelle-Neumann, Elisabeth y Clark, Leslie (1991). «Numeric Values May Change the Meaning of Scale Labels». *Public Opinion Quarterly*, 55: 570-582.

Simon, Hebert A. (1957). *Models of Man*. New York: Wiley.

Sudman, Seymour; Bradburn, Norman y Schwarz, Norbert (1996). *Thinking about Answers*. San Francisco: Jossey-Bass.

Tourangeau, Roger y Rasinski, Kennet (1988). «Cognitive Processes Underlying Context Effects in Attitude Measurement». *Psychological Bulletin*, 103: 299-314.

Tourangeau, Roger; Couper, Mick P. y Conrad, Frederick (2013). «Up Means Good: The Effect of Screen Position on Evaluative Ratings in Web Surveys». *Public Opinion Quarterly*, 77: 69-88.

Yan, Tin y Keusch, Florian (2015). «The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey». *Public Opinion Quarterly*, 79: 145-165.

Yan, Tin; Keusch, Florian y He, Lirui (2018). «The Impacts of Question and Scale Characteristics on Scale Direction Effects». *Survey Practice*, 11(2).

Ye, Cong; Fulton, Jenna y Tourangeau, Roger (2011). «More Positive o more Extreme?». *Public Opinion Quarterly*, 72: 349-365.

RECEPCIÓN: 09/04/2018
REVISIÓN: 25/03/2019
APROBACIÓN: 18/01/2019

doi:10.5477/cis/reis.168.129

Does the Order of Presentation of 0-10 Rating Scales Affect Responses? An Application to Telephone Surveys

¿Influye en la respuesta el orden de administración de escalas valorativas 0-10? Una aplicación en encuestas telefónicas

Vidal Díaz de Rada

Key words

Scale Direction
- Telephone Surveys
- Measurement Scales
- Response Order

Abstract

Despite the long tradition of research on "response effects" in surveys, there is little literature on the impact of the direction of administration of ordinal rating scales. Recent studies have shown that varying the direction of an 11-point scale produces different distributions. This study seeks to analyse to what extent these conclusions, found among North American society, are also applicable to Spanish society. A telephone survey administered to two equivalent samples of an autonomous region provided very similar responses, regardless of the direction of ordinal rating scales.

Palabras clave

Dirección de la escala
- Encuestas telefónicas
- Escalas de medida
- Orden de respuesta

Resumen

Pese a la gran tradición de las investigaciones sobre "efectos de respuesta" en encuestas, hay escasa literatura sobre la influencia de la dirección de administración de escalas ordinales. Investigaciones recientes han demostrado que variar la dirección de una escala de 11 puntos influye en las respuestas obtenidas/logra distribuciones diferentes. El presente trabajo busca analizar hasta que punto estas conclusiones, localizadas en la sociedad norteamericana, se producen también en la sociedad española. Una encuesta telefónica aplicada a dos muestras equivalentes de una comunidad autónoma proporciona una gran similitud en las respuestas con independencia de la dirección de administración de escalas ordinales. De las 14 escalas empleadas, tan solo 3 presentan distribuciones diferentes, influyendo más la edad y el nivel de estudios.

Citation

Díaz de Rada, Vidal (2019). "Does the Order of Presentation of 0-10 Rating Scales Affect Responses? An Application to Telephone Surveys". *Revista Española de Investigaciones Sociológicas,* 168: 129-140. (http://dx.doi.org/10.5477/cis/reis.168.129)

Vidal Díaz de Rada: Universidad Pública de Navarra | vidal@unavarra.es

Introduction

The study of "response effects" is a field with a long tradition in survey research. To a lesser or greater extent, this topic has received the attention of numerous researchers from the mid-1940s to the present day (among others, Kamoen *et al*., 2011). The influence of the order in which the categories are displayed has been explained by considering numerous factors. One of the first interpretations referred to the different types of *memory*, as it held that *long-term* memory is used to retain the categories presented first, whereas the categories presented later are stored in *short-term* memory (among others, Bruce and Papay, 1970).

Other experts (Simon, 1957, among others) have alluded to the desire to complete the questionnaire as quickly as possible, which generates time pressure and makes it difficult to understand the question and choose appropriate answers (Tourangeau and Rasinski, 1988).

A third interpretation, developed by Krosnick and Alwin (1987), advocates that, even if respondent's have settled ideas on the issues in question, the speed at which information is selected means that the first questions create a cognitive framework that will be used to assess those appearing later.

Without neglecting the importance of these aspects, numerous studies have recently shown that the direction of rating scales greatly influences survey responses (among others, Bassili and Krosnick, 2000, Tourangeau, Couper and Conrad 2013; Yan and Keusch, 2015). These experts found a "scale direction effect", that is, they identified variations in responses when a scale starting with the lowest values was used, as opposed to a scale that began with the highest values. This effect, found in other countries, is the basis for this study. The aim is to demonstrate to what extent these findings, initially identified in a survey on economic issues (Yan and Keusch, 2015) and later confirmed in surveys on political topics (Liu and Keusch, 2017; Yan, Keusch and He, 2018), are reflected in Spain.

The study begins with a brief discussion of the most significant studies on the effect of response categories, and on the effects produced by scale direction. The second section contains a description of the survey used to check the presence of this effect in Spain, specifying the questions and data analysis techniques employed. The third section analyses the effects of each type of question, and is followed by the conclusions.

The influence of the order of questionnaire response categories

Out of all the possible effects produced by the order of presentation of responses, this study will focus on rating scales. This type of scales are very commonly used in social and political investigations, and they have been used in surveys in countless areas which have included assessment of politicians and leaders, satisfaction with products or services, probability of political participation, left-right ideology, etc.

Once the use of this type of question has been decided, the researcher must subsequently decide on the number of response categories, the existence/absence of an intermediate option, the use of items at the end of the scale or for each of the categories, the use of an agree-disagree format and of a specific scale, and the order of presentation of response choices (Krosnick and Presser, 2010). The researcher will use each of these elements according to the needs of the analysis, since the decisions taken at this time will determine the data analysis techniques used.

Each of these decisions will affect the degree of effort required on the part of the respondent. For example, the agree-disagree format with 11 categories (0 to 10) with labels only at the ends is very commonly used, but

has been discouraged by numerous experts (among others, Dillman, Smyth and Christian, 2014; Revilla, Saris and Krosnick, 2014) due to the greater cognitive effort that the respondent must make to process information of this type, as proven in numerous studies (among others, Krosnick and Presser, 2010). In addition, many people have difficulty in expressing themselves in numerical terms, which implies a double process: formulating an opinion and subsequently "converting" it into a numerical value.

Some experiments (among others, Tourangeau, Couper and Conrad 2013) have found that numerical scales with labels at both ends are answered more quickly when they start with the most positive-favourable responses (10, in this case); although they have also warned that this form of display causes the first options on the scale to be chosen more often. More recent studies (Dillman, Smyth and Christian, 2014) have not detected any differences in distribution when scales start with the most positive-favourable or the most negative-unfavourable value, confirming that responses are provided more quickly when scales begin with positive-favourable values.

Given the disparity of results (between these and other studies), in 2013 and 2014 the Association for Public Opinion Research -AAPOR held two meetings on the subject, which concluded that the same scale size and equal numerical and/or verbal labels do not guarantee validity, as the direction of the scale influences the perception of, and the answers provided by respondents. In light of these findings, and considering the differences between the application of the scales in North America (from "totally agree" to "strongly disagree") and in the Netherlands (inversely), Yan and Keush (2015) used a scale for rating country development that was applied in both directions: one from ten to zero (from more to less developed countries) (10-0), and another one from zero to ten (from less to more developed countries) (0-10).

The results obtained showed that higher scores were provided for all countries (higher development) when the scale started with the highest number, concluding that "[...] ratings are more variable for developed countries when the scale starts with 0 and for undeveloped countries when the scale begins with 10" (Yan and Keush, 2015).

Seeking to generalise their findings to other subject areas, a similar study was carried out on two surveys on political issues (Liu and Keusch, 2017; Yan, Keusch and He, 2018). The results showed that the direction of the scale had a greater effect on non-attitudinal questions when they were placed in the second half of the questionnaire; and when the scale was long (Yan, Keusch and He, 2018).

While initial studies have indicated that educational level has little influence on response effects (among others, Schuman and Presser, 1981), later research using the meta-analysis technique has shown that educational level and age greatly influence the effect of order of questionnaire response options (Krosnick, Narayan and Smith, 1996). Educational level is related to the use of cognitive skills. After the age of 65 there is a loss of cognitive abilities, mainly a decrease in memory.

Methodological design: application of international findings to a survey conducted in an autonomous region in Spain

Following the rationale of the first experiment by Yan and Keusch (2015), a telephone survey of households with landlines was used. Using the census as a sampling frame, two equivalent samples of 448 people who were stratified according to the area of residence and type of living environment were selected. The municipalities and the respondents were selected randomly, with quotas of sex and age additionally used for selecting respond-

ents. Some 30 repeated call-backs to non-contacted households, and the use of strategies for conversion of "soft" refusals provided a response rate (TR4) of 64%, very similar to the initial study carried out by Yan and Keusch (2015). The importance of the institution conducting the study, and the brevity of the questionnaire contributed to explaining the high cooperation levels.

The decision to choose a sample of landlines was based on the attempt to replicate Yan & Keusch's study as much as possible. Whereas a sample of landlines clearly poses some representativeness problems, (the Spanish Statistics Institute (2017) estimated that 17.4% of the Navarrese population do not have landlines), the objective of the study was not so much to generalise the results to a given population, but the analysis of the differences between experimental treatments.

It is important to mention two factors specific to telephone surveys here, namely stating a preference for the more extreme responses (Ye *et al*., 2011), and the problems that may arise from the use of channel medium of oral communication, in terms of respondents being able to adequately process the first response alternatives. Providing other options means that the first one is processed at a higher speed, thus increasing the number of choices of the last category (Gwartney, 2007; Díaz-de-Rada, 2010).

The questionnaire had two versions (named A and B), which only differed in the order of presentation of the response options, and were administered to two equivalent samples. The samples were similar with regard to sex, age, level of education, relationship with the activity, household size and living arrangement (Cramer's V values of less than 0.08, with levels of significance considerably higher than 0.10). As it was a questionnaire about political attitudes, it also tested the ideology of respondents, their intention to vote, sympathy towards parties, and voting recall in the last regional election (Cramer V values of 0.096, 0.138, 0.211 and 0.108; levels of significant of 0.759, 0.145, 0.086 and 0.580, respectively).

The fieldwork was carried out by 10 interviewers in September 2017. They had no knowledge of the object of the study and, in order to eliminate their possible influence, they did each questionnaire alternately, so that they all administered the same number of A and B questionnaires. In order to eliminate time bias, each interviewer had to carry out questionnaire A and immediately after questionnaire B, and the session would only end when both had been conducted[1].

Regarding the subject of the study, questions that were "usual" in political research were used: the "typical" rating scale (0-10) of political leaders; self-definition of the respondents regarding political ideology; scale to evaluate four important institutions in the Navarre region; assessment of the regional government and president (see Table 1). These were distributed throughout the questionnaire to avoid biased answers (Alvira, 2011). These are common questions in the "political barometers" used by the Centre for Sociological Research (hereinafter CIS), and have also been frequently used in telephone surveys (among others, CIS, 2017). A recommendation made by Alvira was followed: "when evaluating, political leaders, for example a rating scale between 0 and 10 can be used as an aid ..." (Alvira, 2011: 36). In the "A" questionnaire, the scale ran from 0 to 10, 0 being the "worst rating" and 10 the "best rating" (see Table 1). This was reversed in questionnaire B, which ran from 10 to 0, as used by Schuman and Presser (1981) in their classic study on acquiescence in agree-disagree questions.

[1] In other words, the session could not finish after conducting questionnaire A, leaving questionnaire B for the next day. Both had to be carried out in the same work session.

FIGURE 1. - *Questions used*

Rating of parliamentary speakers (on a scale from 0 to 10)

How do you rate the political performance of ____ [NAME OF EACH PARLIAMENTARY SPEAKER], using a scale ...

A. … from 0 to 10, where 0 means "very poor" and 10 "very good"?

B. … from 10 to 0, where 10 means "very good" and 0 "very poor"?

Respondent's ideological self-identification

The terms left and right are often used to talk about politics. Where would you place yourself on a scale ...

A. … from 0 to 10, where 0 means extreme left and 10 extreme right?

B. … from 0 to 10, where 10 means extreme right and 10 extreme left?

Rating of four important institutions in Navarre

How do you rate the performance of the Parliament of Navarre on a scale ...

A. … from 0 to 10, where 0 means "very poor" and 10 means "very good"

B. … from 10 to 0, where 10 means "very good" and 0 means "very poor"

Same scale is applicable for the Ombudsman of Navarre, the Regional Audit Chamber, and the Government of Navarre's Representative.

Rating of the regional government's and the regional president's management

How would you rate the performance of the President of the Government of Navarre, Uxue Barkos, on a scale ...

A. … from 0 to 10, where 0 means that "very poor" and 10 that "very good"?

B. … from 10 to 0, where 10 means that "very good" and 0 "very poor"?

Same scale is applicable for the regional government.

Source: Developed by the author.

It is important to note that numerous studies have found respondents pay more attention to the number (of the scale) than to the text placed at both ends (among others, Sudman, Bradburn and Schwarz, 1996; Schwarz *et al*., 1991). This influence is greater in telephone surveys, due to the greater difficulty in remembering all the options.

Based on previous studies on the subject (among others, Tourangeau; Couper and Conrad 2013), a hypothesis was proposed, namely that higher average ratings would be obtained on a 10 to 0 scale, as a result of the primacy effect, among other reasons. The second hypothesis held that the narrower the scope of the question, the less likely it would be for these findings to be obtained on scales from 0 to 10. The third hypothesis, based on studies by Dillman, Smyth and Christian (2014), postulated that there is a shorter response time when rating scales start with positive response options.

The mean difference test was used to find the effect produced by the different forms of administration, as similar studies have done in other contexts (among others, Chang and Krosnick 2010).

At the end of the first section it was noted that educational level and age have been found to have a strong effect on response effects. These variables were codified into three and four categories in the present study: basic, secondary and higher education; and aged 16-29, 30-49, 50-64 and more than 65. In multi-category variables it is not possible to use the mean difference test, so a one-way ANOVA test was performed, and the Brown-Forsythe test was used to test for distributions with unequal variance. In order to detect the possible joint influence of the type of questionnaire, sex and education level, a two-way ANOVA test was employed.

Results

The results are structured following the rationale described in Figure 1, namely three sections where there is an analysis of the scale used to rate politicians and the respondent's ideological self-identification; in the second one, there is the assessment of several relevant institutions in the Navarre region; and in the third one, the performance rating of the regional government and president.

Rating of parliamentary speakers and respondent's ideological self-identification on a scale from 1 to 10

Parliamentary speakers were, for the most part, the leaders of the political parties that participated in the last regional election. This was true in all cases except for the governing party, whose spokesperson was proposed to run as a member of the Spanish parliament in the general election of December 2015. It should also be noted that the spokesperson for *Podemos* had the least time in the role, as he had been in office since 3 July 2017.

Before requesting that respondents provided a rating for each leader, they were asked to identify each one with the political party to which they belonged. The results are shown in the second column of Table 1[2]. The spokespersons for the *PSOE* and *UPN* were the most recognised, and the spokesperson for *Podemos* was the least recognised (clearly for the reason mentioned above). It was surprising that the spokesperson for Geroa Bai was widely recognised, as he did not run for the regional election, although this could be explained by his participation in the general election. The respondents who correctly identified the candidates with the relevant

[2] The first column shows the number of respondents who refused to answer the question in each sample. There were 14-16 respondents in each sample, representing only 3.8% of the total, a small size which did not affect the study's results. A detailed analysis of this group reveals that 30 did not rate any of the politicians.

TABLE 1. *Rating of parliamentary speakers and ideological self-position (on an 11-point scale)*

	No response	Level of recognition	Nº of cases	Total average	Average 0-10	Order 10-0	Difference
EH-Bildu	16/14	33.6%	257	4.3	4.4	4.2	0.21
PP	17/14	36.6%	271	3.5	3.4	3.6	-0.28
PSOE	17/14	53.0%	395	4.7	4.6	4.8	-0.18
UPN	16/14	51.3%	388	3.9	3.6	4.2	-0.54*
Geroa Bai	15/14	34.1%	260	5.2	5.3	5.0	0.27
Izquierda Esquerra-IE	16/15	13.0%	100	4.6	4.6	4.6	-0.03
Podemos	16/15	10.0%	67	4.5	4.7	4.3	0.32
Ideology	29/26		792	4.5	4.5	4.5	0

* <0.05.

Note: The leaders are shown in the order they were asked about.
Source: Developed by the author.

party were later asked about their assessment of each leader. This involved a reduction in the sample size, notable in the case of the spokespeople for *IE* and *Podemos*. The low sample size of the latter (38 cases in sample A and 29 in sample B), means that it should be treated with extreme caution.

When using the presentation of scales from 1 to 10 (the most commonly used in surveys), the *Geroa Bai*'s spokesperson was the most highly rated, followed by the spokespersons for *Podemos* and *PSOE-UE*. The presentation of scales from 10 to 0 showed lower ratings for the spokespersons of three of the four parties that governed in Navarre when the study was carried out, namely *Geroa Bai, EH-Bildu* and *Podemos*. Only the rating provided for *IE* did not change. These are small differences, except in the case of the *UPN* spokesperson (almost half a point, the only significant difference). The comparison between the two scale directions showed that the rating of the spokespeople varied. The spokespeople for the *PSOE* and *IE* obtained the second and third best rating.

Once the absence of a scale direction effect was verified, the rest of the variables capable of being influenced were analysed. Education level presented a significant difference in the case of the spokespersons for the *PSOE* (*F* 10.5, *p* 0.000) and the *UPN* (*F* 5.9, *p* .003), and a less significant difference for the *PP* (*F* 3.79, *p* 0.049). The three spokespersons achieved high ratings from the respondents with a low educational level (5.5, 4.7 and 4.4, respectively), a score that fell among respondents with a high educational level (4.3, 3.5 and 3.2, respectively). The same leaders also had different ratings when the age of the respondents was taken into account, which showed that as the age increased, the rating of each leader decreased. This happened up to 64 years old, whereas the trend changed among the oldest respondents, who assigned the highest scores. As it was suspected that the variability within each subsample could be causing the differences between samples, each one was analysed separately, and the same trend was found. However, there was no significant difference when all three variables were considered together (questionnaire order, age and education).

The lower part of the table shows that there was no difference in terms of ideology, as identical averages were found in both samples.

TABLE 2. *Rating of the performance carried out by four Navarran institutions*

	No response	Level of recognition	Nº of cases	Total average	Average 0-10	Order 10-0	Difference
Parliament	7/6	41.9	345	5.2	5.3	5.2	0.05
Ombudsman	2/2	40.7	322	5.6	5.4	5.7	-0.25
Audit Chamber	3/2	38.8	291	6.1	6.2	6.0	0.21
Government Representative	2/2	35.2	291	4.3	4.0	4.6	-0.57*

* <0.05

Source: Developed by the author.

Rating of the performance of four important regional institutions

As indicated in Table 1, the second aspect, to be rated was the performance of the regional Parliament of Navarre (*Parlamento de Navarra*), the regional Ombudsman (*Defensor del Pueblo autonómico*), the regional Audit Chamber (*Cámara de Comptos*) and the regional Government's Representative (Delegado del Gobierno). The questions were administered exactly as in the previous case; a total of 4-5 respondents failed to answer, and a total of 13 provided no response to the question about the Parliament.

The average level of knowledge about the Parliament and the Ombudsman was slightly in excess of 40%, whereas it was somewhat lower for the regional Audit Chamber and the regional Government's Representative (Table 2). Ratings were only provided by those who knew each institution. The Government's Representative received the poorest ratings. However, the ratings of this position also showed the greatest differences, half a point between one scale and another. Better ratings were obtained on the scale from 10 to 0, as proposed in the hypothesis. The differences were lower in the case of the Ombudsman and the Audit Chamber, with differences slightly higher than 0.20 points, which was not significant. It was surprising that the Audit Camber obtained an average score that was higher when using the scale from 0 to 10 than the opposite scale.

The study of the differences by level of education and age only provided significant differences by age in the cases of the Audit Chamber and the Government's Representative. The rating of the Audit Chamber improved as the age of the respondents increased, whereas for the Government's Representative, the highest scores were obtained from the extreme groups: those under 30 and those over 65. There was no difference in the joint analysis of education level, age and direction of response.

Rating of the performance of regional government and president

Data were collected from almost the entire sample regarding the third aspect considered[3], as the regional government and president were sufficiently recognised by most of the interviewees. The difference was the lowest of all those considered (see Table 3). Only age showed a significant relationship in both questions, where older people provided higher ratings.

Other possible influential factors

Little variability was seen in the direction of 0-10 scales. The results were different to

[3] With the exception of 16 and 23 interviewees who did not answer both questions, 9 and 13 respectively in sample A; 7 and 10 in sample B.

TABLE 3. *Rating of the Regional Government's and president's performance*

	No response	Nº of cases	Total average	Average 0-10	Order 10-0	Difference
Government performance	9/7	824	4.8	4.9	4.8	0.07
President performance	13/10	801	4.9	4.9	4.8	0.09

Source: Developed by the author.

those obtained by Keusch, which leads to the need to verify to what extent those numerical scales with labels at the ends are answered more quickly when they start with the most positive response options. These results were also shown by Dillman's team (Dillman, Smyth and Christian, 2014).

It should be taken into account that the scales used in the previous sections included only 14 of the 80 variables included in the questionnaire. Therefore, it was decided to compare the ordinal questions from the type-A questionnaire (which began with positive/favourable response options: "very good", "highly favourable", etc.) with those of questionnaire B, in which questions began with negative/unfavourable response options: "very poor", "highly unfavourable", etc. According to the approaches included in the previous paragraph, questionnaire A should have a shorter duration than questionnaire B. The second hypothesis, which postulated that differences decrease when the scope of the question is narrower, needed to consider these questions separately, most of them using scales of 4-5 categories.

Questionnaire A was answered in an average of 11.9 minutes, and B required 12.01. This is a non-significant difference, which confirmed that the direction of the responses had no effect.

Conclusions

There was hardly any difference in the scales from 0 to 10, regardless of whether they were used in one direction or another. Only two items, out of the 14 used, presented differences related to the direction of the scales. The effect of age and educational level were found to be greater. The effects of these variables may have concealed other factors, such as the *recency effect* (characteristic of telephone surveys), the different levels of political culture, and the *strength* of the political *attitudes* held (Bassili and Krosnick, 2000). The first interpretation involves a greater choice of the last response categories among the respondents with low educational levels, which was not seen when each subsample was analysed separately. Nor was an influence of a greater or lower political cultural level seen. The verification of the third interpretative avenue requires an analysis of respondents' interest in politics, as the most interested respondents should present greater differences in their political ratings, something that was not found. The influence of educational level and age was lower in each subsample than when both were compared, although the differences were not significant.

The differences were smaller for the rest of the questions (ordinal questions containing 4 and 5 categories). Finally, regarding the time of administration, the duration was similar for both questionnaires.

At this point the reader may feel discouraged about the results. Especially when considering the large number of studies that pointed in a different direction, as discussed in the first part of the paper. However, I feel the opposite. The stability of the findings confirmed the suitability of the survey tool for the intended context, and "validates" those research findings that used scales ordered in an upward direction (perhaps out of habit).

The absence of the effect indicated in the first section can be explained by taking into account the cultural differences between the country where such effects were detected and the location where this study was carried out.

All studies have limitations, and the main one here is related to the generalisation of the findings, as it was located in a single autonomous region in Spain. It is a common limitation of "survey experiments" that use equivalent samples, many of which are conducted on captive samples such as clients, university students, etc. While several experts have shown that these samples can be generalised,[4] the findings presented here require further research on the subject. Another factor that can help explain these results is the higher level of training of the personnel who carried out the fieldwork, the exhaustive monitoring and the persistence to ensure that the questionnaire was read exactly as it was written, something that was also verified but does not always take place. This area certainly requires further research.

Bibliography

Alvira, Francisco (2011). *La encuesta: una perspectiva general metodológica*. Madrid: CIS.

Bassili, John N. and Krosnick, Jon A. (2000). "Do Strength-related Attitude Properties Determine Susceptibility to Response Effects?". *Political Psychology*, 21(1): 107-132.

Bruce, Darly and Papay, James P. (1970). "Primacy Effects in Single Trial Free Recall". *Journal of Verbal Learning and Verbal Behaviour*, 9: 473-486.

Centro de Investigaciones Sociológicas (2017). *Preelectoral de Cataluña. Elecciones autonómicas 2017*, study 3195. Available at: http://www.analisis.cis.es/cisdb.jsp

Chang, Linchat and Krosnick, Jon A. (2010): "Comparing Oral Interviewing with Self-administered Computerized Questionnaires: An Experiment". *Public Opinion Quarterly*, 74: 154-167.

Díaz de Rada, V. (2010). *Comparación entre los resultados obtenidos por encuestas personales y telefónicas*. Madrid: CIS.

Dillman, Don; Smyth,Jolene D.; Melani Christian, Leah; y Stern, Michael J. (2005). *Comparing check-all and forced-choice question formats in web survey.* Pullman, Washington: SESRO.

Dillman, Don A.; Smyth, Jolene D. and Christian, Leah Melani (2014). *Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method (4th ed.)*. New York: Wiley.

Gwartney, Patricia (2007). *The Telephone Interviewer's Handbook: How to Conduct Standardized Conversations*. San Francisco: Jossey-Bass.

Instituto Nacional de Estadística (2017). *Encuesta de Tecnologías de la Información en los Hogares*. Madrid: INE.

Kamoen, Naomi; Holleman, Bregje; Mak, Pim y Sanders, Ted J.M. (2011). "Agree or Disagree? Cognitive Processes". *Discourse Processes*, 48(5): 355-385.

Krosnick, Jon A. and Alwin, Duane F. (1987). "An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement". *Public Opinion Quarterly*, 51: 201-219.

Krosnick, Jon A. and Presser, Stanley (2010). "Question and Questionnaire Design". En: Marsden, P. V. y Wright, J. D. (eds.). *Handbook of Survey Research*. Bingley: Emerald.

Krosnick, Jon A.; Narayan, Sowmya and Smith, Wendy R. (1996). "Satisficing in Surveys: Initial Evidence". *New Directions for Evaluation*, 70: 29-44.

Liu, Mingnan and Keusch, Florian (2017). "Effects of Scale Direction on Response Style of Ordinal Rating Scales". *Journal of Official Statistics*, 33: 137-154.

Revilla, Melanie A.; Saris, Willem E. and Krosnick, Jon A. (2014). "Choosing the Number of Categories in Agree-Disagree Scales". *Sociological Methods and Research*, 43: 73-97.

Schuman, Howard and Presser, Stanley (1981). *Questions and Answers in Attitude Surveys*. New York: Academic.

Schwarz, Norbert; Knäuper,Bärbel; Hippler, Hans-J.; Noelle-Neumann, Elisabeth y Clark, Leslie (1991).

[4] See, for example, Dillman *et al.* studies (2005), among others, conducted with students at the University of Washington, whose findings have been generalisable to national samples.

"Numeric Values May Change the Meaning of Scale Labels". *Public Opinion Quarterly*, 55: 570-582.

Simon, Hebert A. (1957). *Models of Man*. New York: Wiley.

Sudman, Seymour; Bradburn, Norman y Schwarz, Norbert (1996). *Thinking about Answers*. San Francisco: Jossey-Bass.

Tourangeau, Roger and Rasinski, Kennet (1988). "Cognitive Processes Underlying Context Effects in Attitude Measurement". *Psychological Bulletin*, 103: 299-314.

Tourangeau, Roger; Couper, Mick P. and Conrad, Frederick (2013). "Up Means Good: The Effect of Screen Position on Evaluative Ratings in Web Surveys". *Public Opinion Quarterly*, 77: 69-88.

Yan, Tin and Keusch, Florian (2015). "The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey". *Public Opinion Quarterly*, 79: 145-165.

Yan, Tin; Keusch, Florian and He, Lirui (2018). "The Impacts of Question and Scale Characteristics on Scale Direction Effects". *Survey Practice*, 11(2).

Ye, Cong; Fulton, Jenna and Tourangeau, Roger (2011). "More Positive o more Extreme?". *Public Opinion Quarterly*, 72: 349-365.

RECEPTION: April 9, 2018
REVIEW: March 25, 2019
ACCEPTANCE: January 18, 2019