

# Balancing Act: Effects of Scale Polarity on Measurement

# Randall K. Thomas, Frances M. Barlas, & Larry Osborn



Taking Survey and Public Opinion Research to New Heights!

73rd Annual Conference

May 16 - 19, 2018

DENVER, COLORADO



We often need to make choices about how to measure ideas in questionnaires. One major choice is the response scale. Among the choices we have:

- Should the scales be bipolar or unipolar?
- Should we anchor responses with labels?
- How many responses should we use?



#### Should the scales be bipolar or unipolar?

- The response format could represent a bipolar scale (ranging from a conceptual pole and its antithesis, e.g., like – dislike, with a natural neutral response)
- Or it could be a unipolar scale (with a concept and its absence, e.g., like do not like and no neutral response).

			W		<b>V</b>
5 Category Bipolar	ar Strongly dislike Somewha		Neutral	Strongly like	
5 Category Unipolar	Do not like	Somewhat like	Like	Strongly like	Very strongly like
	Î				1



Many researchers prefer what is called a balanced bipolar scale where there are an equal number of responses on each side of the neutral point.

The idea is that if we want to adequately map out a concept, we need to present both "opposites" of the concept so that people have the opportunity to pick in favor of or opposed to the concept and it should have a midpoint (whether explicit or implicit in the case of even-numbered responses).

# **Introduction – Balanced Bipolar Scales**

#### **Examples of balanced bipolar scales:**

Strongly	Somewhat	Neutral	Somewhat	Strongly
Dislike	Dislike		Like	Like
Strongly	Somewhat	Neutral	Somewhat	Strongly
Disagree	Disagree		Agree	Agree
Very unimportant	Somewhat important	Neither important nor unimportant	Somewhat important	Very important

# Introduction – Unbalanced Bipolar Scales



#### Example of a bipolar scale that is not balanced (asymmetric bipolar):



# **Introduction – Mixed Scale Properties**



Example of a terrible attempt at a bipolar scale that is neither bipolar nor unipolar, which is associated with a lot of respondent confusion:







## Should the scales be fully- or end-anchored?

Another question that arises is whether all responses should have labels for each response (fully anchored) or do responses only need to have the end categories with labels (end anchored only).

Prior research has indicated that fully-anchored scales may have somewhat higher validity than end-anchored scales (Krosnick, 1999).



Further, end-anchored scales may be more susceptible to response bias – people respond more to the endpoints than the dimension of judgment (Schimmack, Böckenholt, & Reisenzien, 2002).

End-anchored scales are used frequently for telephone surveys ("On a scale of 0 to 10 where '0' means 'Dislike very strongly' to '10' means 'Like very strongly' how would you rate X?").

Cross-country studies often employ end-anchored scales to make it easier for translators (only the ends to translate).



# How many responses should we use?

A number of studies have indicated that using more response categories can lead to more reliable and valid results. This is true up to the point that people can reliably distinguish between the meanings of the categories.

In earlier research, Thomas, Krosnick, and Anand (2008) found that 4 to 5 response categories led to the optimal level of validity, at least for bipolar scales.



# **Empirical Characteristics of Scales**

# Study 1



As part of a larger international web-based study, we had 44,757 respondents from a non-probability sample complete a section which asked how much respondents liked doing a variety of activities.

Belgium - Flemish	4017
Belgium - French	2865
France	17995
Germany	4159
Italy	9419
Spain	4268
UK	2034
Total	44757

# Study 1



Respondents were then asked how much they liked to do 10 different behaviors (from drinking coffee to shopping for clothes). They were randomly assigned to 1 of 4 scale conditions, each with 5 response categories:

- Polarity:
  - Unipolar (Do not like at all-Like very much)
  - Bipolar (Dislike very much-Like very much)
- Anchoring
  - End-anchored
  - Fully-anchored





# A fairly typical response distribution for a 5 category bipolar v. unipolar and fully anchored v. end anchored.



Average Response Endorsement by Scale





# Which one has the HIGHEST top 2 box score?







### Which one has the HIGHEST top 2 box score?







# Which one has the LOWEST top 2 box score?







#### Which two scales have labels that made no difference?



# **Study 1 - Results**



# Looking at the criterion-related validity – the Unipolar Fully-Anchored scale explained an average of 4% more variance than the Bipolar Fully-Anchored scale.

**Criterion-related Validity** 



# Study 2



As part of a U.S. web-based study, we had 17,405 respondents who completed a section which asked how much respondents liked a variety of 5 different beverages (e.g., carbonated soft drinks, iced tea, etc.). The scale used was a 5 category response format.

# **Study 2 - Results**



# Different country, different study, but we see similar response patterns for the scales.



# **Study 2 - Results**



# Similar to Study 1 – the Unipolar Fully Anchored scale explained an average of 5% more variance than the Bipolar Fully Anchored scale.

**Criterion-related Validity** 



Bipolar Fully Anchored Unipolar Fully Anchored Bipolar End Anchored Unipolar End Anchored



# What about number of response categories?

# **Scales and Smartphones**





With the increased use of smartphones – even a 5 category fully-anchored response format (a relatively short scale) is often not be feasible without significant leftright scrolling.

The goal of the next study was to explore the impact of using fewer response categories.

# Study 4 Method



This study also used KnowledgePanel sample, but investigated political attitudes and used agreement scales. Similar to Study 3, we had the following conditions:

- 3, 4, or 5 response categories
- Unipolar or bipolar fully-anchored scale

# **Study 4 – Political Issues**



Next we would like to find out what you think about a number of issues. How much do you agree with the following?

- A. The national government should spend more to take care of its poorest citizens.
- B. It is important to reduce the gap in wealth between the rich and poor in this country.
- C. The national government should be allowed to borrow as much money as it thinks is necessary.
- D. The national government should have strict limits placed on its spending.
- E. Our leaders should expand the power of the United States.
- F. Most national taxes should be greatly reduced.
- G. Free trade among nations is good for this country.
- H. A person who plans a murder and carries it out should be put to death.
- I. Illegal immigrants represent a major threat to this country.
- J. Religious faith can provide useful solutions for the problems that our country will face.
- K. This nation would improve if the people were to become more religious.
- L. Global warming is occurring because of human activity.

# **Study 4 – Responses Used**



#### **Respondents were randomly assigned 1 of the following response formats:**

3 Category Unipolar	Do not agree	Agree	Very strongly agree		
4 Category Unipolar	Do not agree	Agree	Strongly agree	Very strongly agree	
5 Category Unipolar	Do not agree	Somewhat agree	Agree	Strongly agree	Very strongly agree
3 Category Bipolar	Disagree	Neutral	Agree		
4 Category Bipolar	Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	
5 Category Bipolar	Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree

36

#### Means again were higher with bipolar scales.







# **Experiment 2 - Average Validity**



No significant differences in validity (predicting party ID) for anchoring or number of response categories.



Average Validity - r<sup>2</sup>

# **Experiment 2 - Average Validity**



No significant differences in validity (predicting party ID) for anchoring or number of response categories.



Average Validity - r<sup>2</sup>

# **Recommendations for Scale Design**



We have been finding that response formats with fewer response categories take less time to complete but also can show as much validity as formats with more response categories, especially when using the unipolar variant.

**Conclusions:** 4 responses in a unipolar fully-anchored format have been associated with:

- Optimum levels of validity
- Differentiation capability similar to bipolar scales with 7 categories
- Still working well with the 'thumbs' environment of the smartphone.



# Thank you!

#### **Randall K. Thomas**

Frances M. Barlas

Randall.Thomas@gfk.com

Frances.Barlas@gfk.com



# New Scales for the Modern Survey Era: Numeric Scales

# **Randall K. Thomas and Frances M. Barlas**



Taking Survey and Public Opinion Research to New Heights!

73rd Annual Conference

May 16 - 19, 2018

DENVER, COLORADO



Rating scales are the most common methodology we use in survey research – designed to help us, as researchers, systematically understand what people think, feel, and want.

We often provide semantic response labels with our scales that are intended to help respondents in selecting their responses along a dimension of judgment.

How much do you like football?

Hate	Dislike	Neutral	Like	Love
------	---------	---------	------	------



**Compare these 11 category scales:** 

- Using a scale from -5 to +5 where -5 means "Strongly dislike" and +5 means "Strongly like" how much do you like football?
- Using a scale from 0 to 10 where 0 means "Strongly dislike" and 10 means "Strongly like" how much do you like football?



Schwarz and colleagues (1991; 1995) found that the use of numbers in response labels can shift the meaning of scales – compared to those given a 0 to 10 scale, those given -5 to +5 chose the top 4 responses more often.

#### **Conclusion:**

The labels associated with the responses can significantly affect response distributions, and the numbers that are used are one part of how we interpret the meaning of the responses.



# In contrast to the luxury we used to have, 11 category scales are a thing of the past.

As more and more people take online surveys with their smartphones, the search is on for scales with fewer and smaller labels that still provide meaningful differentiation and can be used easily by respondents.

Since most smartphone respondents take online surveys in the portrait orientation, even the horizontal presentation of just five categories with full semantic labels can extend off the screen.





We considered an alternative that might work better on smartphone screens – clickable cells:

How much do you like or dislike drinking coffee?

Strongly Dislike	Somewhat Dislike	Somewhat Like	Strongly Like	-2	-1	+1	+2
---------------------	---------------------	------------------	------------------	----	----	----	----

Prior research tested numeric labeling also provided semantic end labels (Schwarz et al., 1991).



Semantic labels can take up more space on a screen than just using numbers. We wanted to see how respondents used numeric responses when these semantic definitions were eliminated to reduce the space that responses occupied. The only clues that the respondent had that could be used to discern the meaning of the numbers was: 1) the item stem and 2) the positive/negative sign preceding the number.

**Research Question:** Do respondents define the meaning and distances of the numeric responses in a similar way as they do for semantically labeled responses.



# Method





- Joint study by GfK and the Advertising Research Foundation (ARF)
- 3,584 respondents completed this experiment all from non-probability sample providers
- Randomly assigned respondents to either:
  - o banked format or grid format
  - semantically labeled or numerically labeled scales
  - bipolar scale or unipolar scale
  - **3, 4, or 5 category response format**





#### **Unipolar** item stem:

#### How much do you like doing each of the following?

Unipolar	Semantic	3 category	Do not like	Like	Strongly like		
		4 category	Do not like	Like	Strongly like	Very strongly like	
		5 category	Do not like	Somewhat like	Like	Strongly like	Very strongly like
	Numeric	3 category	0	+1	+2		
		4 category	0	+1	+2	+3	
		5 category	0	+1	+2	+3	+4





#### **Bipolar** item stem:

#### How much do you like or dislike doing each of the following?

Bipolar	Semantic	3 category	Dislike	Neutral	Like		
		4 category	Strongly dislike	Somewhat dislike	Somewhat like	Strongly like	
		5 category	Strongly dislike	Somewhat dislike	Neutral	Somewhat like	Strongly like
	Numeric	3 category	-1	0	+1		
		4 category	-2	-1	+1	+2	
		5 category	-2	-1	0	+1	+2



# Method

#### Items used:

- A. Drink coffee
- B. Shop for clothes for myself
- C. Read scientific articles
- D. Play video games.
- E. Drink tea
- F. Eat beef
- G. Attend religious services.
- H. Eat chocolate
- I. Smoke cigarettes
- J. Exercise vigorously for at least 15 minutes at a time



# **Results**

# **Results – Time to Complete**





#### 1 0.8 0.6 0.4 0.2 0 Unipolar Unipolar **Bipolar Bipolar** Semantic Numeric

# **Results – Mean Values of Items**

# **Results - Validity**



- Next, to examine validity of the different formats, we looked at how the ratings of liking predicted self-reported behavior (number of days in the past 30 days).
- Since the behavioral variable can be skewed with the 0 to 30 days, we grouped responses (0, 1-4, 5-10, 11+), and performed regression analyses. The next screen reports the average adjusted r<sup>2</sup> across the 10 items.

# **Results – Average Validity**



# **Results – Self-ratings of Format**





#### © GfK Custom Research 2018



# **Discussion**

# **Conclusions**



- The numeric scales took less time for respondents to answer.
- This was especially reflected in the results for the 4 and 5 category response formats where semantic took much longer.
- Unipolar scales had lower means than bipolar scales for both semantic and numeric formats.
- As we have consistently found in previous research, the validity of the unipolar scales was higher than the bipolar scales – for both the semantic and numeric scales.

# **Discussion**



- Overall, presenting numeric responses without semantic labels and without respondent instructions yielded reliable and valid results and more quickly than semantic labels.
- This type of format has the possibility of replacing semantic labels in some contexts, being:
  - easier to use on smartphones
  - reducing respondent burden
  - easier to implement across countries translations of numbers are not normally needed.

# **Discussion - Limitations**



- We believe that numeric responses may work for many of the evaluative and intensity measures we commonly use in surveys.
- Numeric response labels may not work for some types of scales (such as frequency measurement).
- While promising, we believe that these results need to have extensive replication to help identify under which conditions they will produce valid and reliable results.
- We are fielding additional studies to further explore the range of applicability of numeric responses.



# Thank you!

#### **Randall K. Thomas**

Frances M. Barlas

Randall.Thomas@gfk.com

Frances.Barlas@gfk.com



# Direction of Agree-Disagree Rating Scales and Data Quality

Ting Yan, Westat Florian Keusch, University of Mannheim

**AAPOR 2018** Taking Survey and Public Opinion Research to New Heights

# **Pros and Cons of Agree-Disagree Scales**

**Pros** (Krosnick, 2012)

- Ease of administration
  - Rs are about 2/3 faster to this format than other formats
  - Fewer "don't know" than yes/no format
  - Rs prefer this format

**Cons** (Converse and Presser, 1981; Krosnick, 2012; Holbrook 2013)

- Cognitively difficult
  - Dimension<>scale
- Subject to "acquiescence" bias
- Subject to straightlining
- Multi-barreled
- Disagree end is ambiguous
  - "I am seldom depressed."



# **Direction of Agree-Disagree Scales**

- A practical question:
  - Shall we start with "agree" or "disagree?"
    - What do you think?
- General guidelines:
  - Starting with the least desirable (Sudman and Bradburn, 1982)
  - Usually better to list from lower to higher (Bradburn, Sudman, and Wansink, 2004)
  - Logical progression from top to bottom (Tourangeau et al., 2004)
- Culture norm or industry standards
  - US: agree->disagree
  - Netherlands: disagree->agree (Hofmans et al. 2007)

# This talk

- Does scale direction affect...
  - Acquiescence?
  - Extreme response style (ERS)?
  - Mid-point response style (MRS)?
  - Straightlining?
  - Internal consistency?



# Conclusions

- Scale direction only affects level of ERS for overall and for web
  - Higher level of ERS when scale starts with 'agree' side
- Scale direction does NOT affect
  - Acquiescence
  - MRS
  - Straighlining
  - Reliability
  - ERS for face-to-face

