Evaluating Qualifiers in Rating Scales

Thursday 4:00 PM – 5:30 PM

July 18, 2019

Room D22

Morgan Earp

Jean Fox

Robin Kaplan

Overview

- Background
- Motivation
- MTurk Study
- Case Studies
- Conclusions
- Limitations

Background

- We often use surveys to collect data on things like attitudes, experiences, and expectations using rating scales.
  - Can collect data from a lot of people in a systematic way
- Lots of research about writing good survey questions
  - It's easy to write surveys, but hard to write good surveys.
- One of the many challenges is deciding on the response options for rating scales.

Selecting Rating Scale Options

- You want the options to:
  - Be appropriate conversational answers to the question asked
  - Cover the full range of situations
  - Be equally distributed across the full range of the construct

- Our research explores if and when varying response options cover the full scale, as well as how the response options are distributed

Definitions

- Qualifiers in scales
  - Strength/Intensity (e.g., Not at all, Somewhat, Very)
  - Frequency (e.g., Never, Sometimes, Often)
  - Evaluation (e.g., Bad, Good, Great)

- Bi-polar vs unipolar
  - Focusing on unipolar here

Motivation

- Explore the “quantity” that commonly used qualifiers represent
- Explore the relative values of closely related qualifiers to understand how they compare to one another

MTurk Study

Participants (*N* = 355)

- Online study with participants from MTurk
- Mean age = 35.2 (SD = 10.7)
- Education:
  - High school: 14.8%
  - Some college: 19.5%
  - Associate's/Bachelor's: 57.9%
  - Graduate degree: 7.8%
- Gender
  - 59.3% Male, 40.4% female

Slider Task

- Participants rated on scales from 0 to 100 “how much” each of the terms meant
  - 15 Quality terms (e.g., Excellent, Good, Average, Poor)
  - 18 Amount terms (e.g., Completely, Very, Moderately, A little)
  - 22 Frequency terms (e.g., Often, Frequently, Occasionally, Rarely)
- Terms were presented in randomized order
- Selected commonly used terms for task

Example

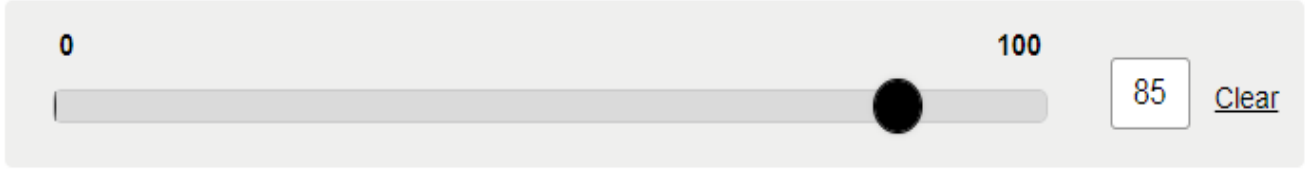
How much do each of the terms represent?

Using the sliders below, please assign a value to each of the terms, where:

- **0 means "the least you can imagine," and**
- **100 means "the most you can imagine."**

Very

Strongly



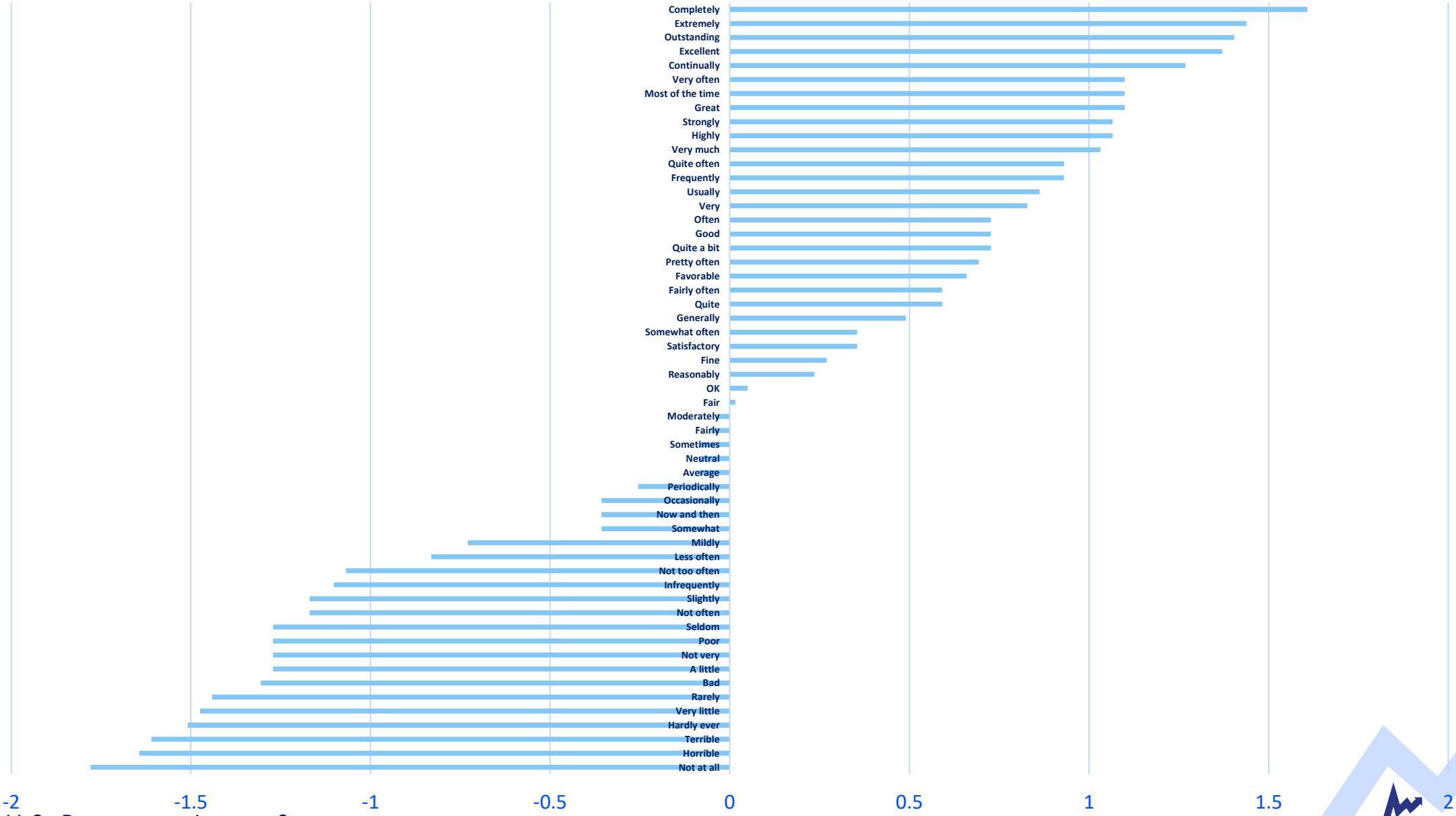
Comparing Quantifiers

Descriptive Statistics						
	N	Median	Std. Deviation			Z value
Not at all	393	0.00	25.991	Grand Mean	52.54	-1.78
Horrible	393	4.00	23.374	Std. Dev	29.53	-1.64
Terrible	393	5.00	22.730			-1.61
Hardly ever	393	8.00	21.287			-1.51
Rarely	393	10.00	20.166			-1.44
Very little	396	9.00	21.473			-1.47
Bad	394	14.00	20.388			-1.31
Not very	397	15.00	20.347			-1.27
Poor	393	15.00	19.361			-1.27
Seldom	391	15.00	19.885			-1.27
A little	395	15.00	20.630			-1.27

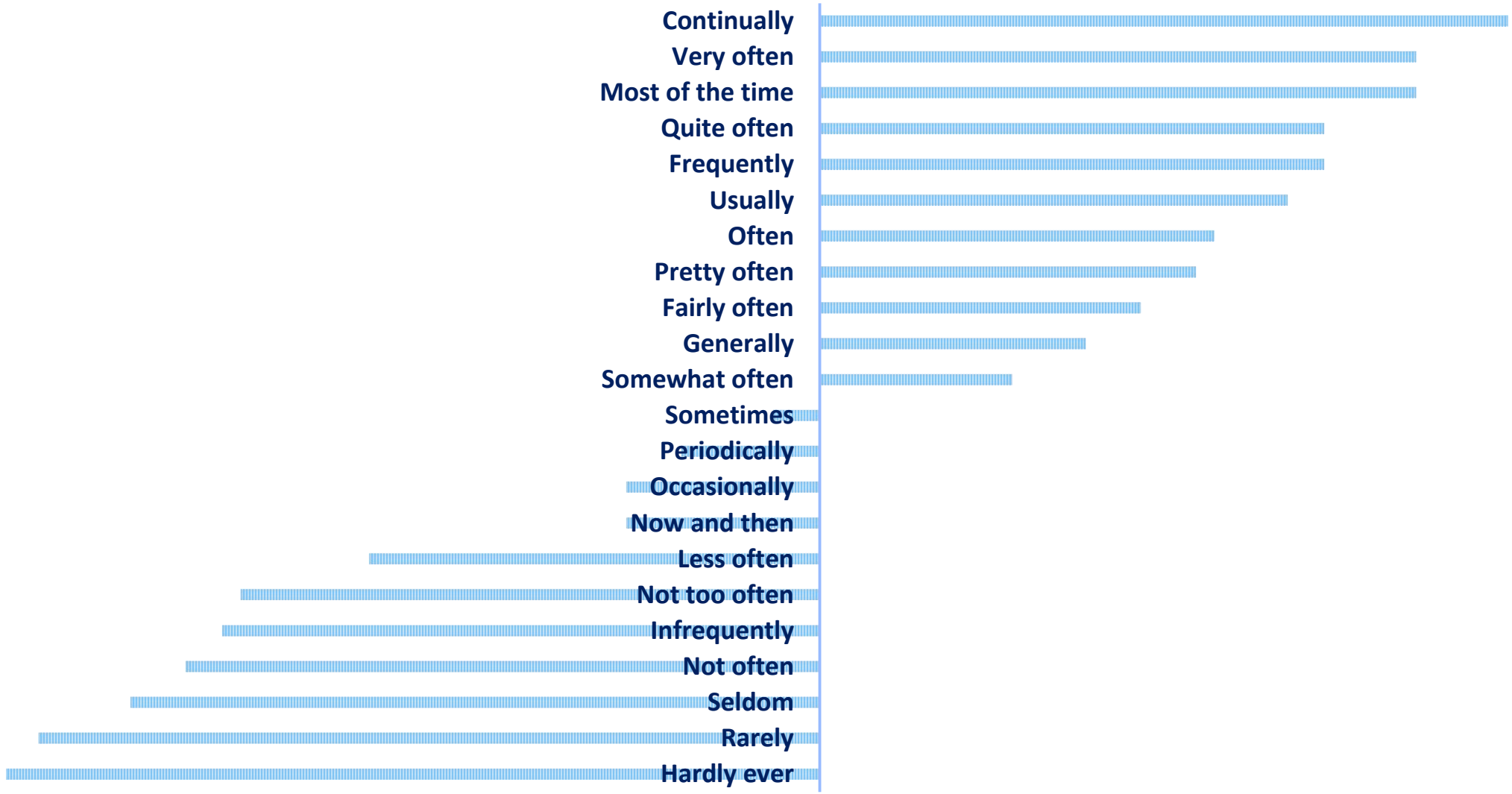
...

Continually	391	90.00	20.424			1.27
Excellent	394	93.00	14.577			1.37
Outstanding	394	94.00	16.468			1.40
Extremely	397	95.00	18.019			1.44
Completely	396	100.00	18.104			1.61

All Terms

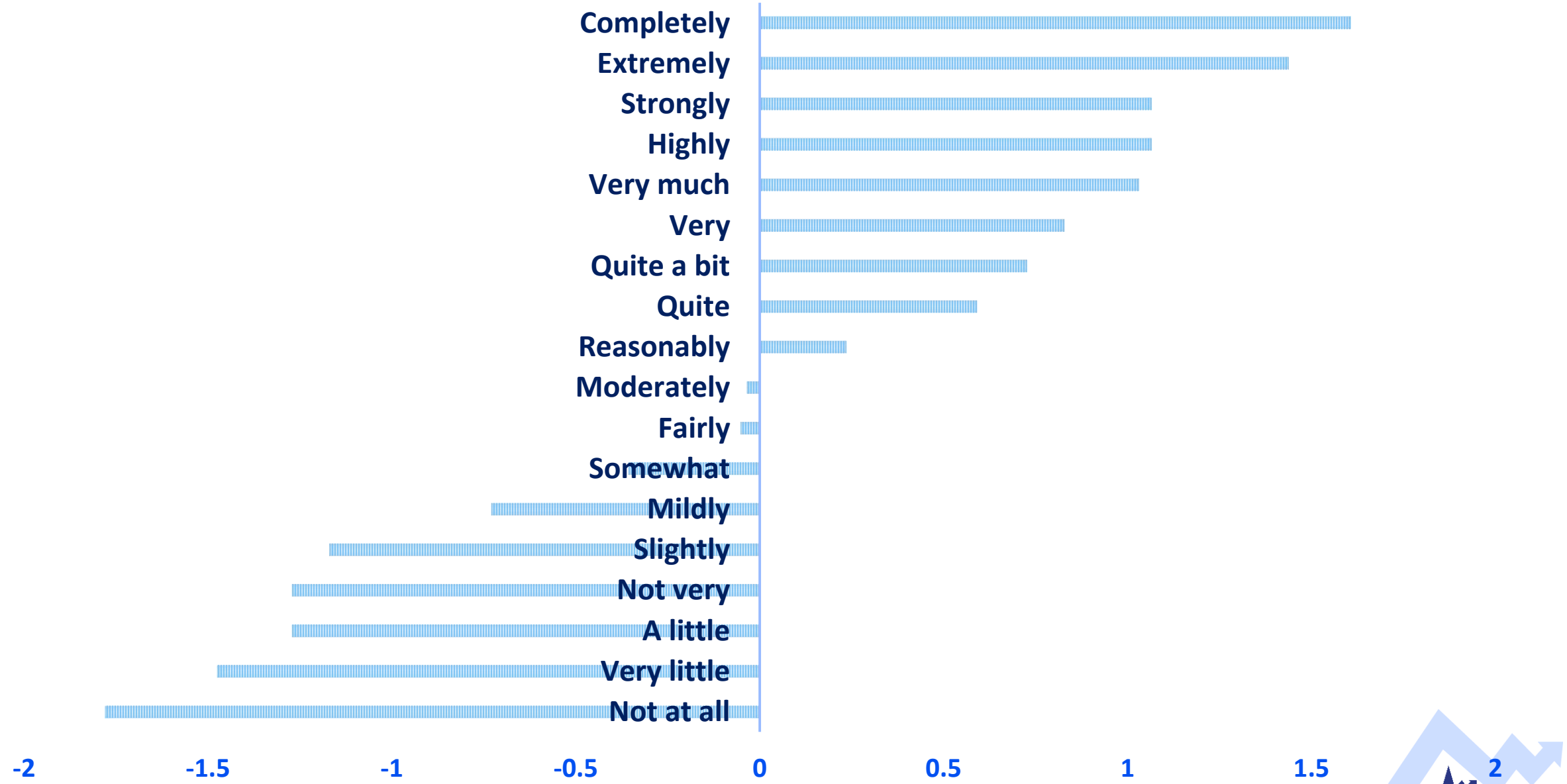

Completely
Extremely
Outstanding
Excellent
Continually
Very often
Most of the time
Great
Strongly
Highly
Very much
Quite often
Frequently
Usually
Very
Often
Good
Quite a bit
Pretty often
Favorable
Fairly often
Quite
Generally
Somewhat often
Satisfactory
Fine
Reasonably
OK
Fair
Moderately
Fairly
Sometimes
Neutral
Average
Periodically
Occasionally
Now and then
Somewhat
Mildly
Less often
Not too often
Infrequently
Slightly
Not often
Seldom
Poor
Not very
A little
Bad
Rarely
Very little
Hardly ever
Terrible
Horrible
Not at all
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Frequency Terms



Quality Terms

Amount terms



Paired Comparisons

- Selected similar terms and asked participants to select the one that suggests “more” of that construct
  - 14 Quality pairs (e.g., Excellent vs. Outstanding)
  - 19 Amount pairs (e.g., Completely vs. Extremely)
  - 17 Frequency pairs (e.g., Often vs. Usually)
- Presented one at a time, grouped by construct

Example

Next, you will look at pairs of words that represent different amounts:

- Your task is to select the word that you think suggests more, or a greater quantity.
- Some of the pairs of words may be very close in meaning.
- Please do your best to determine which of the words suggests more, or a greater quantity.

REMEMBER: go with your initial impression.

Which word suggests more, or a greater quantity?

- ○ Completely
- ○ Extremely

Case Studies

Case Studies

- We solicited previous internal studies that might have useful data using a variety of response scales
  - Needed enough responses
  - Wanted unipolar data only
  - Items had to have good item fit in relation to the construct they were specified to measure

Case Studies

- We found 7 studies with data we could use as case studies
- Measured 10 constructs
  - Burden
  - Concern
  - Confidence
  - Frequency
  - Importance
  - Likelihood
  - Persuasiveness
  - Sensitivity
  - Trust
  - Usefulness

Case Studies

- We found 7 studies with data we could use
- Measured 10 constructs using multiple scales
  - Burden
  - Concern
  - Confidence
  - Frequency
  - Importance
  - Likelihood
  - Persuasiveness
  - Sensitivity
  - Trust
  - Usefulness

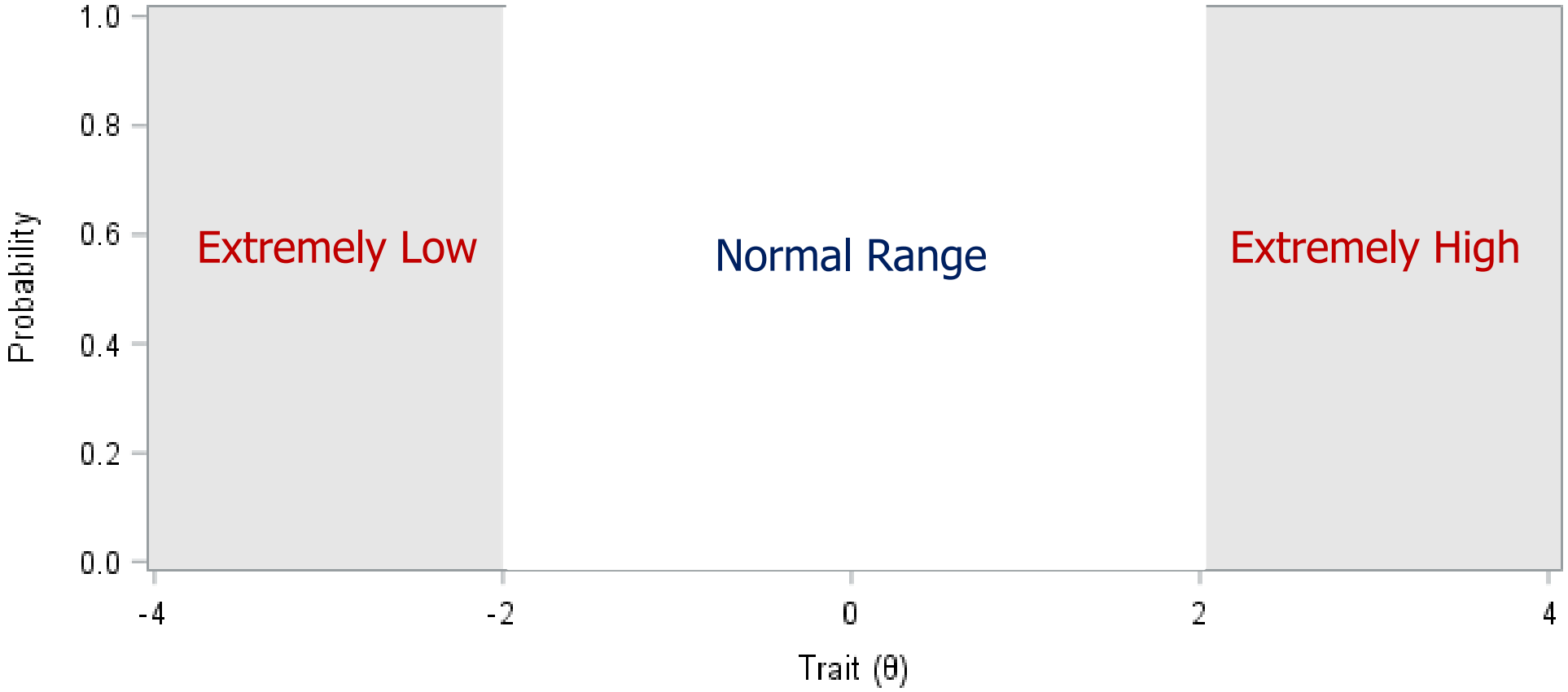
Case Study Response Scales

- Case Study 1 *Persuasive* (n=…)
  - Not at all
  - A little
  - Somewhat
  - <no qualifier>
    "Persuasive"
  - Very
- Case Study 2 *Concern* (n=…)
  - Not at all
  - A little
  - Moderately
  - Very
  - Extremely
- Case Study 3a *Burden* (n=…)
  - Not at all
  - A little
  - Moderately
  - Very
  - Extremely
- Case Study 3b *Burden* (n=…)
  - Not at all
  - Somewhat
  - Moderately
  - Very
  - Extremely

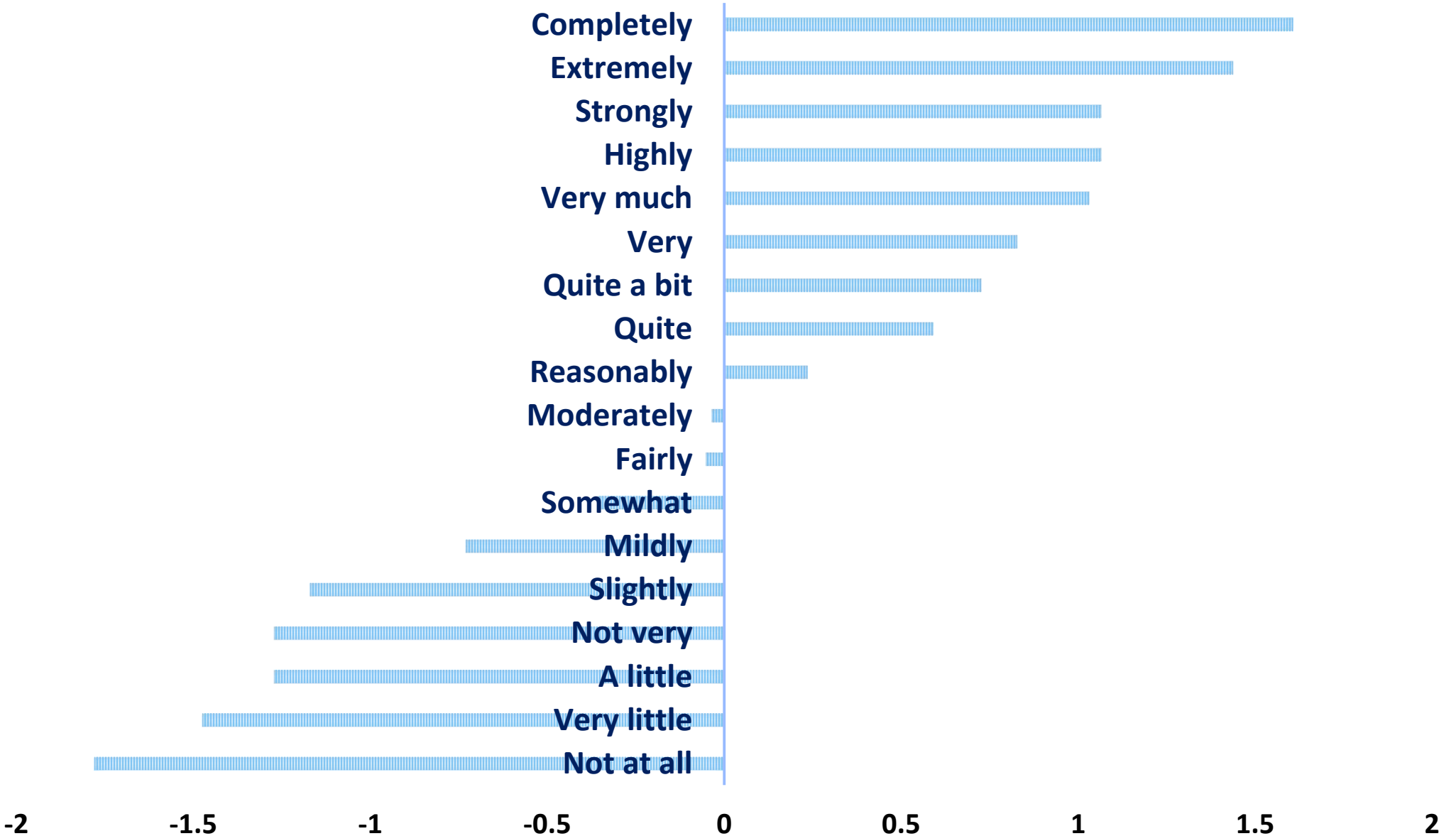
Case Study Response Scales

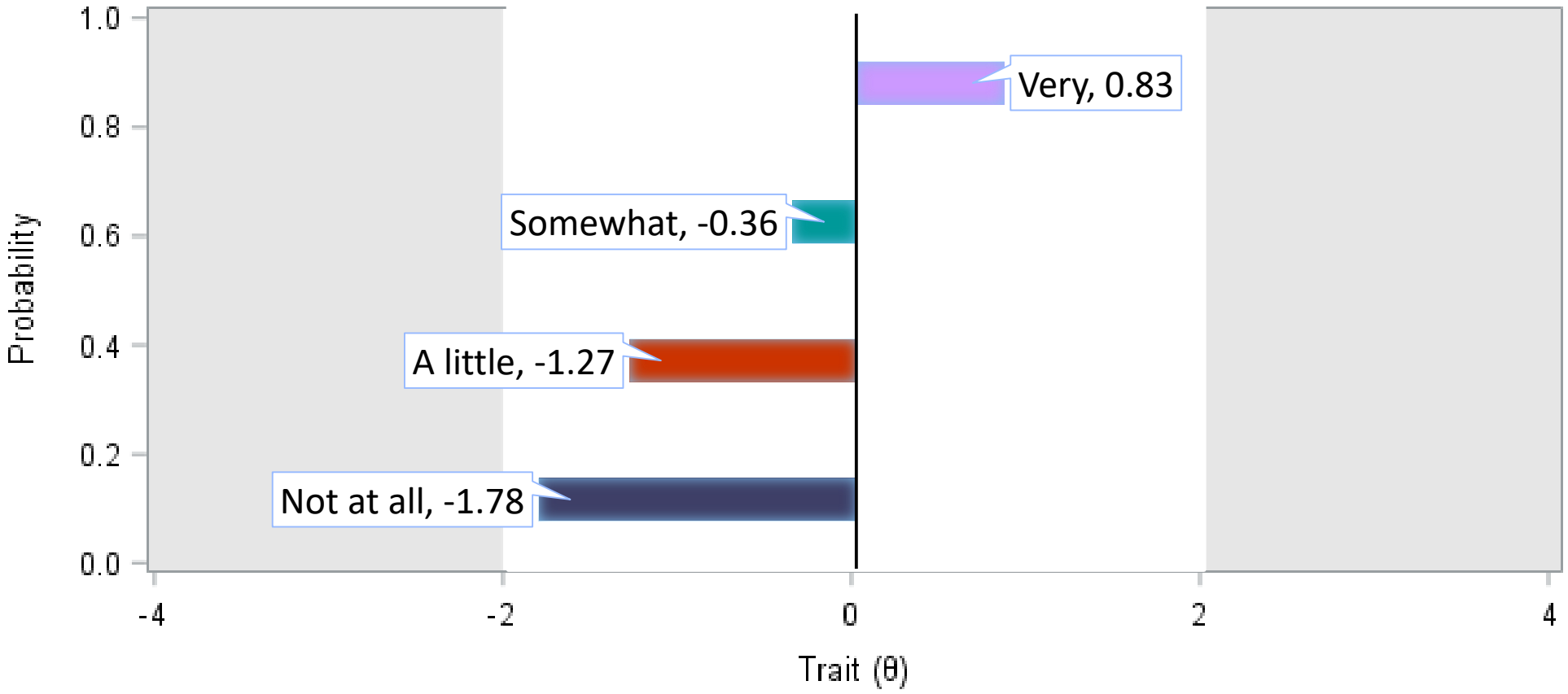
- Case Study 1 (Persuasive)
  - Not at all
  - A little
  - Somewhat
  - <no qualifier>

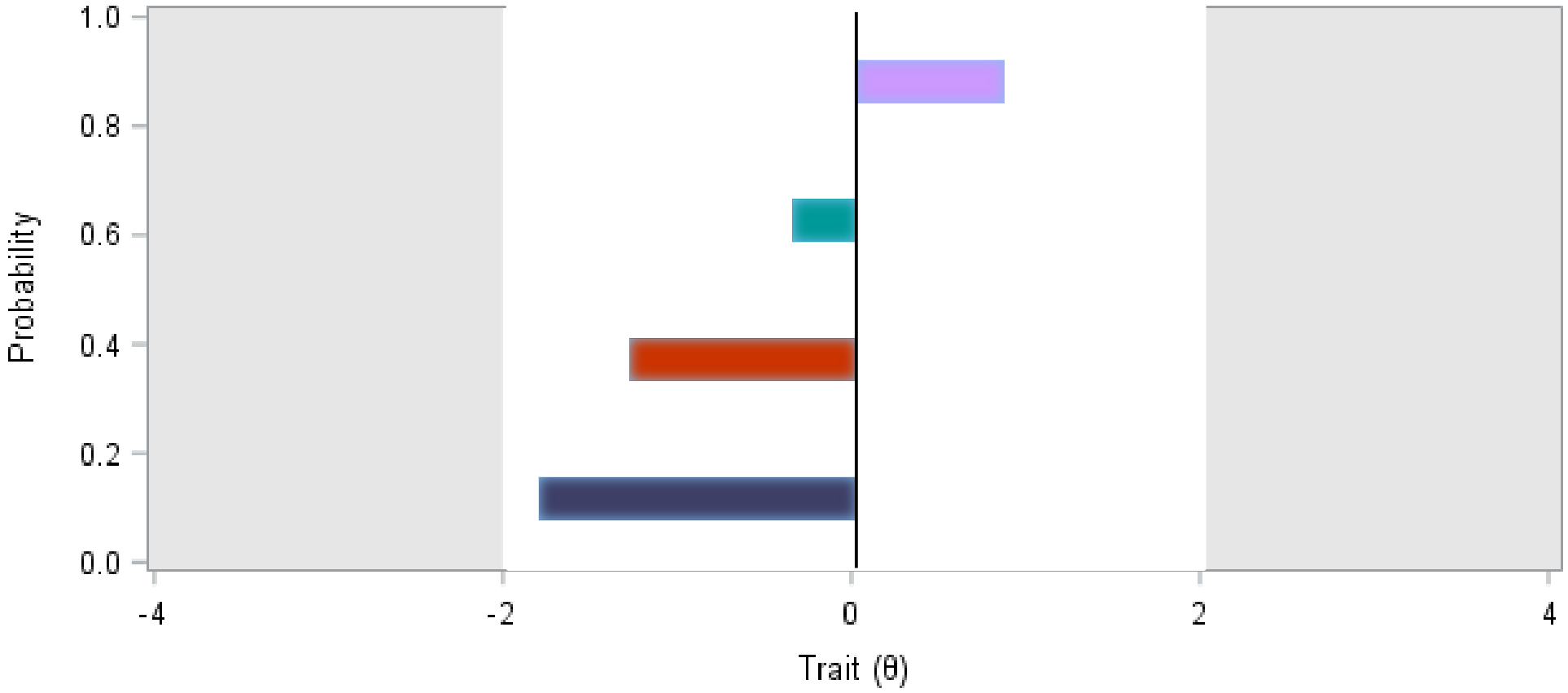
    "Persuasive"
  - Very

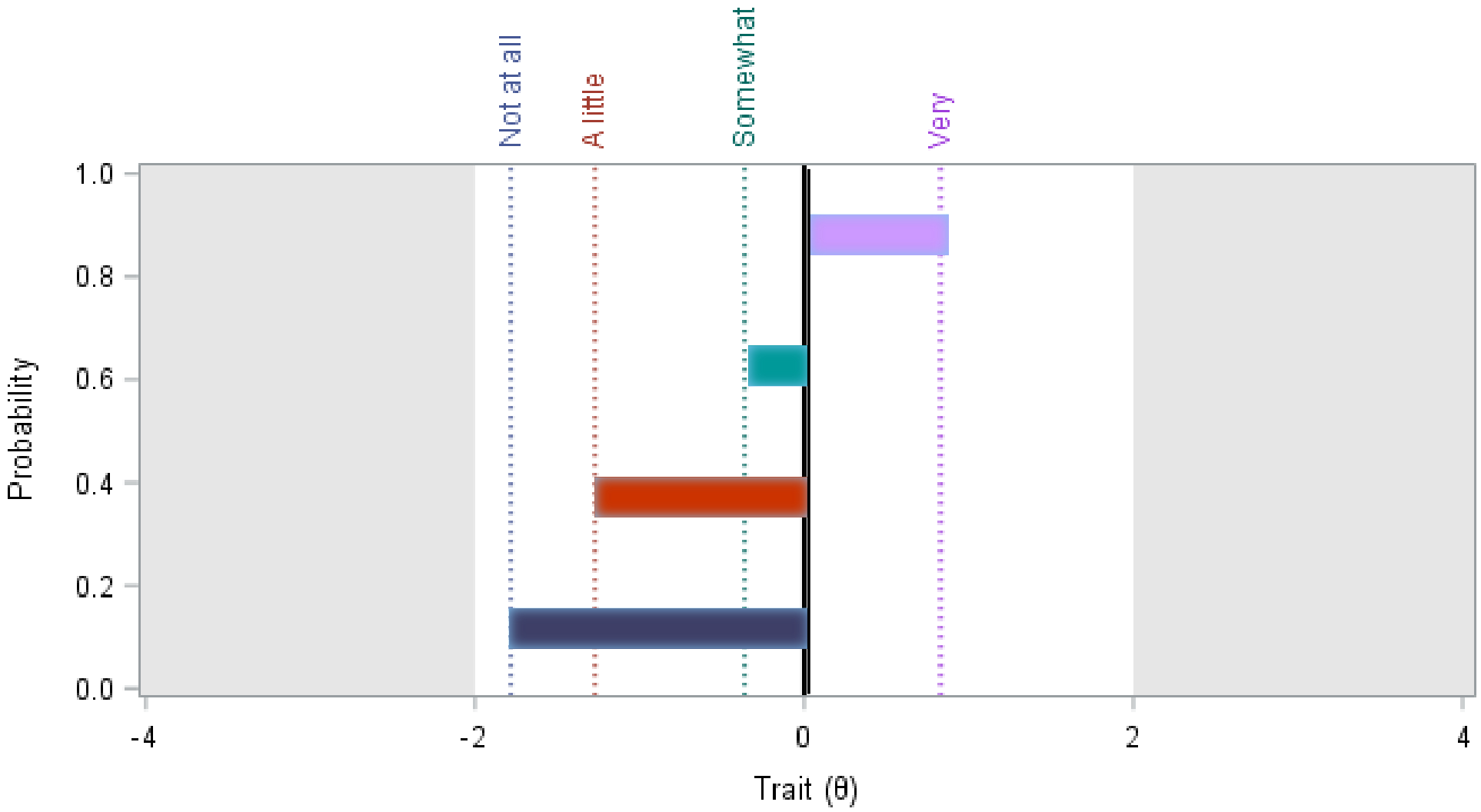

1.0
0.8
0.6
0.4
0.2
0.0
Probability
Extremely Low
Normal Range
Extremely High
-4
-2
0
2
4
Trait (θ)

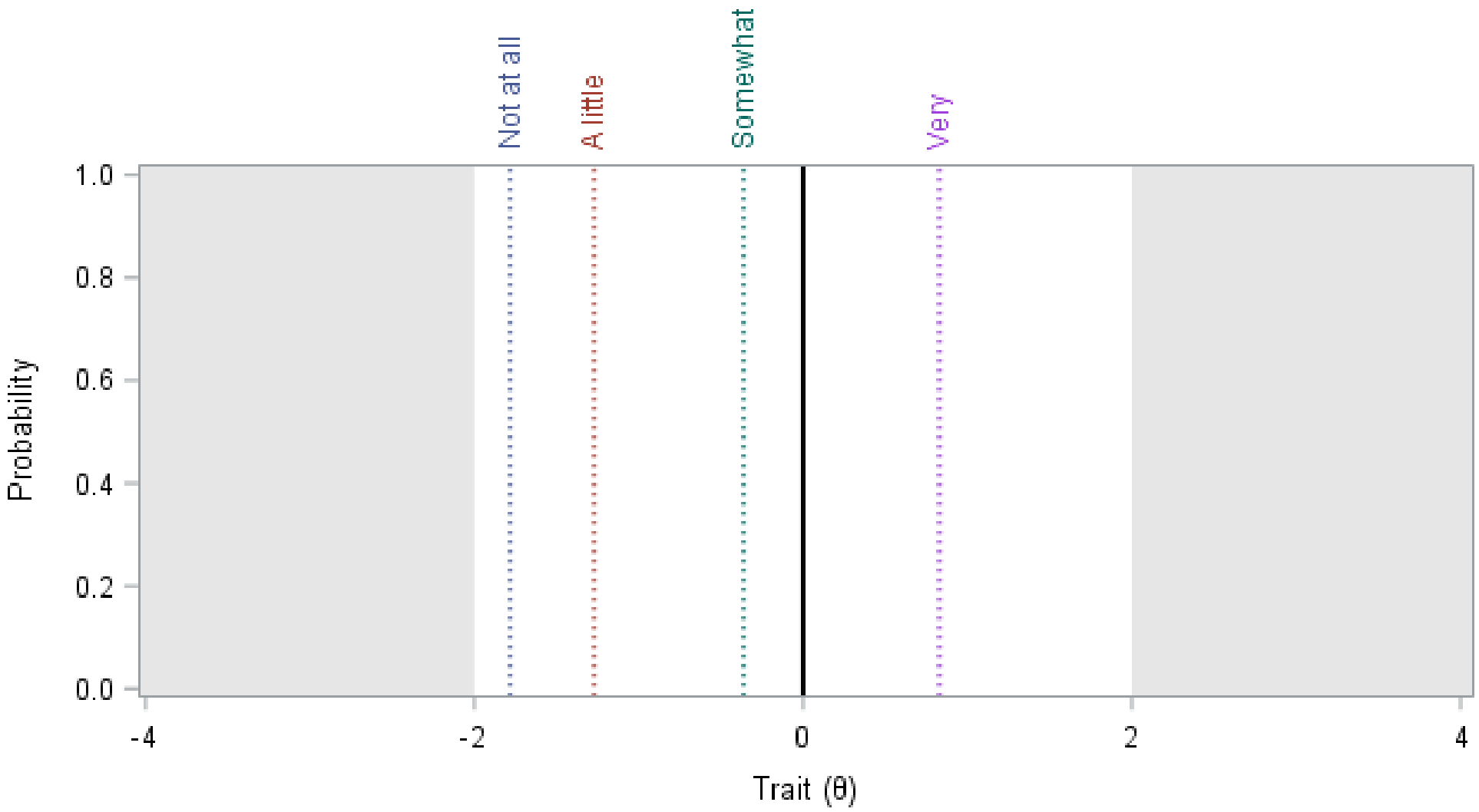
Amount terms

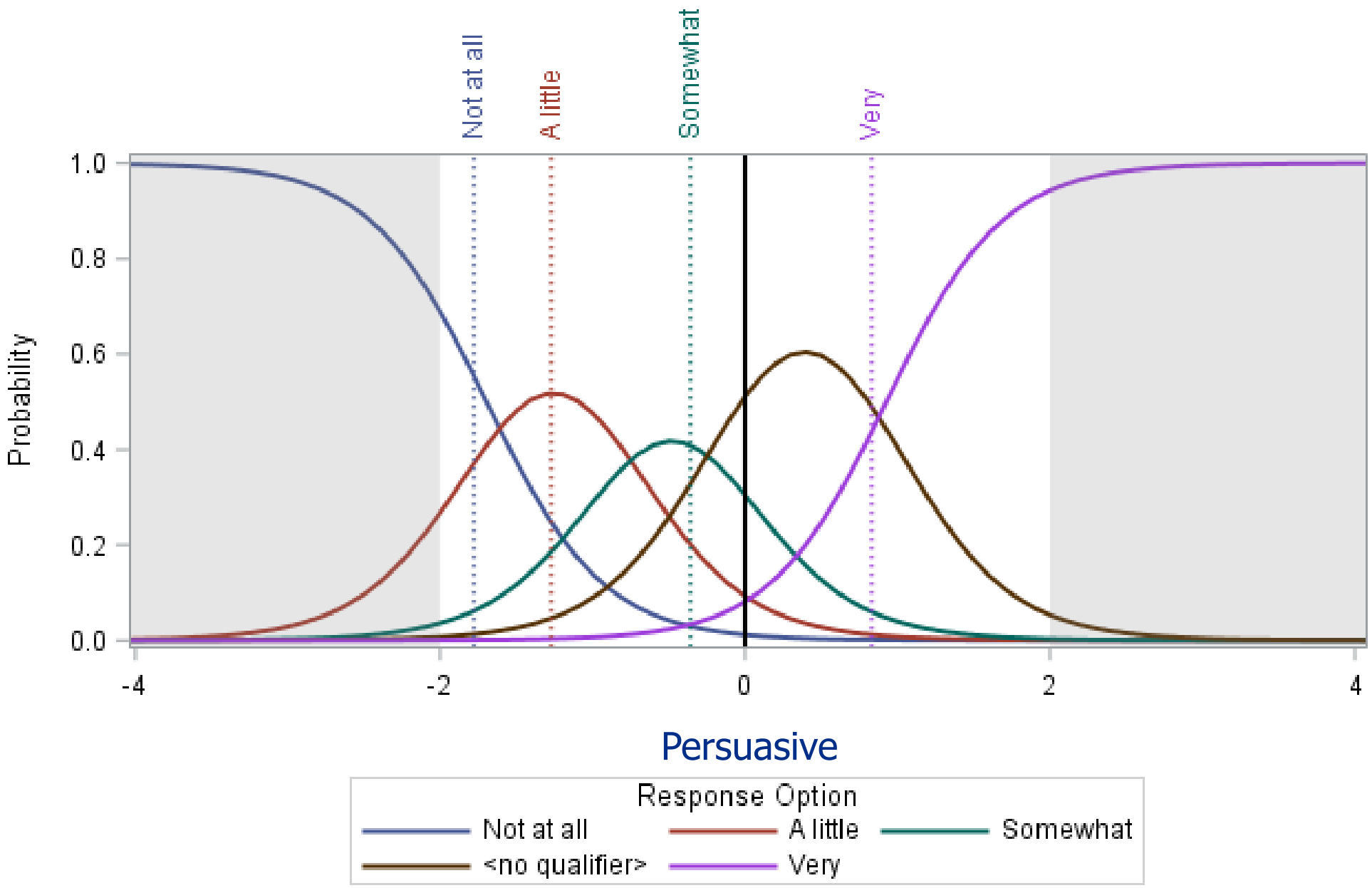


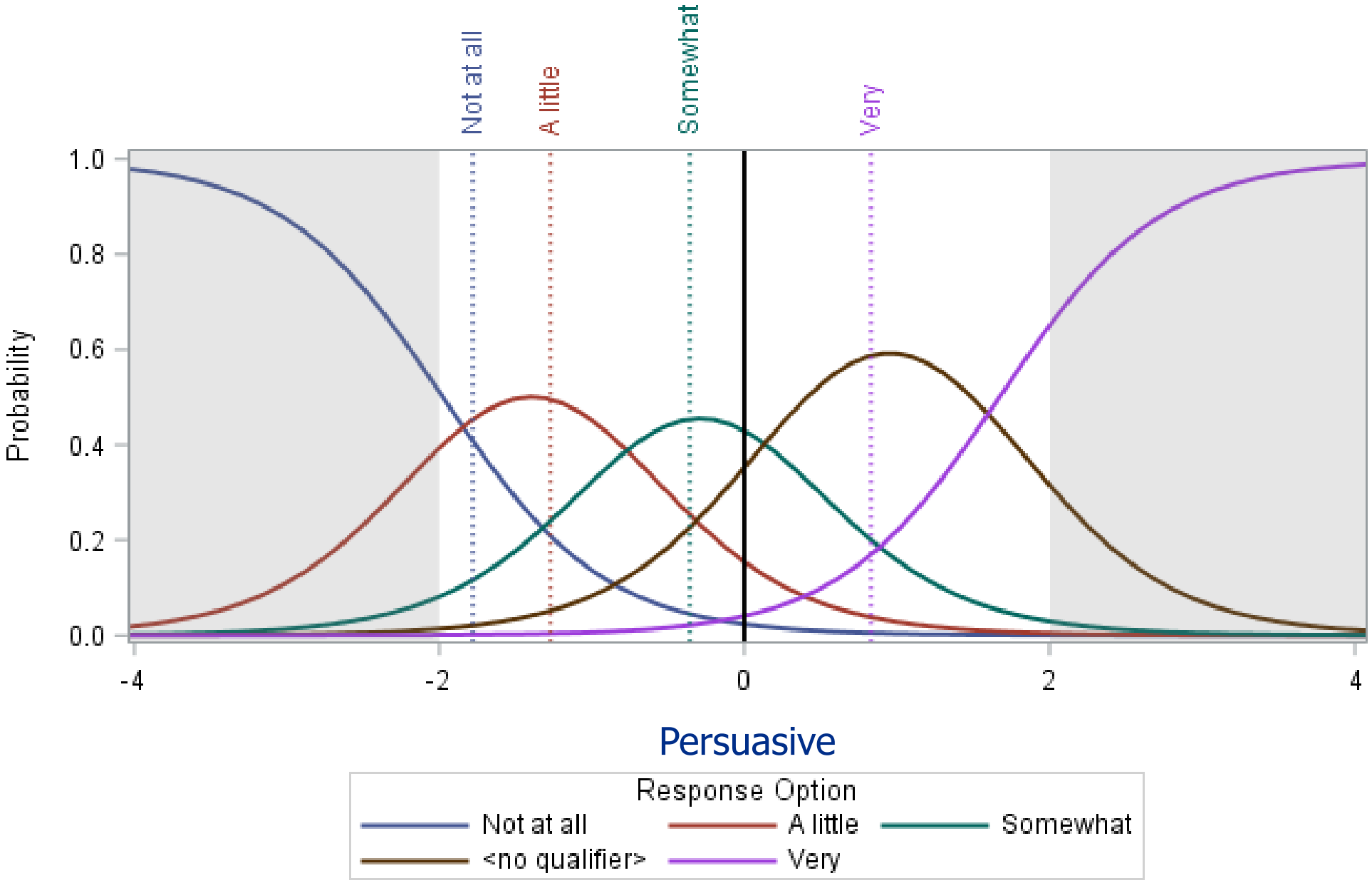

1.0
0.8
0.6
0.4
0.2
0.0
Probability
Very, 0.83
Somewhat, -0.36
A little, -1.27
Not at all, -1.78
-4
-2
0
2
4
Trait (θ)

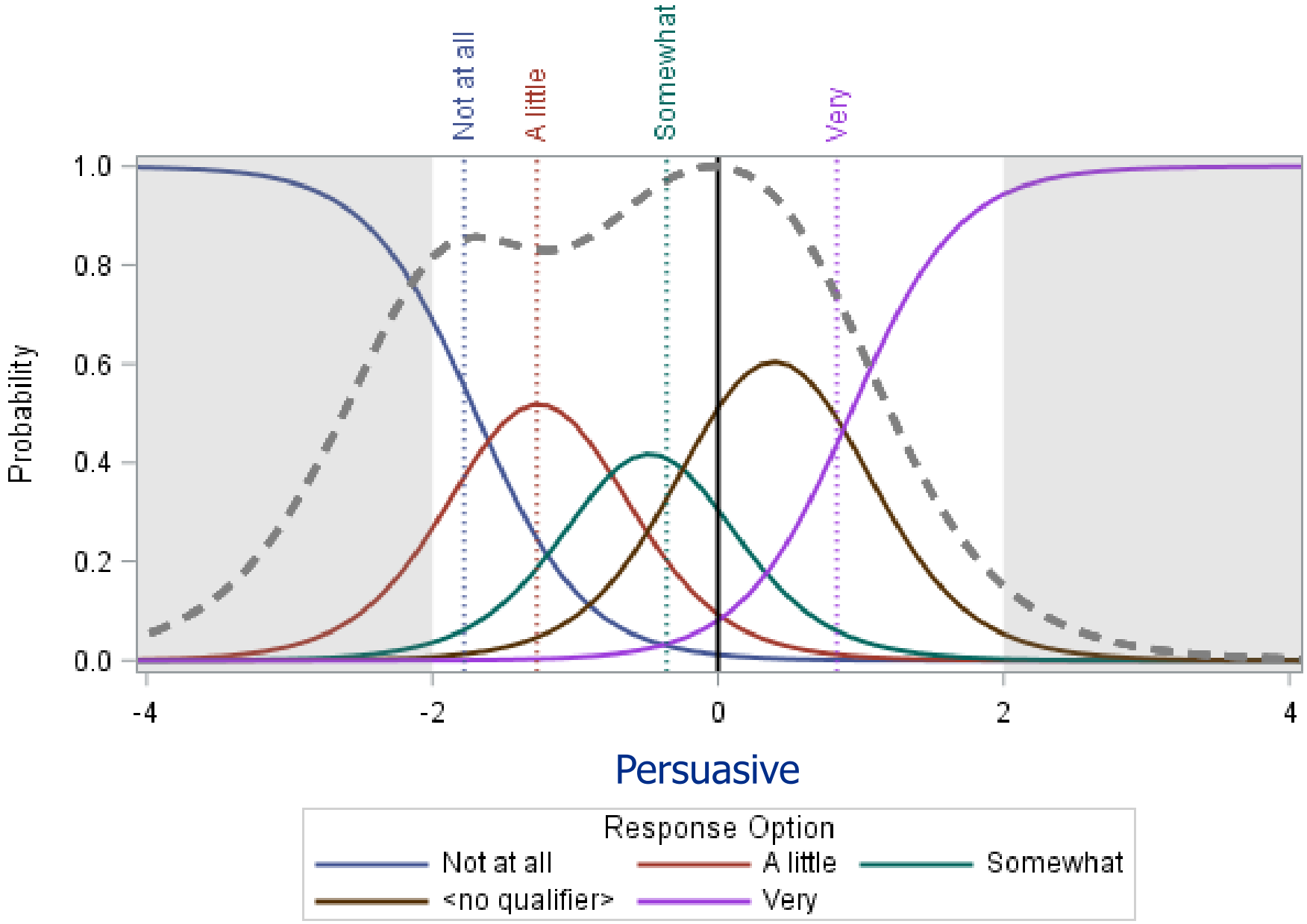

1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Trait (θ)

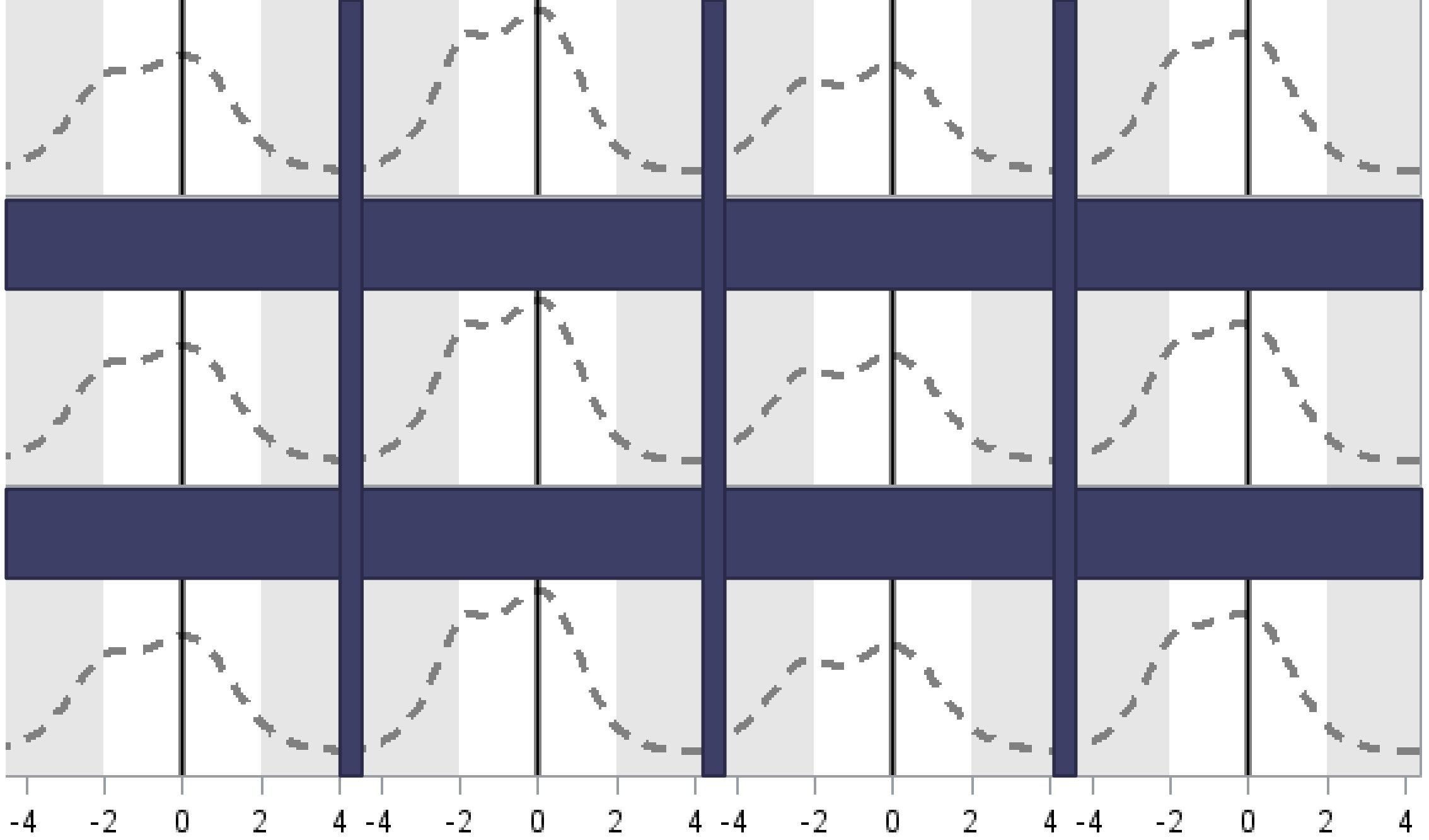

Not at all
A little
Somewhat
Very
1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Trait (θ)


Not at all
A little
Somewhat
Very
1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Trait (θ)

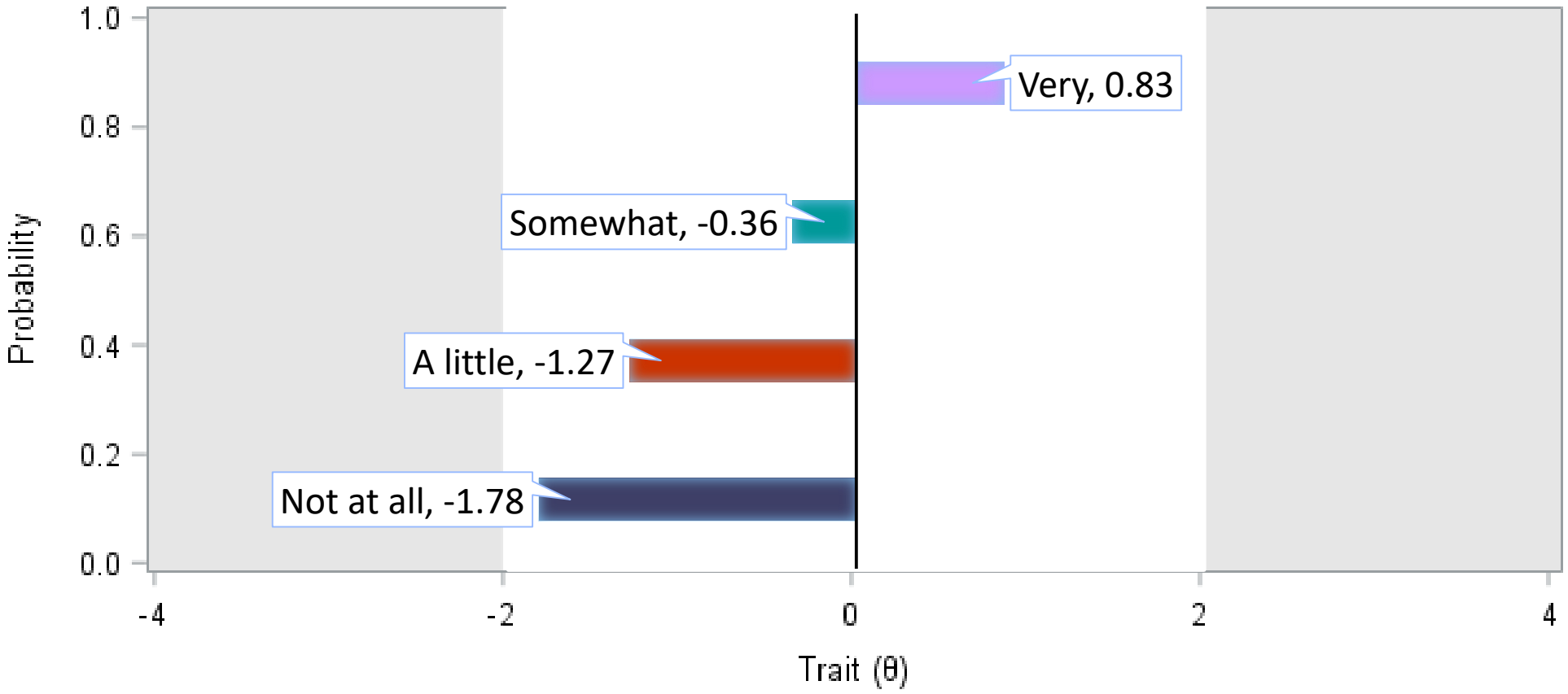

Not at all
A little
Somewhat
Very
1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Persuasive
Response Option
Not at all
A little
Somewhat
<no qualifier>
Very

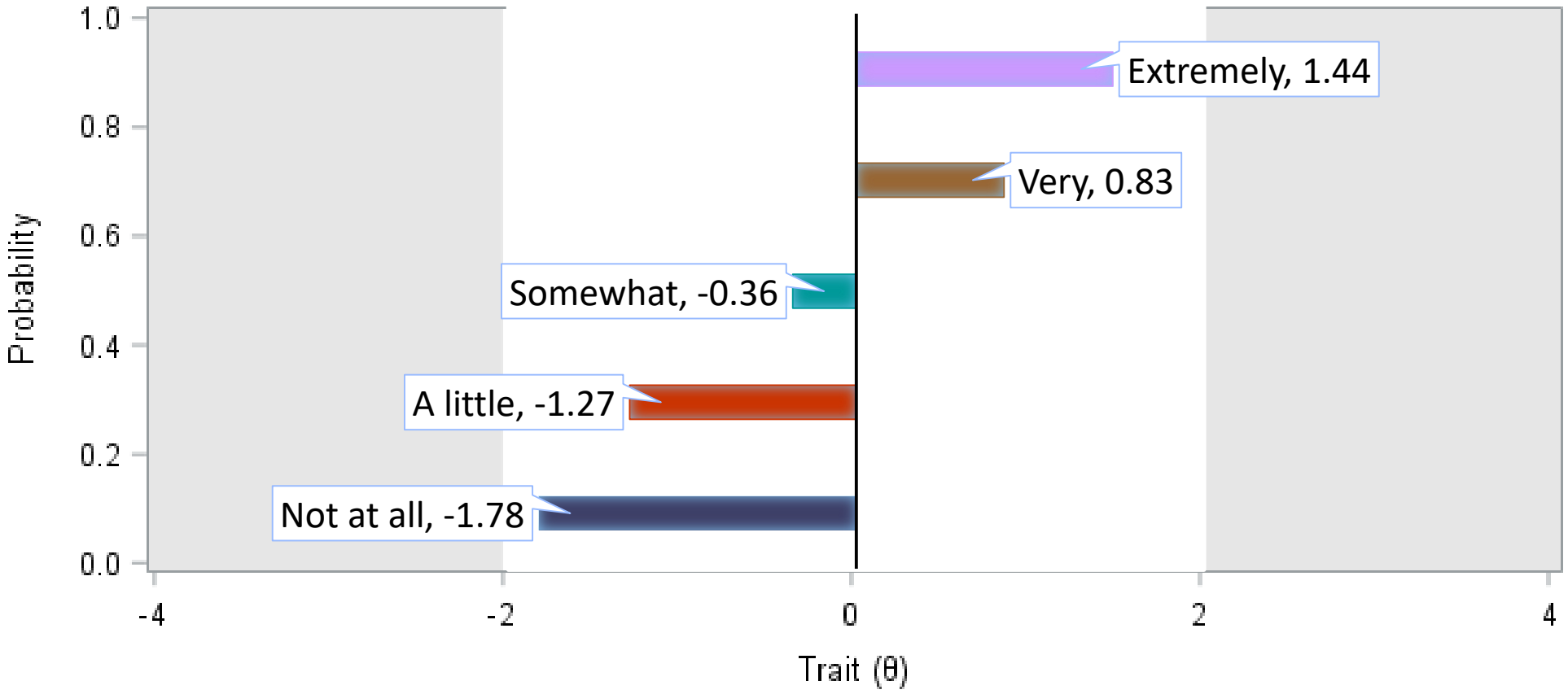

Not at all
A little
Somewhat
Very
1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Persuasive
Response Option
Not at all
A little
Somewhat
<no qualifier>
Very


Not at all
A little
Somewhat
Very
1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Persuasive
Response Option
Not at all
A little
Somewhat
<no qualifier>
Very



Persuasive


1.0
0.8
0.6
0.4
0.2
0.0
Probability
Very, 0.83
Somewhat, -0.36
A little, -1.27
Not at all, -1.78
-4
-2
0
2
4
Trait (θ)


Extremely, 1.44
Very, 0.83
Somewhat, -0.36
A little, -1.27
Not at all, -1.78
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-4
-2
0
2
4
Trait (θ)

Very vs. Extremely mTurk Comparison

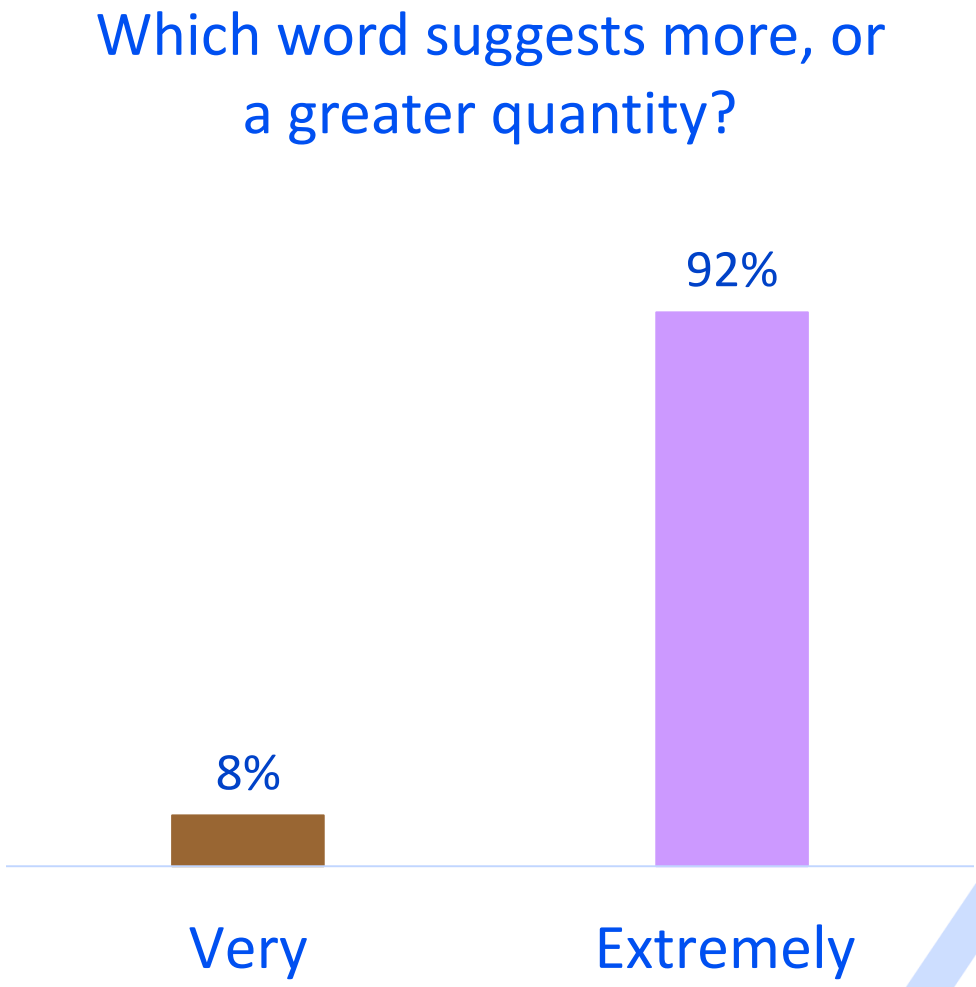
Which word suggests more, or a greater quantity?

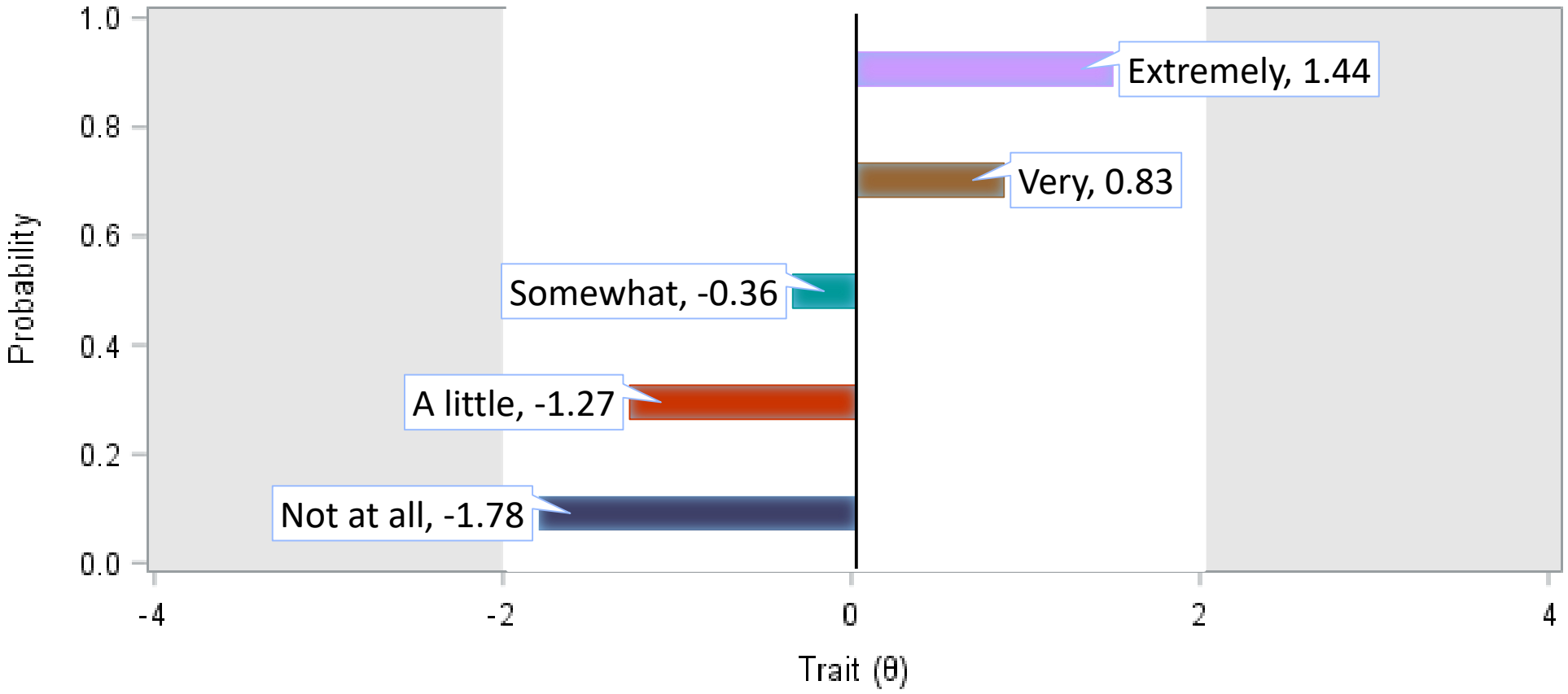
- ◯ Very
- ◯ Extremely

Very vs. Extremely mTurk Comparison

Which word suggests more, or a greater quantity?

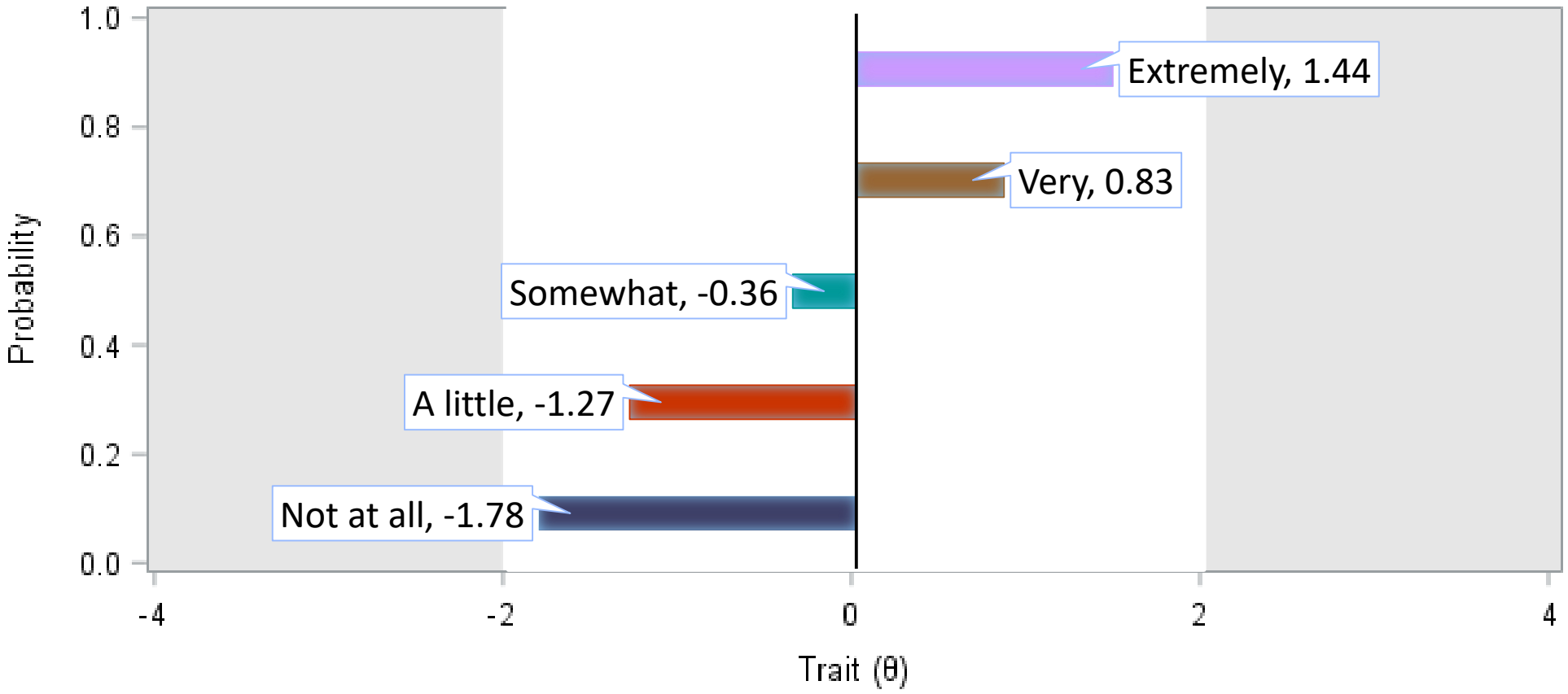
- Very
- Extremely

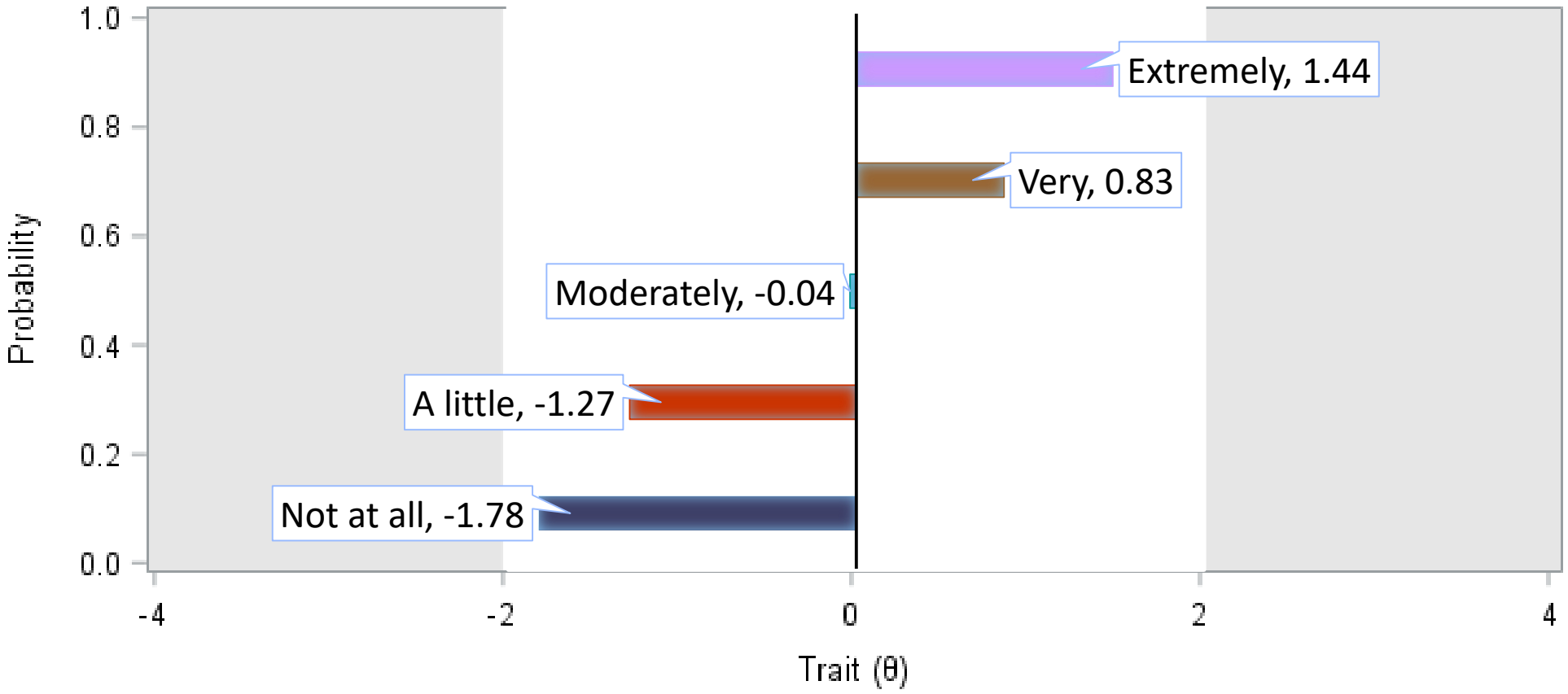



Extremely, 1.44
Very, 0.83
Somewhat, -0.36
A little, -1.27
Not at all, -1.78
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-4
-2
0
2
4
Trait (θ)

Case Study Response Scales

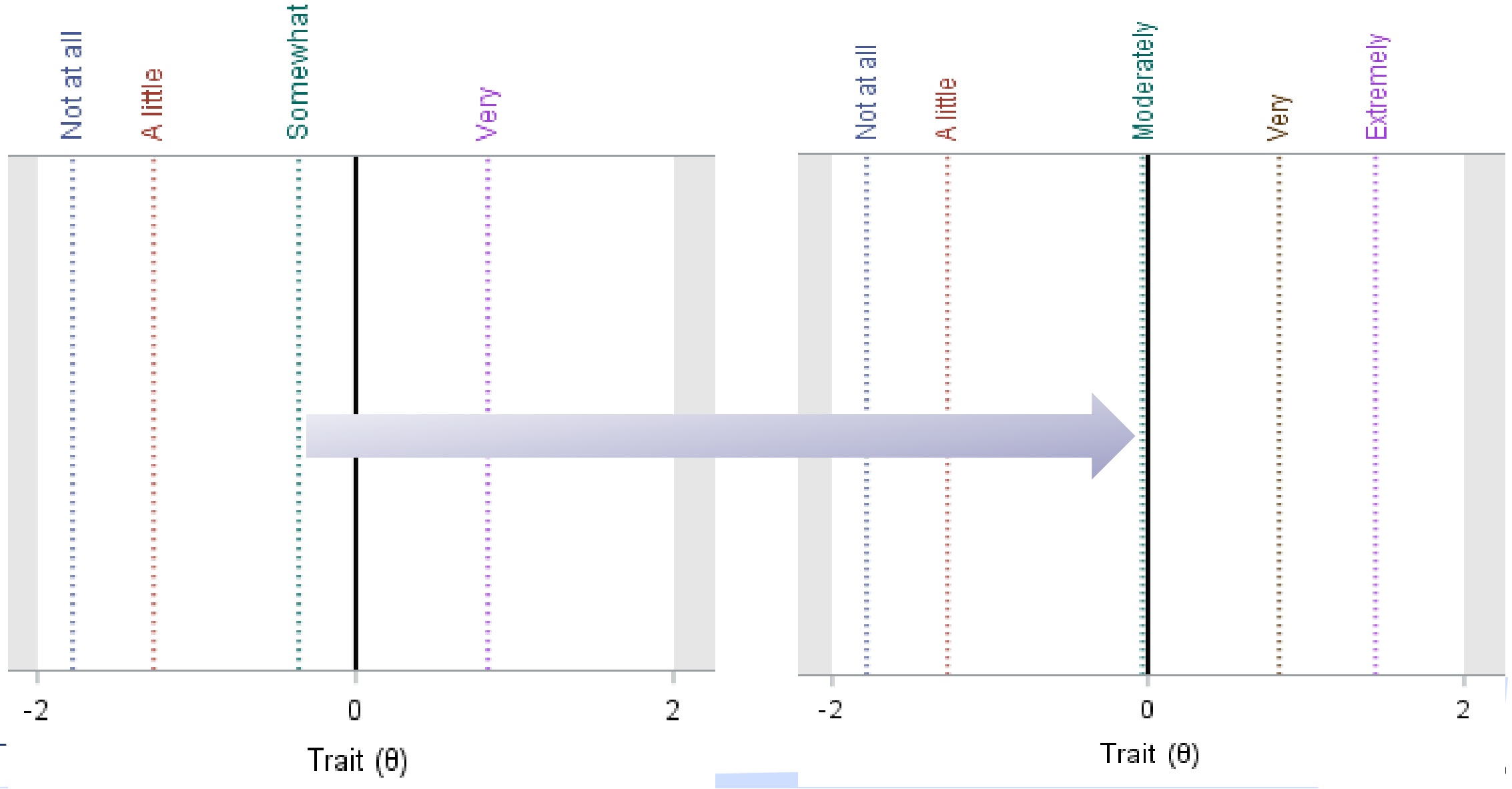
- Case Study 1 (Persuasive)
  - Not at all
  - A little
  - Somewhat
  - <no qualifier>
    "Persuasive"
  - Very
- Case Study 2 (Concern)
  - Not at all
  - A little
  - Moderately
  - Very
  - Extremely
- Case Study 3a (Burden)
  - Not at all
  - A little
  - Moderately
  - Very
  - Extremely
- Case Study 3b (Burden)
  - Not at all
  - Somewhat
  - Moderately
  - Very
  - Extremely

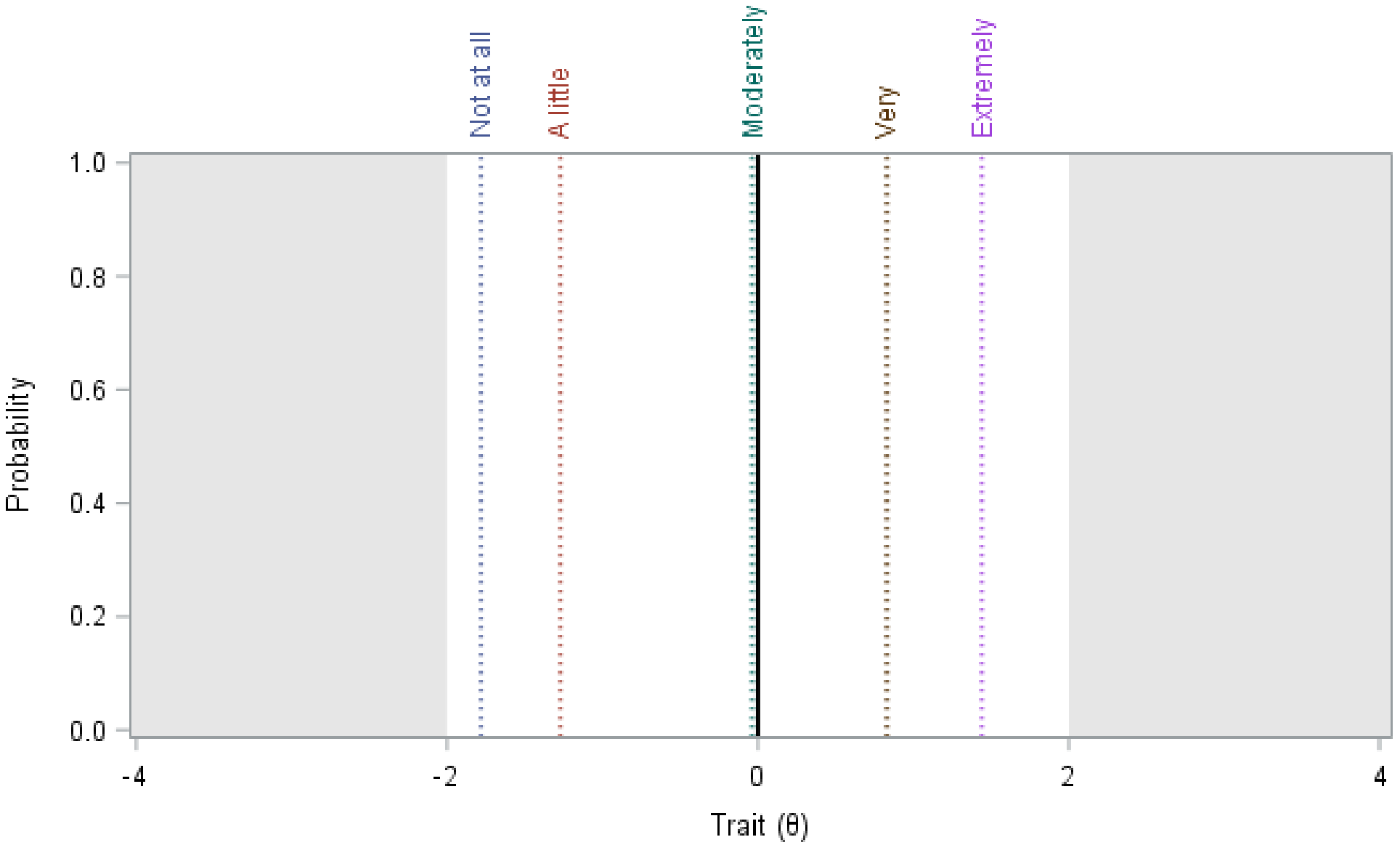

Extremely, 1.44
Very, 0.83
Somewhat, -0.36
A little, -1.27
Not at all, -1.78
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-4
-2
0
2
4
Trait (θ)

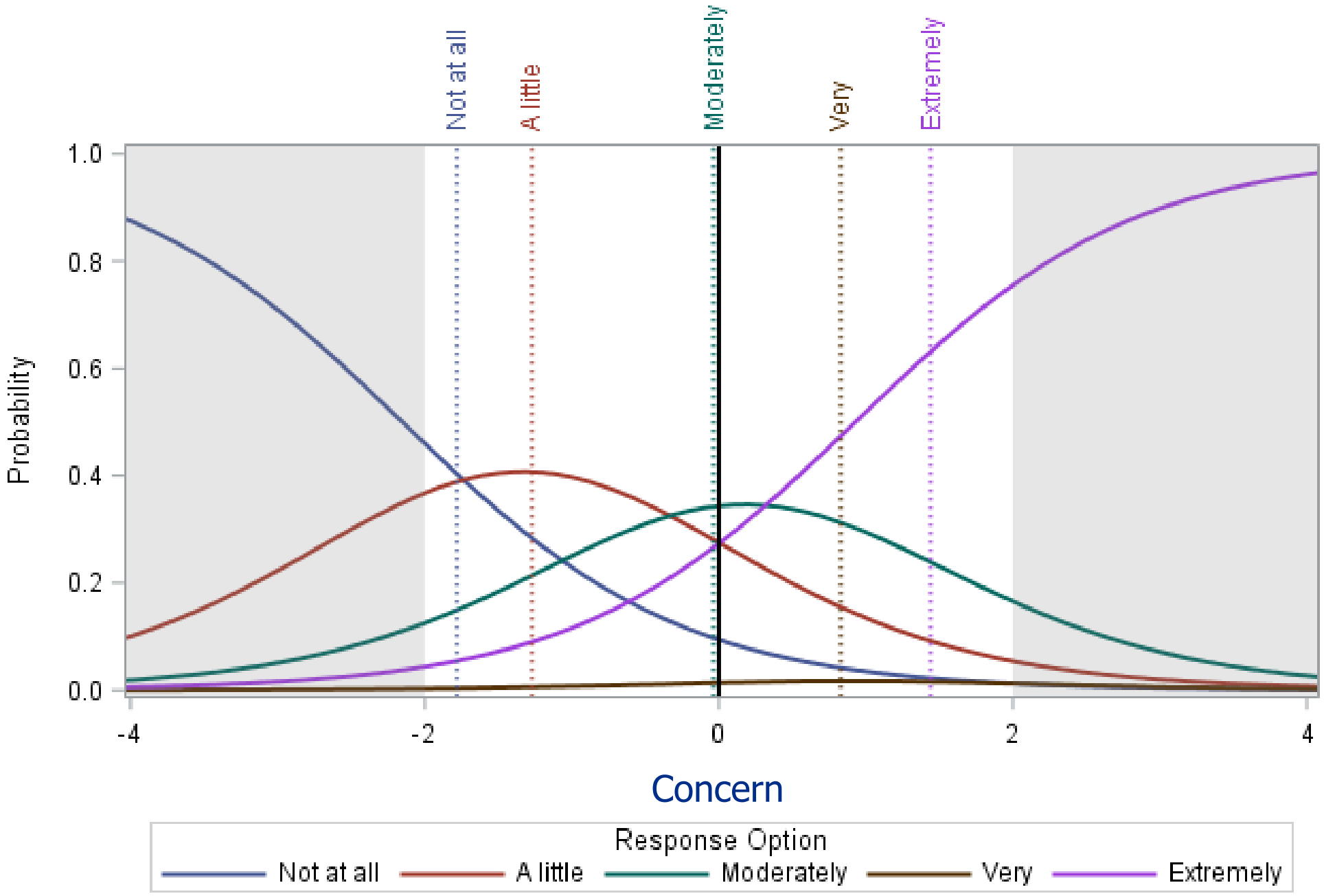

Extremely, 1.44
Very, 0.83
Moderately, -0.04
A little, -1.27
Not at all, -1.78
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-4
-2
0
2
4
Trait (θ)

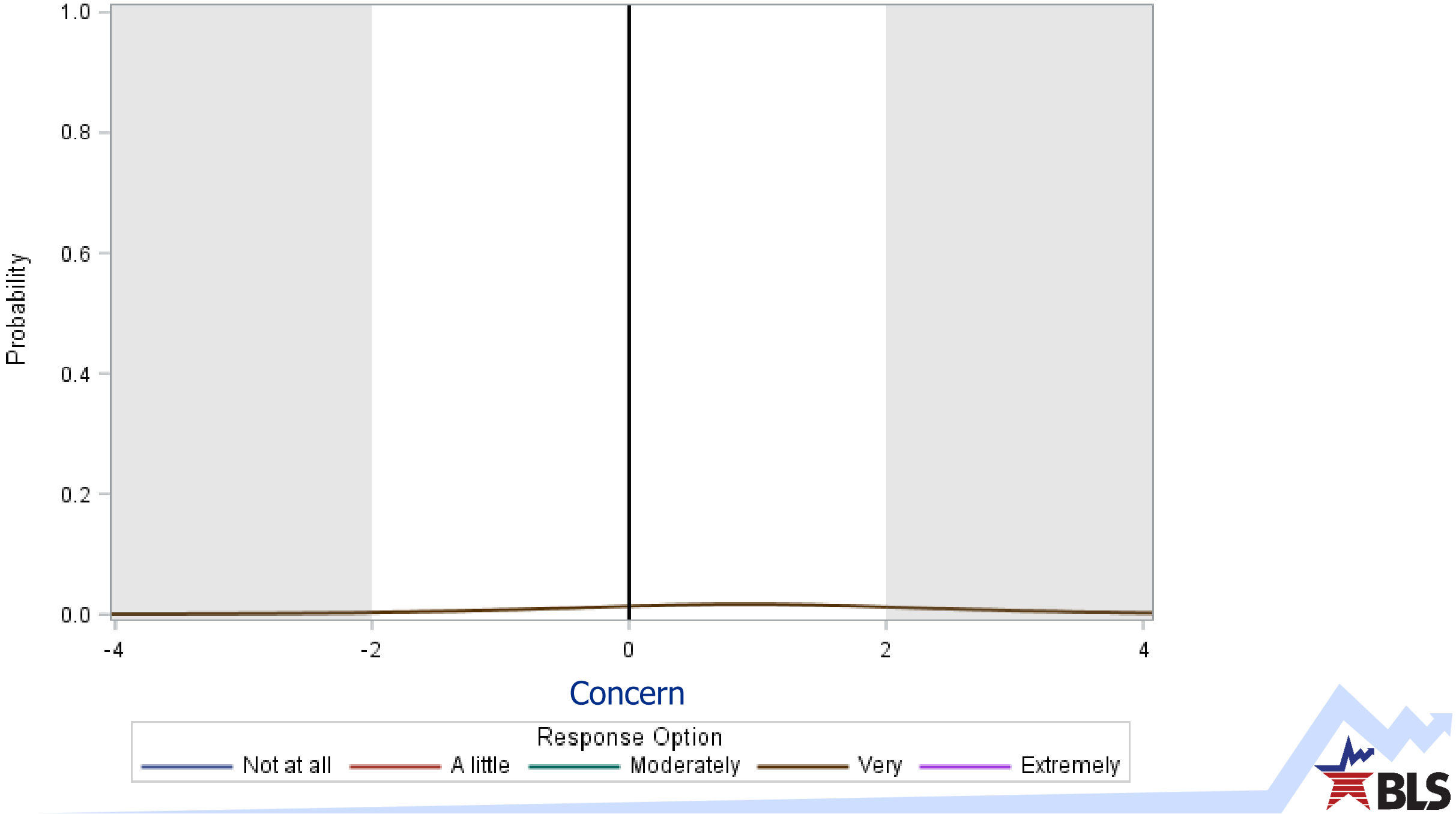
Case Study 1 (Persuasive)

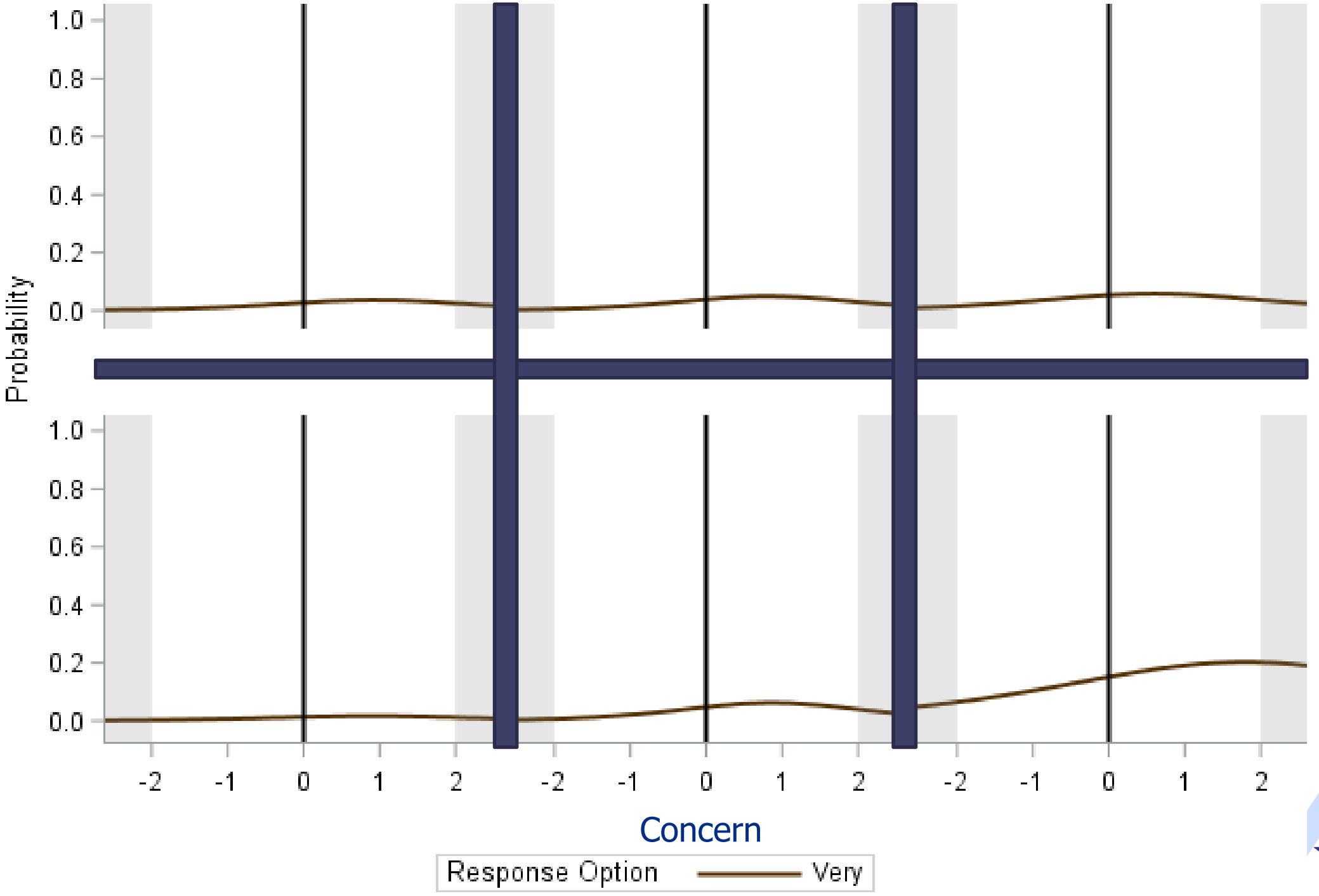
Case Study 2 (Concern)

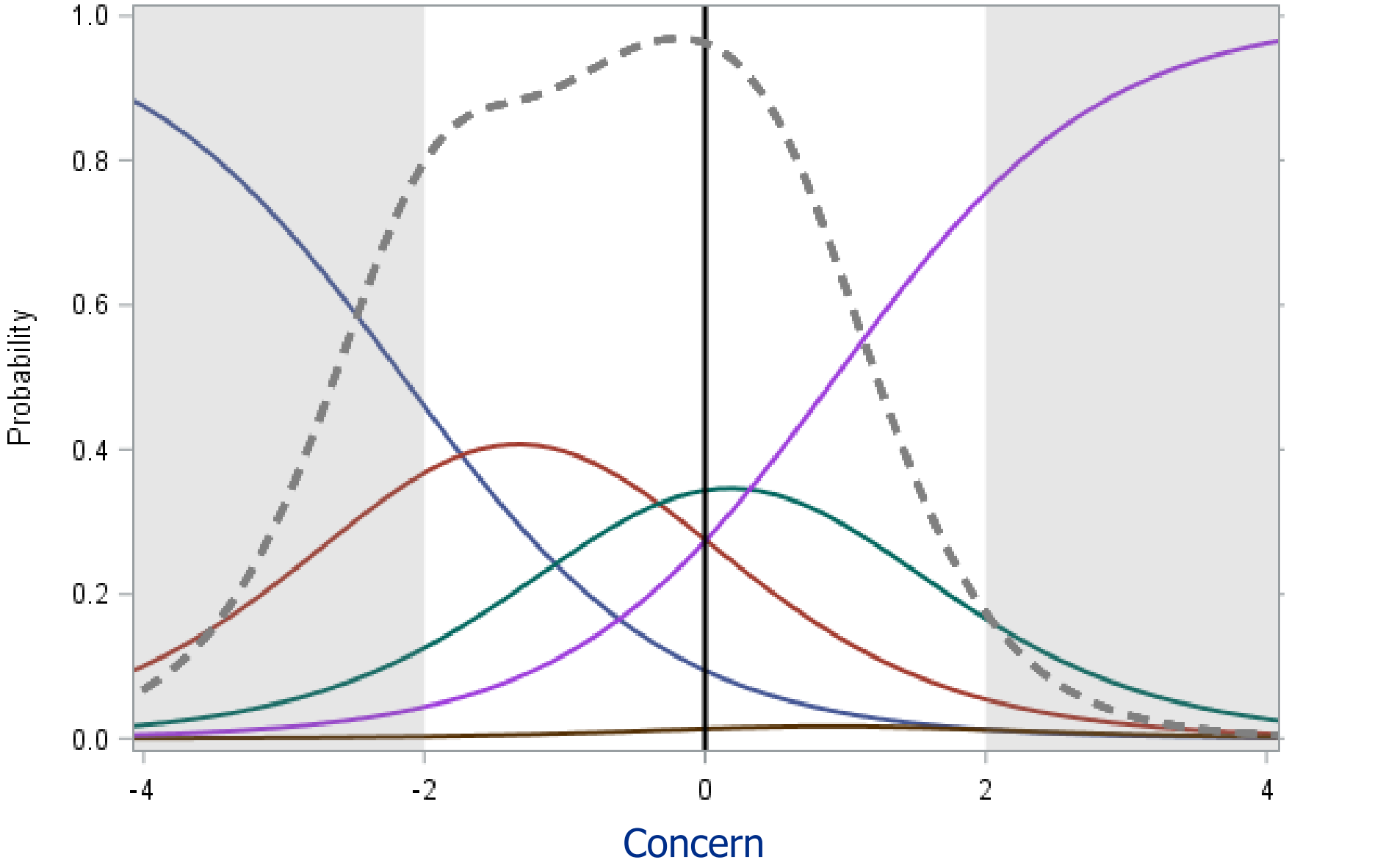


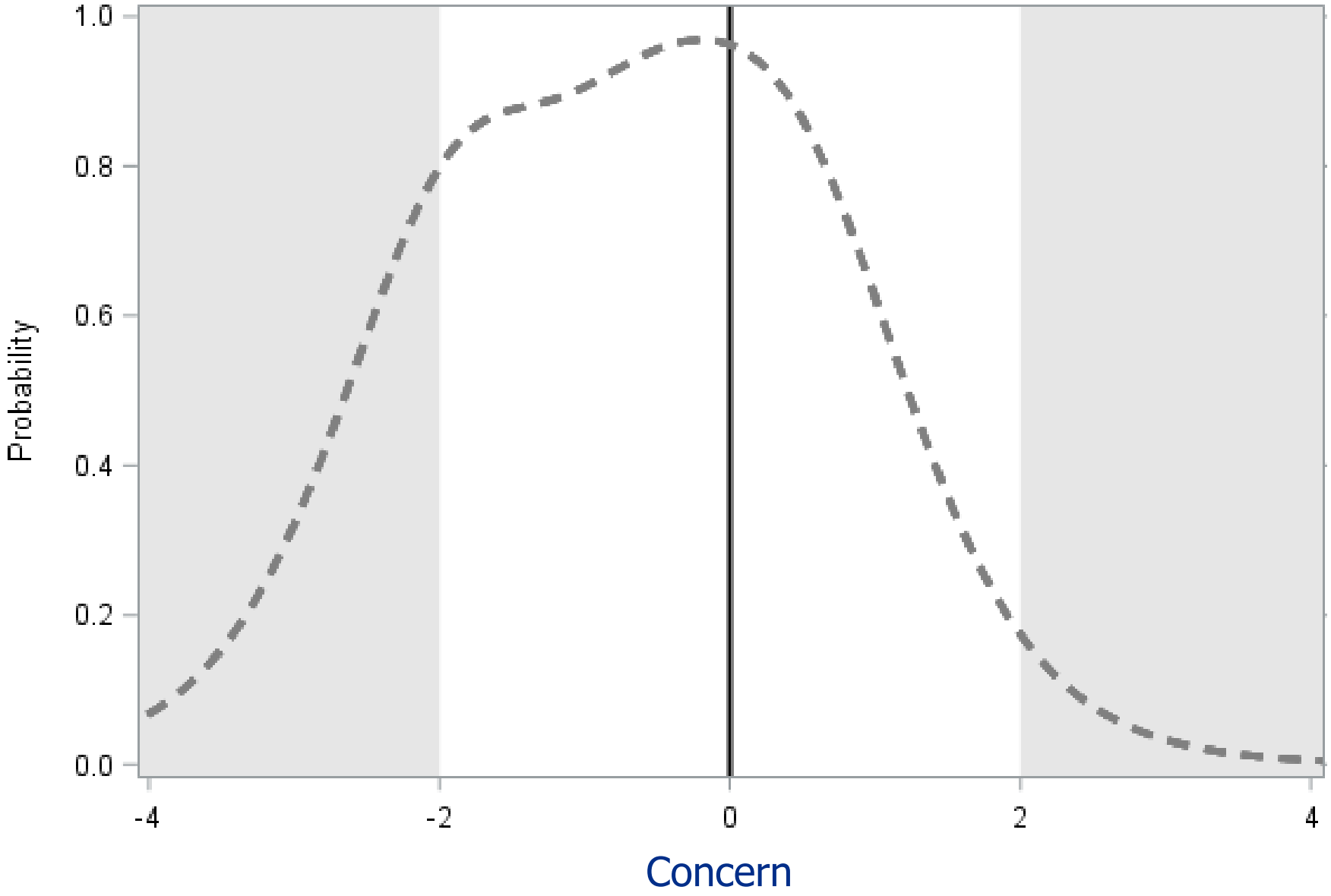

Not at all
A little
Moderately
Very
Extremely
1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Trait (θ)

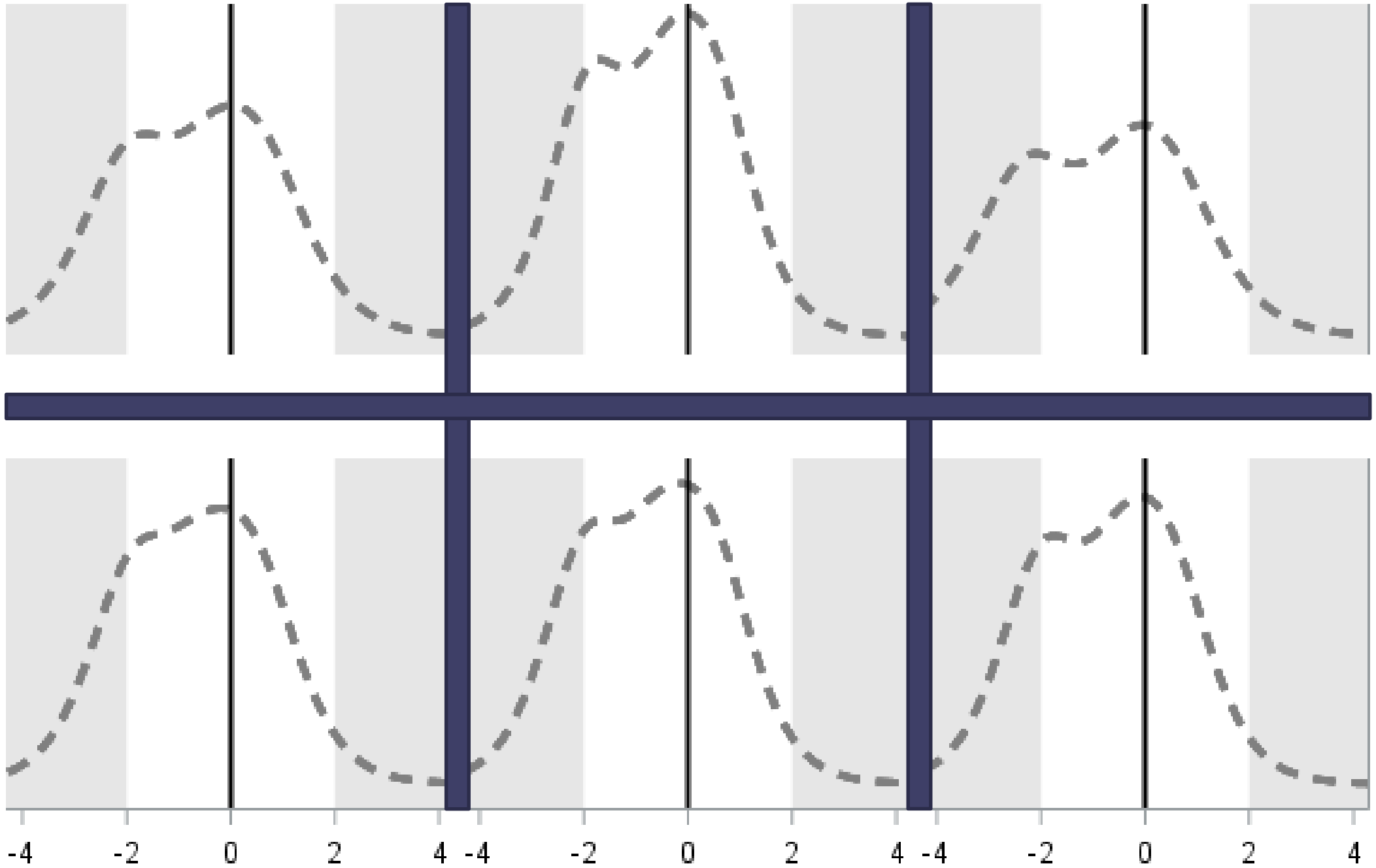

Not at all
A little
Moderately
Very
Extremely
1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Concern
Response Option
Not at all
A little
Moderately
Very
Extremely


1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Concern
Response Option
Not at all
A little
Moderately
Very
Extremely


Probability
1.0
0.8
0.6
0.4
0.2
0.0
-2
-1
0
1
2
Concern
Response Option
Very

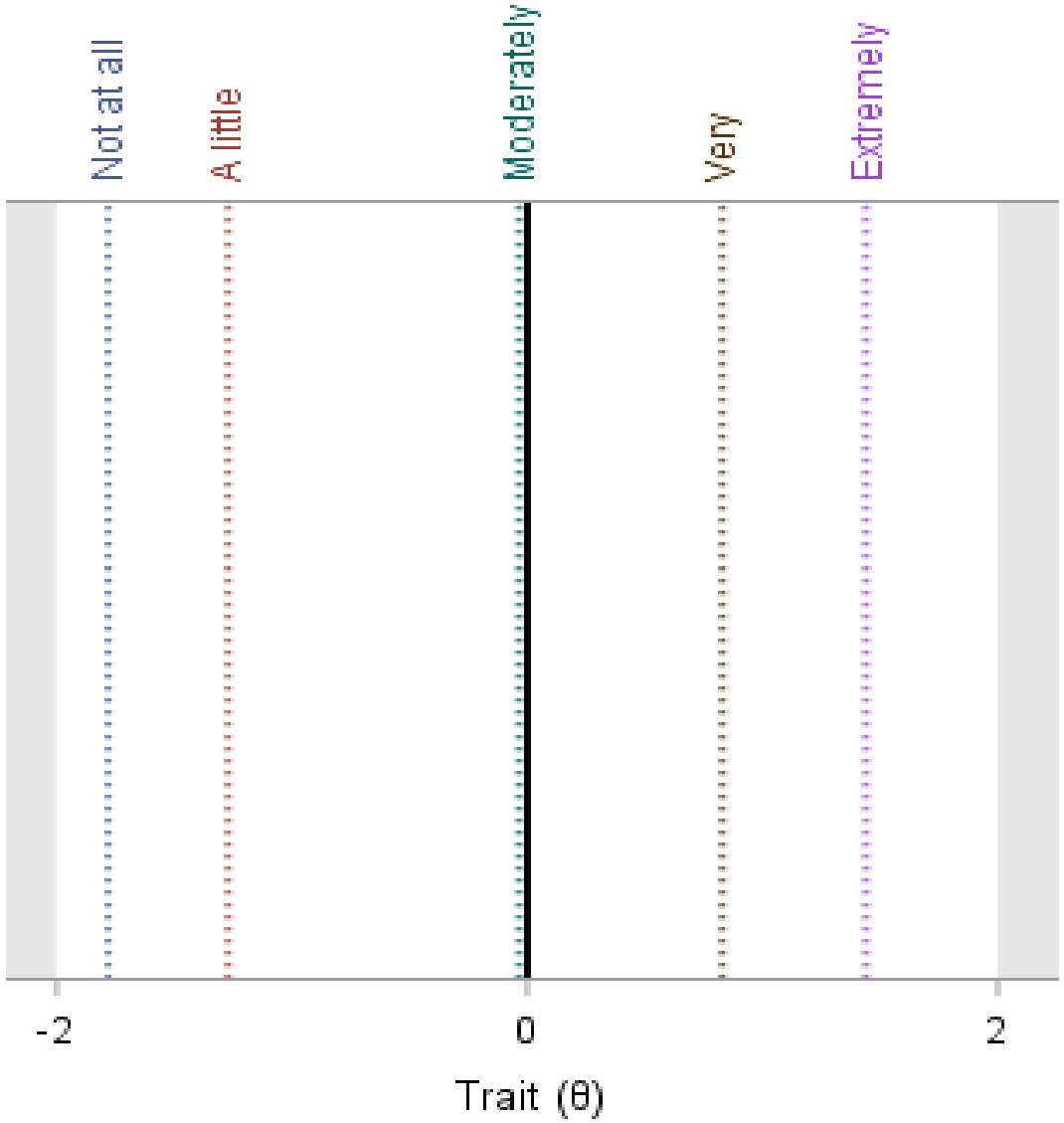

1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Concern
Response Option
Not at all
A little
Moderately
Very
Extremely


1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Concern

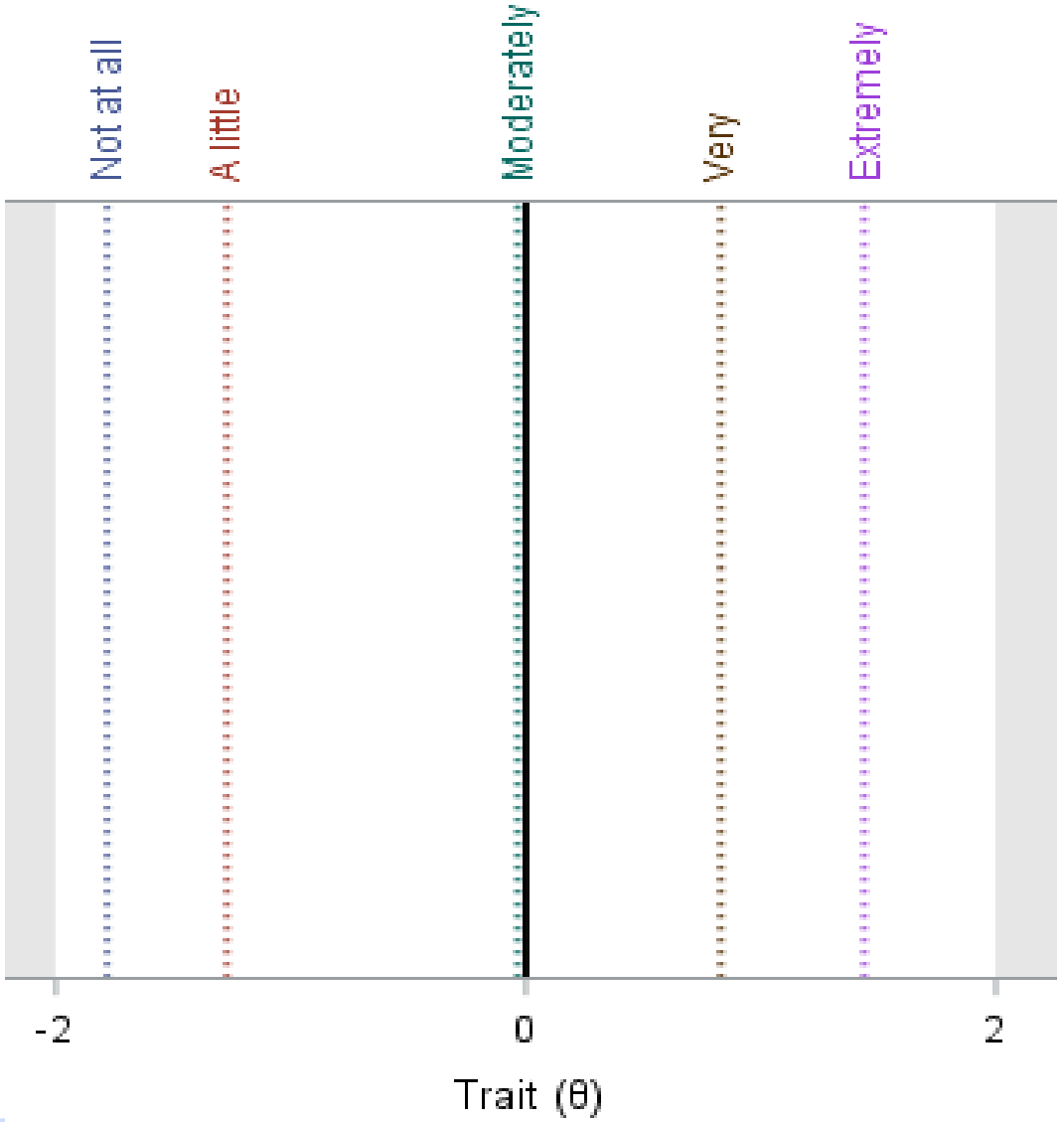


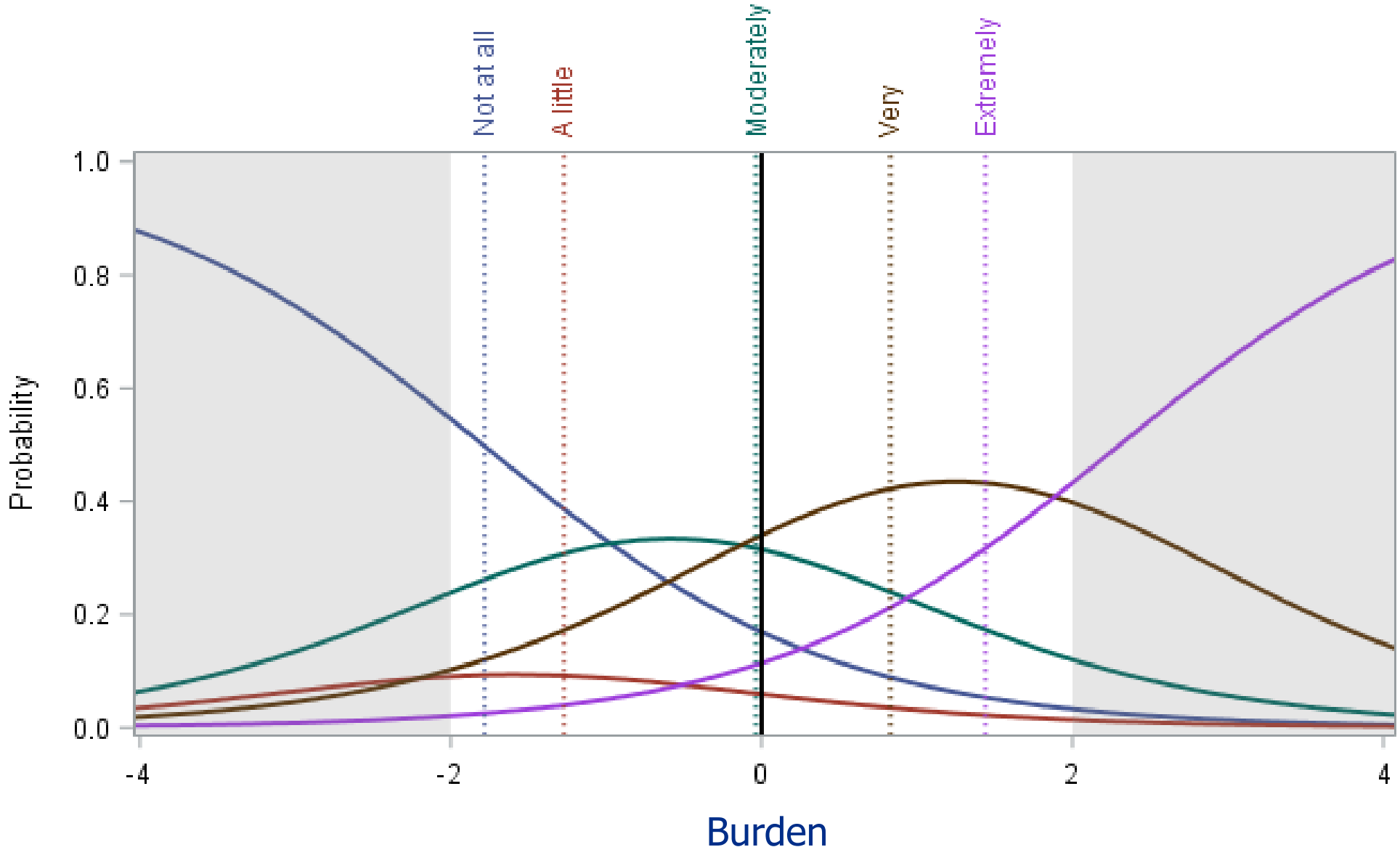
Concern

Case Study 2 (Concern)

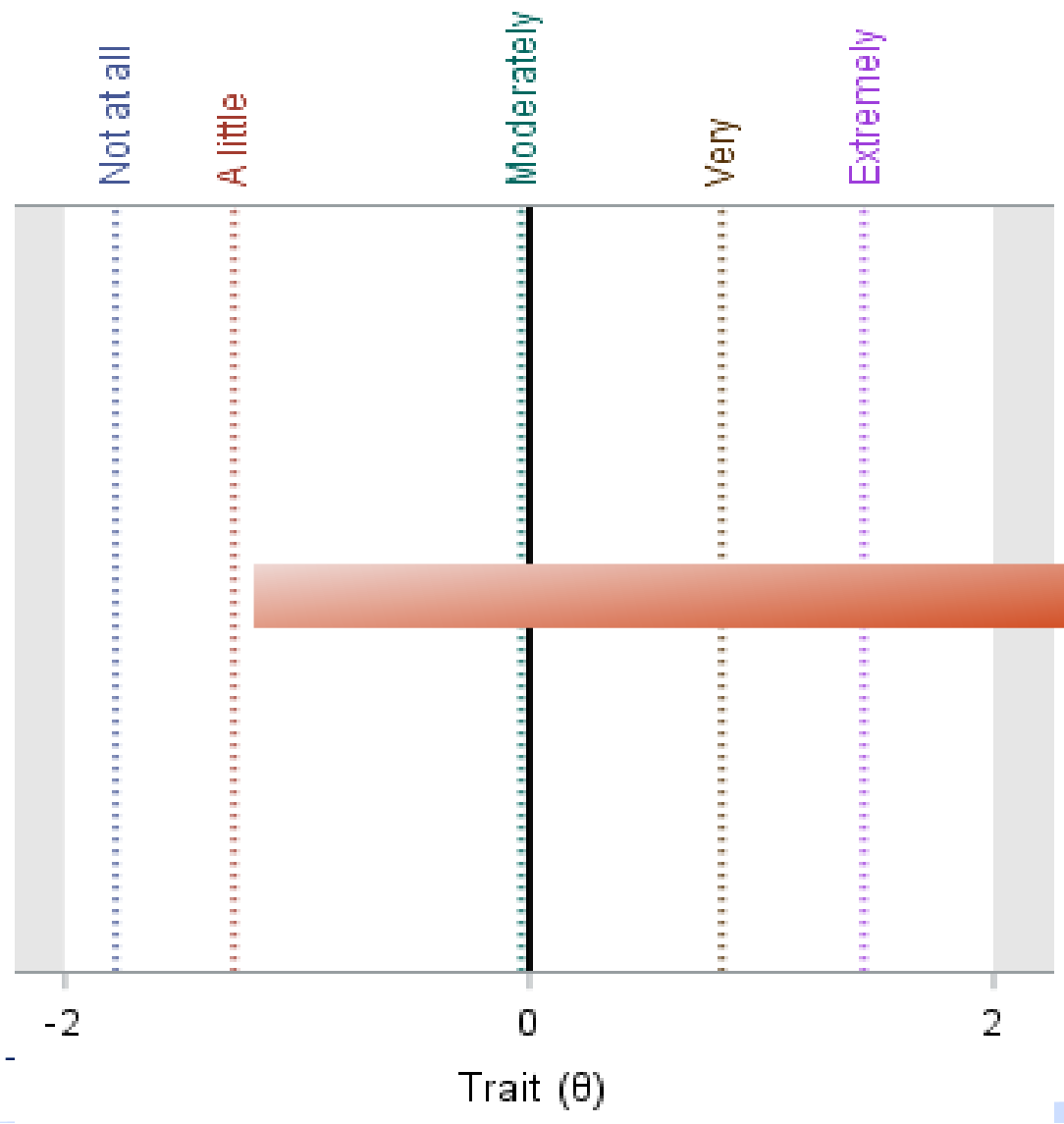


Case Study 3 (Burden)

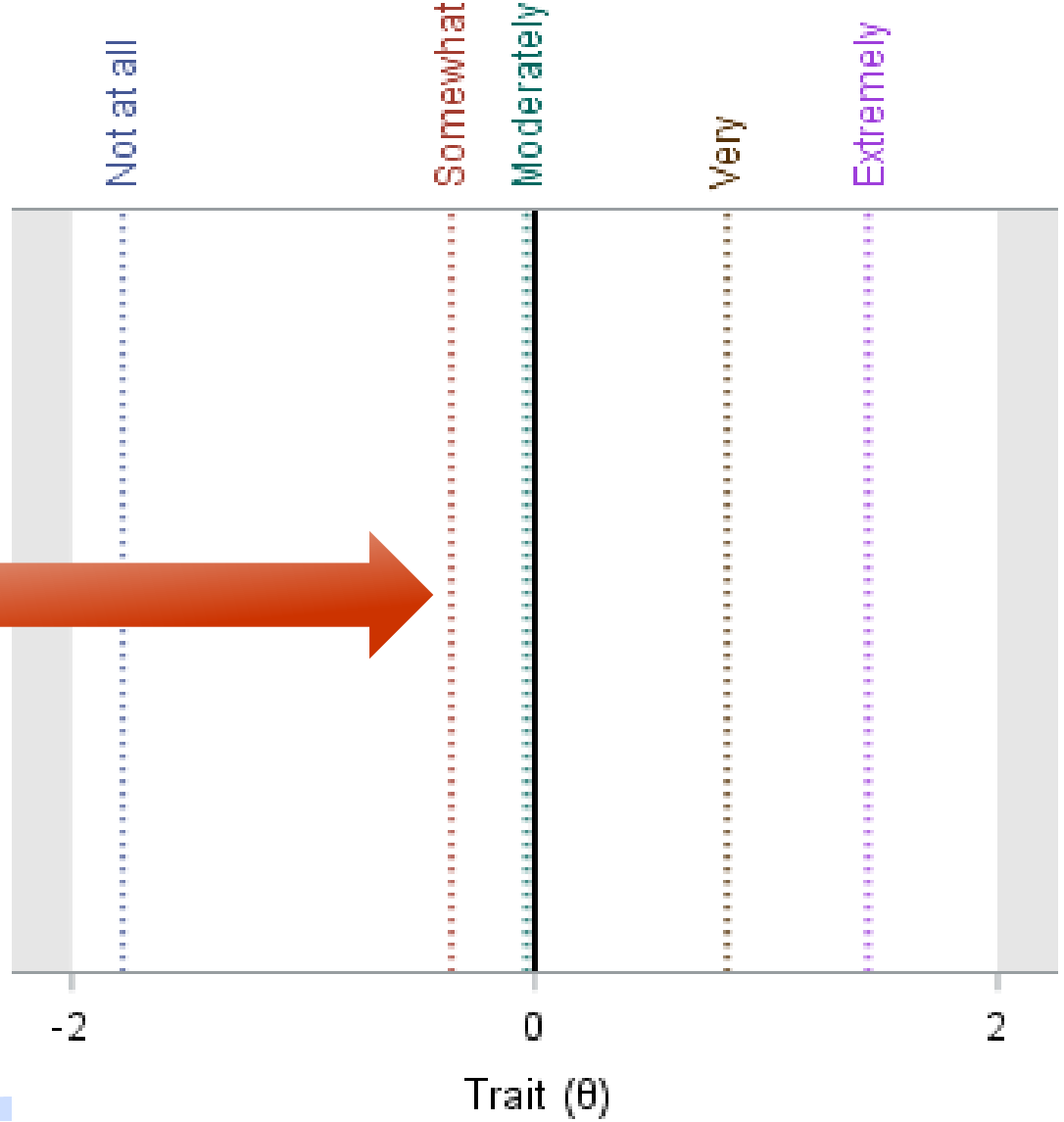


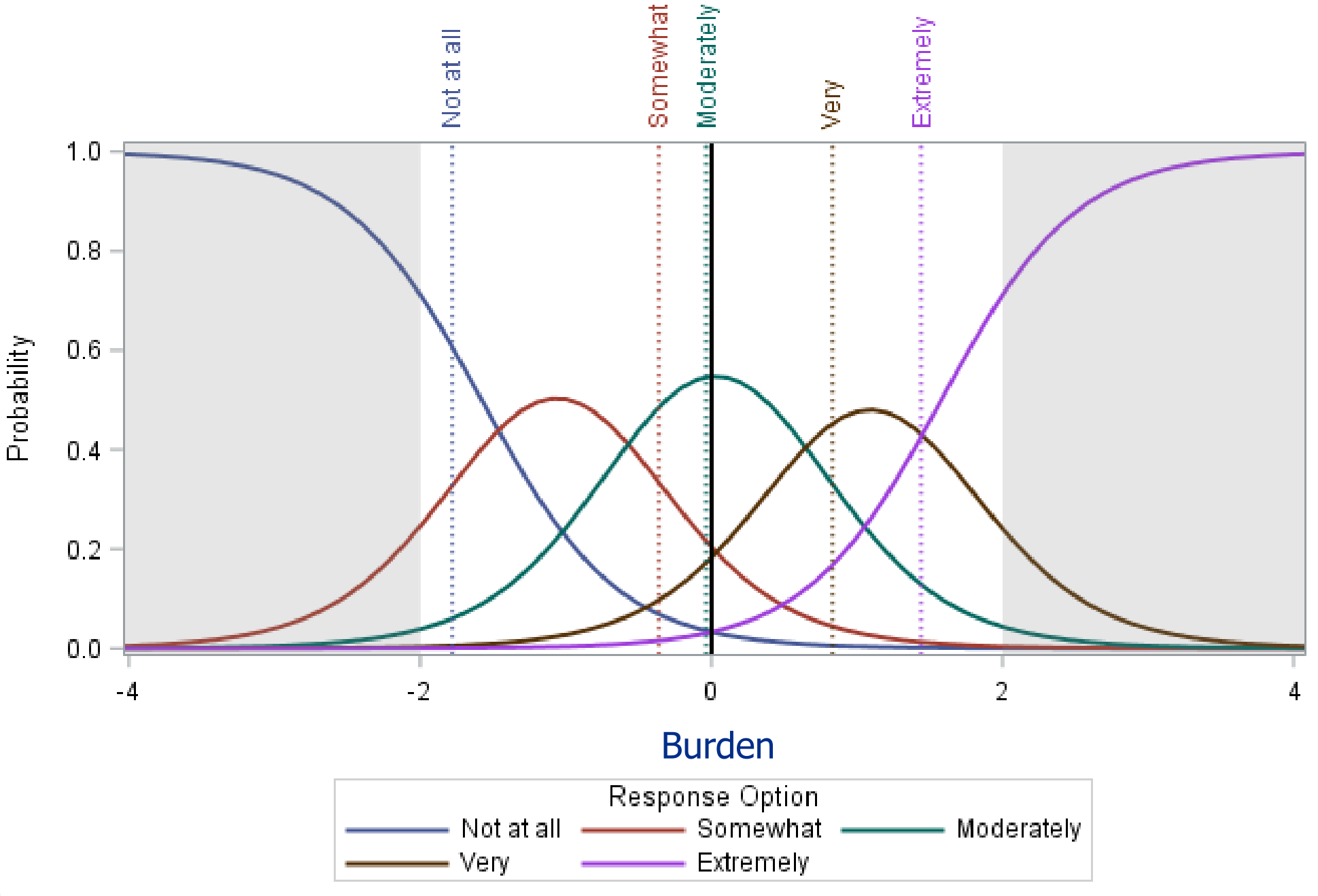

Not at all
A little
Moderately
Very
Extremely
1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Burden
Response Option
Not at all
A little
Moderately
Very
Extremely

Case Study 3a (Burden)



Case Study 3b (Burden)




Not at all
Somewhat
Moderately
Very
Extremely
1.0
0.8
0.6
0.4
0.2
0.0
Probability
-4
-2
0
2
4
Burden
Response Option
Not at all
Somewhat
Moderately
Very
Extremely

Conclusions

- The response option probability distributions tended to follow the same order we observed in the MTurk study

- Specific findings
  - "Very" as an endpoint may not capture the full range of responses, but
  - Adding "Extremely" may suppress people using "Very"
  - Looking at "a little" vs "somewhat," the value assigned to a qualifier by a respondent may depend on the other responses in the scale.

Conclusions

- BUT the data in the case studies did not always match the expectations set by the values from MTurk study
  - Some scales that should have been well-distributed based on the MTurk findings were not, and
  - Some scales that should not have been well-distributed were.
- Factors that may impact the interpretation of individual scale items
  - The construct
  - The other response items used in the scale
  - The context of the survey item
  - The respondent population

Limitations

- We did not test every possible response option in our MTurk study, so we were limited in the case studies we could examine as a follow-up
- While we identified some interesting patterns between the MTurk and the case studies we had available, the sample size of case studies and constructs was extremely limited

Next Steps

- We would like to dig a little deeper into this, but we need more data to identify if there are consistent effects across contexts
  - Constructs
  - Response options
  - Populations
- Do you have publicly available data that uses some of the response options we assessed in the MTurk study?
  - Please contact Jean Fox fox.jean@bls.gov

Morgan Earp

earp.morgan@bls.gov