The (empirical) case for analyzing Likert data with parametric tests

Heather Wojton, PhD
Institute for Defense Analyses

04/xx/2017
Roadmap

- Evaluating human-system interactions in OT
- The great debate
- Let the data decide
- Recommendations for analysts
- I stop talking
Human-system interaction affects mission success
Likert data is commonly used to measure HSI in OT

Data can be collected using a single item in the Likert response format...

How easy was it to navigate the interface?

Very Difficult: 1 2 3 4 5
Very Easy: 6 7

...or using a Likert Scale (Likert, 1932; Likert & Hayes, 1957)
Likert Scales *typically* include 8 or more items

Furr & Bacharach (2014)
Testers disagree on appropriate analysis methods for Likert data
Roadmap

- Evaluating human-system interactions in OT
- The great debate
- Let the data decide
- Recommendations for analysts
- I stop talking
Not all numbers are created equal

Stevens (1946) proposed 4 levels of measurement

<table>
<thead>
<tr>
<th>Property of Numbers</th>
<th>Levels of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nominal</td>
</tr>
<tr>
<td>Identity</td>
<td>X</td>
</tr>
<tr>
<td>Order</td>
<td>X</td>
</tr>
<tr>
<td>Quantity</td>
<td></td>
</tr>
<tr>
<td>Rational Zero</td>
<td></td>
</tr>
</tbody>
</table>

Example

<table>
<thead>
<tr>
<th></th>
<th>Sex</th>
<th>Education</th>
<th>Memory</th>
<th>Behavior</th>
</tr>
</thead>
</table>

Researchers have criticized this classification system (Mitchell, 1986; Velleman & Wilkinson, 1993)

More nuanced classification systems exist (Chrisman, 1998; Mosteller & Tukey, 1977; van den Berg, 1991)
The *ordinal-ist* argument is grounded in Stevens’ levels of measurement

How easy was it to navigate the interface?

<table>
<thead>
<tr>
<th>Very Difficult</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Very Easy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B</td>
</tr>
</tbody>
</table>

*How do you know the distance between points is equal?*

A = B  OR  A ≠ B
Ordinal-ists argue you can’t guarantee distances are equal and thus, Likert data is ordinal.
Ordinal-ists argue that Likert data violate the assumptions of parametric tests.

They argue Likert data is not continuous or normally distributed.

Argument is easier to apply to single Likert items than Likert scales.
Normality assumption is often misunderstood

www.lock5stat.com
Normality assumption is often misunderstood

Mean from 1 sample

www.lock5stat.com
Normality assumption is often misunderstood

samples = 101
mean = 42.627
std. error = 5.923

www.lock5stat.com
Normality assumption is often misunderstood

www.lock5stat.com
Normality assumption is often misunderstood

The Sampling Distribution of Sample Means must be normally distributed

www.lock5stat.com
Ordinal-ists argue that we will experience higher error rates using parametric statistics on Likert data because of these violations.

(Nunnally, 1967; Jaimeson, 2004)
The Problem: Ordinal-ists are all theory and no data
Roadmap

• Evaluating human-system interactions in OT
• The great debate
• Let the data decide
• Recommendations for analysts
• I stop talking
The distance between scale points is an empirical question

Visual Analogue Scale (VAS) scores correspond to actual changes in intensity of stimuli and are widely recognized as interval data (Bolognese et al., 2003; Joyce et al., 1975; Myles et al., 1999; Price et al., 1983)
Evidence indicates the distance between scale points is roughly equal

Multiple studies have replicated this effect
(Baggaley & Hull, 1983; Carifio, 1976; Carifio 1978; Davey et al., 2007; Mauret & Pierce, 1998; Parker et al., 2002)
The effect of non-normality on error rates is an empirical question

The F-test is **robust** to violations of normality (Bartlett, 1937; Boneau, 1960; Box & Anderson, 1955; David & Johnson, 1951; Glass et al., 1972; Gombolay & Shah, 2016; Lindquist, 1953; Norton, 1952; Pearson, 1931)

Glass (1972) examined the effect of scale length on type I error rate in F-tests

The F-test controlled type I error rates for scales with at least 5 points

Skewness, kurtosis, and moderate heterogeneity of variance had little impact
We do not risk higher error rates when analyzing Likert data with the F-test or t-test
The effect of non-normality on error rates is an empirical question

The Pearson correlation is **robust** to violations of normality
(Pearson, 1931, 1932a, 1932b; Dunlap, 1931; Havelick & Peterson, 1976; Murray, 2013)

Norman (2010) asked participants to complete 8, 10-point Likert format questions on 2 occasions

Computed Pearson and Spearman correlations for responses across the 2 occasions

Predicted the Spearman correlation from the Pearson correlation

Pearson performed equivalent to Spearman even when data was severely non-normal
We do not risk higher error rates when analyzing Likert data with the Pearson correlation or Regression
Roadmap

• Evaluating human-system interactions in OT

• The great debate

• Let the data decide

• Recommendations for analysts

• I stop talking
Use **parametric statistics** to analyze your Likert data

Likert data approximates interval data

Greater power to detect an effect

Error rates are **not** higher
Roadmap

• Evaluating human-system interactions in OT

• The great debate

• Let the data decide

• Recommendations for analysts

• I stop talking
In science, data trumps theory

Questions?