Data Bootcamp Analysis

Katie Conger, data curtesy of Geeraert, Newman, & Baayen (2017)

March 30, 2023

## 1. Introduction

The purpose of this project is to demonstrate an ability to analyze data using R. To satisfy the requirements of this project, the portfolio must address three elements. The paper follows this order. Section 1 identifies my research question or hypothesis. Section 2 explains how data was obtained and analyzed using R. Finally, section 3 evaluates of the dataset with respect to one of the FAIR principles or best practices for data publishing more generally.

My dissertation addresses factors that impact idiom comprehension. An idiom is a phrase whose meaning cannot be derived by combining the literal meaning of its component parts (Hockett 1958, Katz 1973, Weinreich 1969, Pulman 1993, Marlies 1995, Mel’čuk 1995). Unlike literal phrases whose composite meaning results from combining the meaning of the individual words, all idioms are non-compositional because their phrasal meaning is not the sum of their parts. For example, cold turkey does not refer to “a low temperature bird” but instead to “the abrupt cessation of a habitual activity”. Because the meaning of an idiom is not the sum of its parts, all idioms are also formulaic. Thus, we must also know the conventionalized “formula” behind an idiomatic phrase, or which words must appear together to evoke the idiomatic meaning (Nunberg et al. 1994). The goal of this work is to investigate the roles of idiomaticity, noncompositionality and formulaicity in how we think about and understand idioms. Idioms are often discussed as if they form a discrete class, cleanly separated from non-idioms by their noncompositional nature and their conventionalized, formulaic meaning. However, the defining features of idioms, noncompositionality and formulaicity, are not unique to these phrases. Additionally, there is a great degree of variation within the class “idiom”, with some idioms seemingly more similar to non-idioms than other idioms. Thus, treatment of idioms as a discrete class may be problematic as there is little reason to assume that there is a binary distinction between a cleanly defined class “idiom” and a separate class “non-idiom”. Within linguistics, many have advocated for a prototype-based account of idioms, with more prototypically idiomatic phrases and prototypically literal phrases at opposite ends and less prototypical members of each class situated closer to the middle (cf. Penttila 2010).

Within psycholinguistics, clear delineation of the class “idiom” is a point of longstanding contention. With respect to how we think about and understand idioms, some theorize that all members of the class “idiom” are treated the same (cf. Bobrow & Bell 1973, Swinney & Cultler 1979, Cacciari & Tabossi 1988). Others theorize that similarities between idioms and other types of language, as well as differences within the class “idiom”, render a “one-size-fits-all” treatment impossible. However, little is known about which factors impact how we think about and comprehend idioms and whether there are interactions between any such features.

One such factor is “frequency”, which is operationalized in a number of ways, referring to phrasal predictability, subjective (perceived) frequency within specific communities of practice, pointwise mutual information (co-occurrence frequency). My personal research employs pointwise mutual information (PMI). Recently, I found relationships between a number of experimental factors and PMI, with strong methodological implications for future experimental work. While the scope of my work is necessarily limited to PMI, at least for the moment, this project offered an opportunity to examine whether similar relationships exist between subjective frequency, which is known for having several sub-constructs. One study collecting relevant data was conducted by Geeraert (2016), who investigated the relationship between idiom subjective frequency and idiom flexibility, or the degree to which a phrase can be modified without the idiomatic meaning being blocked (e.g., pay through the roof allows for lexical substitution, such as pay through the ceiling but kick the bucket allows for no such substitution \*kick the pail).

My goal with this project was to reexamine the data collected by Geeraert et al. (2016) to investigate the plausibility of underlying latent variables not necessarily identified or addressed in their initial analysis. More specifically, the analysis presented by Geeraert et al. (2016) focuses on correlations between idiom modifications and the willingness of participants to accept these idiom variations. However, their data allows for latent variable analysis of “frequency”, identifying constructs that may play into a cognitive conceptualization of a term that is not nearly as simple as simply counting single word occurrence.

In their study, Geeraert et al. (2016) collected a variety of ratings on idioms. Idioms were presented to subjects in one of seven forms: 1. Canonical: the expected, usual form of an idiom (e.g., Although these were new stocks, they suddenly went through the roof.)

1. Literal: a literal usage of an idiom that on its own is ambiguous as to whether one should interpret it literally or figuratively. (e.g., While the guys were re-shingling, they suddenly went through the roof.)
2. Lexical Variation: a word modification or addition made to the canonical condition form (e.g., Although these were new stocks, they suddenly went through the ceiling.)
3. Partial: replacement of an component of the canonical idiom form with a pronoun (e.g., Although these were new stocks, they suddenly went through it.)
4. Integrated Concept: a novel embellishment to a canonical form based on an underlying metaphor (e.g., Although these were new stocks, they suddenly went through the investment roof.) \*Of note: not all idioms were metaphorically based which makes this questionable.
5. Blend: two idioms combined into one (e.g., Although these were new stocks, they suddenly went through the charts. Blend: went through the [roof] and [off the] charts) (Geeraert et al. 2016) Ratings on a scale from 1-100 were collected for idiom acceptability, how often a participant perceived an idiom to be used by others, how often they personally used a given idiom, and how much they liked the idiom. Participants were also asked to indicate whether they knew an idiom or not using a categorical yes/no response. Reaction times for response items were also measured. Using mixed effects linear regression, Geeraert et al. (2016) found a significant correlation between types of idiom variations and acceptability. They used these results as evidence that idioms are not comprehended as phrasal collocations.

While their statistical methods seem sound, many assumptions underlie their conclusions. Of import for this analysis is the neglected role of probability, prediction, and/or familiarity across conditions. It is my belief that the variables of idiom acceptability and whether one knows an idiom or not should be indicative of the construct “frequency”, where “frequency”in regards to an idiomatic phrase is often conceptualized as a collapsed version of any of the following: cloze probability, prediction, entropy, surprisal, mutual information, and general probability. One may also expect an interaction between how well a participant knows an idiom, how often they have heard an idiom, how often they perceive an idiom to be used by others, and how much they like an idiom. This project will proceed in three stages with the goal of identifying any patterning of variance between these measures that could point to a reflection of the construct of “familiarity”.

## 2. R Code and Analysis Output

This section includes the R code and analysis outputs. A brief explanation of each step is included. The steps of the analysis are:

#### 2.1: Ordinal .RESP analysis

The first stage will look at ratings of acceptability, how often an idiom is personally used, and how often an idiom is perceived to be used by others. All responses use a 1-100 rating scale. Factor analysis will be performed to see if there is indeed a patterning a variance between these variables.

#### 2.2: Numeric reaction times to response questions

All response time factors will be compared to see if there is a patterning a variance between the amount of time required to respond to idioms with regards to acceptability, personal knowledge, personal use, use by others, and how well an idiom is liked. Factor analyses will be performed to see if there is a patterning of variance in response times. Importantly, the variable “known idiom” can be included in this analysis as all .RT measures are numeric whereas the “known idiom” response type was categoric, precluding it from inclusion in Stage 1. Reaction times to a yes/no question versus a scalar rating may still not be entirely comparable if only response time is compared. However, if one is looking for variance patterns versus simply significant mean differences, this is not a problem.

#### 2.3: Homogenized data types

The final stage will attempt to perform a principal component analysis on all factors by first homogenizing the data across types to make numeric, categoric, and ordinal responses comparative.

### 2.1: Ordinal .RESP analysis

knitr::opts\_chunk$set(echo = TRUE)  
  
library("tidyverse")  
library("lavaan")  
library("psych")  
library("corrplot")  
library("GPArotation")  
  
rm(list = ls())

#### Load the clean data

The idiom ratings are described in the paper by Geeraert, Newman, & Baayen (2017), uploaded with these files. Clearly defined explanations of each condition are available by request from the authors.

1. Subject
2. Age: Age of subject
3. Gender (two levels): female and male
4. Hand: (two levels): handedness of subject: left and right
5. NativeLang: (14 levels): subject’s native language
6. Idiom: The idiom
7. Condition: Type of idiom or phrase + Canonical + Conceptu + Blend + Lexical + Literal + Partial
8. AcceptRating.RESP: acceptability rating of phrase’s ability to be used literally (which corresponds in some degree to transparency)
9. KnowIdiom.RESP: (two levels): yes, no
10. HowOftenUse.RESP: how often the subject uses the idiom
11. HowOftenOthersUse.RESP: how often other people use the idiom
12. LikeUsingIdioms.RESP: how much the subject likes using the idiom

The idioms were rated on a 0-100 scale except KnowIdiom.RESP which has two levels, “yes” or “no”. The data set included here has 17 variables: Subject, Condition, Age, Gender and Handness. The item ratings names all end in .RESP and have corresponding reaction time measures that end in .RT.

# read in the clean data  
df <- read.csv("Idiom\_Data.csv")

#### Create data frames

#This selects just the rating response items  
response.items <- df %>%   
 select(ends\_with("RESP"))  
  
#For the purposes of this analysis, the subject column is not needed. This removes the subjects column from df.  
RInoSUB <-df %>%   
 select(-contains("Subject")) %>%  
 select(-contains("Expression"))   
   
#Transform categoric DF columns to numbers. This data frame will be used throughout the analysis.   
RI<-RInoSUB %>% ##Change from words to numbers for gender, language, ext.  
 mutate(Condition = case\_when(  
 Condition == "Literal" ~ 1,   
 Condition == "Canonical" ~ 2,  
 Condition == "Lexical" ~ 3,  
 Condition == "Partial" ~ 4,  
 Condition == "Blend" ~ 5,  
 Condition == "Concept" ~ 6)) %>%  
   
 mutate(KnowIdiom.RESP = case\_when(  
 KnowIdiom.RESP == "yes" ~ 1,   
 KnowIdiom.RESP == "no" ~ 2)) %>%  
  
#Note: Chinese, Cantonese, and Mandarin are collapsed since "Chinese" is the language family including Cantonese and Mandarin. What "Chinese" includes is unknown.   
 mutate(NativeLang = case\_when(  
 NativeLang == "English" ~ 1,   
 NativeLang == "Chinese" ~ 2,  
 NativeLang == "Mandarin" ~ 2,  
 NativeLang == "Cantonese" ~ 2,  
 NativeLang == "Turkish" ~ 3,  
 NativeLang == "Tagalog" ~ 4,  
 NativeLang == "Russian" ~ 5,  
 NativeLang == "Arabic" ~ 6,  
 NativeLang == "Polish" ~ 7,  
 NativeLang == "Vietnamese" ~ 9,  
 NativeLang == "Punjabi,English" ~ 8)) %>%  
   
   
#Though unlikely to matter, mutating handedness and gender for the sake of being thorough.   
   
 mutate(Gender = case\_when(  
 Gender == "female" ~ 1,   
 Gender == "male" ~ 2)) %>%  
   
 mutate(Hand = case\_when(  
 Hand == "right" ~ 1,   
 Hand == "left" ~ 2)) %>%  
  
   
#RI<-df %>% ##Adds a column for literal versus figurative of all forms. This could be used for later analysis but is not strictly necessary here.   
 mutate(Idiom\_Type = case\_when(  
 Condition == 1 ~ 1,   
 Condition == 2 ~ 2,  
 Condition == 3 ~ 2,  
 Condition == 4 ~ 2,  
 Condition == 5 ~ 2,  
 Condition == 6 ~ 2))

### Factor relationships

There is reason to believe that certain explored factors are not reflective of separate variables. Instead, it is likely that some, such as frequency, are also related to other factors. Here, I will explore relationships between factors not originally analyzed by Geeracts et al. (2017).

Notes: RI includes all variables, including idiom\_type. RInoSUB includes all variables except idiom\_type. RInoSUB allows for testings across parts of speech.

#### Basics: Getting a feel for the data

Basic information

describe(RI)

## vars n mean sd median trimmed mad min

## Condition 1 7887 2.54 1.67 2 2.32 1.48 1

## Age 2 7887 19.99 3.28 19 19.46 1.48 17

## Gender 3 7887 1.24 0.42 1 1.17 0.00 1

## Hand 4 7887 1.08 0.28 1 1.00 0.00 1

## NativeLang 5 7559 1.84 1.78 1 1.33 0.00 1

## AcceptRating.RESP 6 7887 59.51 33.53 68 61.57 40.03 0

## KnowIdiom.RESP 7 7887 1.32 0.47 1 1.28 0.00 1

## HowOftenUse.RESP 8 7887 54.66 28.83 67 56.64 25.20 0

## HowOftenOthersUse.RESP 9 7887 58.75 22.98 63 60.32 20.76 1

## LikeUsingIdioms.RESP 10 7887 68.13 24.91 70 71.14 25.20 0

## AcceptRating.RT 11 7887 6896.66 4052.44 5991 6287.19 2575.28 81

## KnowIdiom.RT 12 7887 1995.64 1345.45 1624 1758.92 644.93 404

## HowOftenUse.RT 13 7887 5085.24 2025.41 4559 4886.50 1771.71 1823

## HowOftenOthersUse.RT 14 7887 4064.40 1748.18 3453 3816.80 1135.67 1880

## LikeUsingIdioms.RT 15 7887 3047.91 1325.95 2543 2844.31 985.93 1591

## Idiom\_Type 16 7887 1.64 0.48 2 1.67 0.00 1

## max range skew kurtosis se

## Condition 6 5 0.86 -0.60 0.02

## Age 43 26 4.98 31.38 0.04

## Gender 2 1 1.24 -0.46 0.00

## Hand 2 1 3.02 7.11 0.00

## NativeLang 9 8 2.71 6.54 0.02

## AcceptRating.RESP 100 100 -0.39 -1.27 0.38

## KnowIdiom.RESP 2 1 0.76 -1.43 0.01

## HowOftenUse.RESP 100 100 -0.58 -0.96 0.32

## HowOftenOthersUse.RESP 100 99 -0.57 -0.09 0.26

## LikeUsingIdioms.RESP 100 100 -0.93 0.41 0.28

## AcceptRating.RT 73971 73890 3.70 29.68 45.63

## KnowIdiom.RT 37577 37173 5.74 84.58 15.15

## HowOftenUse.RT 13105 11282 1.19 2.14 22.81

## HowOftenOthersUse.RT 8889 7009 1.23 0.64 19.68

## LikeUsingIdioms.RT 8349 6758 1.47 2.34 14.93

## Idiom\_Type 2 1 -0.57 -1.68 0.01

summary(RI)

## Condition Age Gender Hand

## Min. :1.000 Min. :17.00 Min. :1.000 Min. :1.000

## 1st Qu.:1.000 1st Qu.:18.00 1st Qu.:1.000 1st Qu.:1.000

## Median :2.000 Median :19.00 Median :1.000 Median :1.000

## Mean :2.545 Mean :19.99 Mean :1.236 Mean :1.083

## 3rd Qu.:4.000 3rd Qu.:21.00 3rd Qu.:1.000 3rd Qu.:1.000

## Max. :6.000 Max. :43.00 Max. :2.000 Max. :2.000

##

## NativeLang AcceptRating.RESP KnowIdiom.RESP HowOftenUse.RESP

## Min. :1.000 Min. : 0.00 Min. :1.000 Min. : 0.00

## 1st Qu.:1.000 1st Qu.: 28.00 1st Qu.:1.000 1st Qu.: 27.50

## Median :1.000 Median : 68.00 Median :1.000 Median : 67.00

## Mean :1.842 Mean : 59.51 Mean :1.323 Mean : 54.66

## 3rd Qu.:2.000 3rd Qu.: 91.00 3rd Qu.:2.000 3rd Qu.: 77.00

## Max. :9.000 Max. :100.00 Max. :2.000 Max. :100.00

## NA's :328

## HowOftenOthersUse.RESP LikeUsingIdioms.RESP AcceptRating.RT KnowIdiom.RT

## Min. : 1.00 Min. : 0.00 Min. : 81 Min. : 404

## 1st Qu.: 46.00 1st Qu.: 56.00 1st Qu.: 4462 1st Qu.: 1268

## Median : 63.00 Median : 70.00 Median : 5991 Median : 1624

## Mean : 58.75 Mean : 68.13 Mean : 6897 Mean : 1996

## 3rd Qu.: 77.00 3rd Qu.: 89.00 3rd Qu.: 8104 3rd Qu.: 2257

## Max. :100.00 Max. :100.00 Max. :73971 Max. :37577

##

## HowOftenUse.RT HowOftenOthersUse.RT LikeUsingIdioms.RT Idiom\_Type

## Min. : 1823 Min. :1880 Min. :1591 Min. :1.000

## 1st Qu.: 3669 1st Qu.:3001 1st Qu.:2099 1st Qu.:1.000

## Median : 4559 Median :3453 Median :2543 Median :2.000

## Mean : 5085 Mean :4064 Mean :3048 Mean :1.636

## 3rd Qu.: 6170 3rd Qu.:4626 3rd Qu.:3636 3rd Qu.:2.000

## Max. :13105 Max. :8889 Max. :8349 Max. :2.000

##

describe(RInoSUB)

## vars n mean sd median trimmed mad min

## Condition\* 1 7887 3.64 1.61 4 3.67 1.48 1

## Age 2 7887 19.99 3.28 19 19.46 1.48 17

## Gender\* 3 7887 1.24 0.42 1 1.17 0.00 1

## Hand\* 4 7887 1.92 0.28 2 2.00 0.00 1

## NativeLang\* 5 7887 5.00 2.55 4 4.53 0.00 1

## AcceptRating.RESP 6 7887 59.51 33.53 68 61.57 40.03 0

## KnowIdiom.RESP\* 7 7887 1.68 0.47 2 1.72 0.00 1

## HowOftenUse.RESP 8 7887 54.66 28.83 67 56.64 25.20 0

## HowOftenOthersUse.RESP 9 7887 58.75 22.98 63 60.32 20.76 1

## LikeUsingIdioms.RESP 10 7887 68.13 24.91 70 71.14 25.20 0

## AcceptRating.RT 11 7887 6896.66 4052.44 5991 6287.19 2575.28 81

## KnowIdiom.RT 12 7887 1995.64 1345.45 1624 1758.92 644.93 404

## HowOftenUse.RT 13 7887 5085.24 2025.41 4559 4886.50 1771.71 1823

## HowOftenOthersUse.RT 14 7887 4064.40 1748.18 3453 3816.80 1135.67 1880

## LikeUsingIdioms.RT 15 7887 3047.91 1325.95 2543 2844.31 985.93 1591

## max range skew kurtosis se

## Condition\* 6 5 -0.18 -1.45 0.02

## Age 43 26 4.98 31.38 0.04

## Gender\* 2 1 1.24 -0.46 0.00

## Hand\* 2 1 -3.02 7.11 0.00

## NativeLang\* 14 13 1.78 2.61 0.03

## AcceptRating.RESP 100 100 -0.39 -1.27 0.38

## KnowIdiom.RESP\* 2 1 -0.76 -1.43 0.01

## HowOftenUse.RESP 100 100 -0.58 -0.96 0.32

## HowOftenOthersUse.RESP 100 99 -0.57 -0.09 0.26

## LikeUsingIdioms.RESP 100 100 -0.93 0.41 0.28

## AcceptRating.RT 73971 73890 3.70 29.68 45.63

## KnowIdiom.RT 37577 37173 5.74 84.58 15.15

## HowOftenUse.RT 13105 11282 1.19 2.14 22.81

## HowOftenOthersUse.RT 8889 7009 1.23 0.64 19.68

## LikeUsingIdioms.RT 8349 6758 1.47 2.34 14.93

summary(RInoSUB)

## Condition Age Gender Hand

## Length:7887 Min. :17.00 Length:7887 Length:7887

## Class :character 1st Qu.:18.00 Class :character Class :character

## Mode :character Median :19.00 Mode :character Mode :character

## Mean :19.99

## 3rd Qu.:21.00

## Max. :43.00

## NativeLang AcceptRating.RESP KnowIdiom.RESP HowOftenUse.RESP

## Length:7887 Min. : 0.00 Length:7887 Min. : 0.00

## Class :character 1st Qu.: 28.00 Class :character 1st Qu.: 27.50

## Mode :character Median : 68.00 Mode :character Median : 67.00

## Mean : 59.51 Mean : 54.66

## 3rd Qu.: 91.00 3rd Qu.: 77.00

## Max. :100.00 Max. :100.00

## HowOftenOthersUse.RESP LikeUsingIdioms.RESP AcceptRating.RT KnowIdiom.RT

## Min. : 1.00 Min. : 0.00 Min. : 81 Min. : 404

## 1st Qu.: 46.00 1st Qu.: 56.00 1st Qu.: 4462 1st Qu.: 1268

## Median : 63.00 Median : 70.00 Median : 5991 Median : 1624

## Mean : 58.75 Mean : 68.13 Mean : 6897 Mean : 1996

## 3rd Qu.: 77.00 3rd Qu.: 89.00 3rd Qu.: 8104 3rd Qu.: 2257

## Max. :100.00 Max. :100.00 Max. :73971 Max. :37577

## HowOftenUse.RT HowOftenOthersUse.RT LikeUsingIdioms.RT

## Min. : 1823 Min. :1880 Min. :1591

## 1st Qu.: 3669 1st Qu.:3001 1st Qu.:2099

## Median : 4559 Median :3453 Median :2543

## Mean : 5085 Mean :4064 Mean :3048

## 3rd Qu.: 6170 3rd Qu.:4626 3rd Qu.:3636

## Max. :13105 Max. :8889 Max. :8349

#### Exploratory Factor Analysis (unidimensional)

Scores in FA are an empirically weighted sums of item scores. The weights are determined by the correlation between items.

In this section, RInoKnow is introduced. It includes only the 4 response variables: acceptability ratings, how often a participant believes they use a phrase, how often a participant believes others use a phrase, and whether they “like” using the phrase (whether it sounds “correct” to them). This is the same as RInoSUB except that demographic information has been removed such that only the variables of direct interest remain.

EFA 1: Explores factors for ordinal response items only

*#1.) Removes categorical KnownIdiom.RESP*

RInoKnow <-response.items %>%

select(-contains("Know"))

*# Unidimensional EFA of RESP items*

EFA\_Model1<-fa(RInoKnow, nfactors = 2)

*# View the factor loadings*

EFA\_Model1$loadings

##

## Loadings:

## MR1 MR2

## AcceptRating.RESP 0.130

## HowOftenUse.RESP 0.929

## HowOftenOthersUse.RESP 0.719 0.187

## LikeUsingIdioms.RESP 0.680

##

## MR1 MR2

## SS loadings 1.843 0.059

## Proportion Var 0.461 0.015

## Cumulative Var 0.461 0.476

*# Create a path diagram of the items' factor loadings*

fa.diagram(EFA\_Model1)

*Diagram

Description automatically generated*

*# Loading on EFA model 1*

print(EFA\_Model1$loadings, cutoff = 0.2)

##

## Loadings:

## MR1 MR2

## AcceptRating.RESP

## HowOftenUse.RESP 0.929

## HowOftenOthersUse.RESP 0.719

## LikeUsingIdioms.RESP 0.680

##

## MR1 MR2

## SS loadings 1.843 0.059

## Proportion Var 0.461 0.015

## Cumulative Var 0.461 0.476

*# factor plot for EFA model 1*

factor.plot(EFA\_Model1, title = "Unrotated Factor Loadings for Responses")

*Chart

Description automatically generated*Above results: A strong common pattern of variance is seen between the ratings of how often a participant uses an idiom, how often they perceive others use an idiom, and how much they like an idiom. Factor 1 accounts for 93% of the variance of Howoftenuse.RESP, 71.9% of howoftenothersuse.RESP, and 68% of Likeuseingidioms.RESP. Frequency of idiom exposure appears to be the underlying factor uniting these measures. AcceptRating.RESP does not converge or correlate with other factors.

#### Communalities for EFA 1 of RESP items

This shows the extent to which all items corrolate.

round(EFA\_Model1$communality, 2)

## AcceptRating.RESP HowOftenUse.RESP HowOftenOthersUse.RESP

## 0.02 0.85 0.60

## LikeUsingIdioms.RESP

## 0.45

#### 

#### Rotate the Loadings

There is only one significant factor so rotation is not necessary. From this, frequency is a theoretically sound explanation for the loading of factors on item 1 and was expected.

EFA\_Model1r <- fa(EFA\_Model1$loadings, nfactors = 2, rotate = "varimax", fm = "ml")

factor.plot(EFA\_Model1r, title = "Varimax Rotated Factor Loadings")

Chart, histogram

Description automatically generated

#### FA of RESP

*# use fa from psych package*

RInoKnown\_FA <- fa(RInoKnow, nfactors = 2, rotate = "none", fm = "ml")

*# rotated solution*

RInoKnown\_FaRot <- fa(RInoKnow, nfactors = 2, rotate = "varimax", fm = "ml")

factor.plot(RInoKnown\_FA, title = "FA of RESP Unrotated")

Chart, box and whisker chart

Description automatically generated

factor.plot(RInoKnown\_FaRot, title = "FA of RESP Rotated")

Chart, scatter chart

Description automatically generated

#### Split the data

# Establishing two sets of indices to split the dataset  
N <- nrow(RI)  
indices <- seq(1,N)  
RInoKnown\_indices\_EFA <- sample(indices, floor((.5\*N)))  
RInoKnown\_indices\_CFA <- indices[!(indices %in% RInoKnown\_indices\_EFA)]  
  
  
# Using indices to split the dataset into halves for a EFA and CFA  
RInoKnown\_EFA <- RInoKnow[RInoKnown\_indices\_EFA, ]  
RInoKnown\_CFA <- RInoKnow[RInoKnown\_indices\_CFA, ]

#### Bayesian exploratory analysis: used as a check to see if a different method yields similar results

**library**("MPsychoR")

**library**("corrplot")

**library**("BayesFM")

Privstd <- scale(RInoKnow)

corrplot(cor(RInoKnow))

Chart, bubble chart

Description automatically generated

### 2.2 Numeric reaction times to response questions

#### EFA 2: Explores RT scores

EFA was used to explore the loadings of RT. This was done first with no specified factor number and again with a factor of 3 (which was judged by R to be an appropriate number of factors to use

#### EFA with one factor (unspecified factor number)

*#Select RT items only*

RT <- RI %>%

select(contains(".RT"))

*# Unidimensional EFA with one factor*

EFA\_Model3 <- fa(RT)

*# View the factor loadings*

EFA\_Model3$loadings

##

## Loadings:

## MR1

## AcceptRating.RT 0.342

## KnowIdiom.RT 0.301

## HowOftenUse.RT 0.517

## HowOftenOthersUse.RT 0.486

## LikeUsingIdioms.RT 0.767

##

## MR1

## SS loadings 1.299

## Proportion Var 0.260

*# Create a path diagram of the items' factor loadings*

fa.diagram(EFA\_Model3)

Diagram

Description automatically generated

*# Loading on EFA model 2*

print(EFA\_Model3$loadings, cutoff = 0.2)

##

## Loadings:

## MR1

## AcceptRating.RT 0.342

## KnowIdiom.RT 0.301

## HowOftenUse.RT 0.517

## HowOftenOthersUse.RT 0.486

## LikeUsingIdioms.RT 0.767

##

## MR1

## SS loadings 1.299

## Proportion Var 0.260

*## factor plot for EFA model 2*

factor.plot(EFA\_Model3, title = "Unrotated Factor Loadings for RT with No Specified Factor Number")

Chart

Description automatically generated

Using a single factor, RTs to how well liked an idiom is and how often an idiom is personally used load onto the same factor. This is not the case in the 3 factor analysis in 2.2.2.

#### Communalities for EFA 3

round(EFA\_Model3$communality, 1)

## AcceptRating.RT KnowIdiom.RT HowOftenUse.RT

## 0.1 0.1 0.3

## HowOftenOthersUse.RT LikeUsingIdioms.RT

## 0.2 0.6

#### Rotate the Loadings

This is unnecessary due to the fact that there is only one factor.

*# rotated solution*

EFA\_Model3r <- fa(RT, rotate = "varimax", fm = "ml")

factor.plot(EFA\_Model3, title = "Single Factor RT Unrotated")

Chart

Description automatically generated

factor.plot(EFA\_Model3r, title = "Single Factor RT Rotated")

Chart

Description automatically generated

#### Split practice

# Establish two sets of indices to split the dataset  
N <- nrow(RI)  
indices <- seq(1,N)  
RT\_indices\_EFA <- sample(indices, floor((.5\*N)))  
RT\_indices\_CFA <- indices[!(indices %in% RT\_indices\_EFA)]  
  
# Use those indices to split the dataset into halves for your EFA and CFA  
RT\_EFA <- RT[RT\_indices\_EFA, ]  
RT\_CFA <- RT[RT\_indices\_CFA, ]

#### This chunk shows that there is only 1 significant shared patterning of variance.

# Calculate the correlation matrix first (Done earlier, but not for split)  
RT\_EFA\_cor <- cor(RT\_EFA, use = "pairwise.complete.obs")  
  
# Then use that correlation matrix to calculate eigenvalues  
eigenvals <- eigen(RT\_EFA\_cor)  
  
# Look at the eigenvalues returned  
eigenvals$values

## [1] 1.9757445 0.9136322 0.8471516 0.8003871 0.4630846

#### Below scree plot shows 1 significant factor.

*# Calculate the correlation matrix first*

RT\_EFA\_cor <- cor(RT\_EFA, use = "pairwise.complete.obs")

*# Then use that correlation matrix to create the scree plot*

scree(RT\_EFA\_cor, factors = FALSE)

Chart, line chart

Description automatically generated

### EFA with 3 factor

*#Select RT items only*

RT <- RI %>%

select(contains(".RT"))

*#Unidimensional EFA with three factors to compare the difference*

EFA\_Model3.3<-fa(RT, nfactors = 3)

*# View the factor loadings*

EFA\_Model3.3$loadings

##

## Loadings:

## MR1 MR2 MR3

## AcceptRating.RT 0.192 0.413

## KnowIdiom.RT 0.128 0.124 0.269

## HowOftenUse.RT 0.683

## HowOftenOthersUse.RT 0.681 -0.110

## LikeUsingIdioms.RT 0.533 0.383

##

## MR1 MR2 MR3

## SS loadings 0.773 0.679 0.265

## Proportion Var 0.155 0.136 0.053

## Cumulative Var 0.155 0.290 0.343

*# Create a path diagram of the items' factor loadings*

fa.diagram(EFA\_Model3.3)

Diagram

Description automatically generated

*# Loading on EFA model 2*

print(EFA\_Model3.3$loadings, cutoff = 0.2)

##

## Loadings:

## MR1 MR2 MR3

## AcceptRating.RT 0.413

## KnowIdiom.RT 0.269

## HowOftenUse.RT 0.683

## HowOftenOthersUse.RT 0.681

## LikeUsingIdioms.RT 0.533 0.383

##

## MR1 MR2 MR3

## SS loadings 0.773 0.679 0.265

## Proportion Var 0.155 0.136 0.053

## Cumulative Var 0.155 0.290 0.343

*## factor plot for EFA model 2*

factor.plot(EFA\_Model3.3, title = "Unrotated Factor Loadings for RT with 3 Factors")

Chart, scatter chart

Description automatically generatedBased on the analysis with 3 factors, RTs for how often others use an idiom and how well liked an idiom is load onto one factor. The relationship between these two and how often an idiom is used is quite different than the single factor loading in 2.2.1.

#### Split data and set up matrix for 3 factor rotation

# Establish two sets of indices to split the dataset  
N <- nrow(RI)  
indices <- seq(1,N)  
RT3\_indices\_EFA <- sample(indices, floor((.5\*N)))  
RT3\_indices\_CFA <- indices[!(indices %in% RT3\_indices\_EFA)]  
  
# Use those indices to split the dataset into halves for your EFA and CFA  
RT3\_EFA <- RT[RT3\_indices\_EFA, ]  
RT3\_CFA <- RT[RT3\_indices\_CFA, ]

#### This shows that there is only 1 significant and one marginally significant shared patterning of variance even with 3 proposed factor loadings.

# Calculate the correlation matrix first (Done earlier, but not for split)  
RT3\_EFA\_cor <- cor(RT3\_EFA, use = "pairwise.complete.obs")  
  
# Then use that correlation matrix to calculate eigenvalues  
eigenvals <- eigen(RT3\_EFA\_cor)  
  
# Look at the eigenvalues returned  
eigenvals$values

## [1] 1.9444794 0.9241287 0.8463243 0.8125261 0.4725415

#### Below Scree plot shows 1 significant factor and one marginally significant.

# Calculate the correlation matrix first  
RT3\_EFA\_cor <- cor(RT3\_EFA, use = "pairwise.complete.obs")  
  
# Then use that correlation matrix to create the scree plot  
scree(RT3\_EFA\_cor, factors = FALSE)

Chart, line chart

Description automatically generated

#### Bayesian Exploratory: used as a check to see if a different method yields similar results

As seen with the EFA, there is a patterning of variance between reaction times when responding to questions about how well liked an idiom is and how often an idiom is used There is also a marginally significant patterning shared by how often an idiom is perceived to be used by others. Interestingly, response times for how well known an idiom is personally and how acceptable a participant judged a given variation do not load on the same factor. This appears to clear up some of the confusion regarding loading of how well an idiom is liked, how often others use it, and how often a given participant uses an idiom.

Privstd2 <- scale(RT)  
corrplot(cor(RT))

Chart, bubble chart

Description automatically generated

## 2.3 Homogenized data types

The analysis in 2.3 is not valid as all data is treated as if it were of one type. My goal was simply to gain a better understanding of the impact that not accounting for data type would have. If this were reflective of actual patterning, it appears that 9 factors are needed to account for the variance between the 16 factors. As expected, the responses to how often an idiom is used, how often others use it, and marginally, how well liked an idiom is are still shown as likely measuring the same item. There is a relationship between condition and idiom type. While expected, the fact that they aren’t identical when this is simply a recoding of the same category is interesting. Literal uses of idioms are often criticized as invalid comparisons due to a.) a different cognitive meaning computation process for literal versus figurative words and b.) because it is claimed that one will prioritize the figurative interpretation over the literal. In reality, I believe this is due more to the awkwardness of many literal phrases, such as “I was watering the plant and thought I heard it through the grapevine that it needed more water.” Further investigation of the different types of idioms within condition may be interesting. Response times for how often others use an idiom and how well liked an idiom is appear to be more related than how often one personally uses an idiom. This could be reflective of a longer thought process when evaluating personal usage frequency but would also be interesting to delve deeper into (…at least until you open the can of worms that is RT validity in a post-perceptual test.)

### 2.3.1 EFA of all data:

#### FA of RI, including categoric, numeric, ordinal data all in number format but without accounting for differing variable types

*# Unidimensional FA*

*# use fa from psych package*

RI\_all\_FA <- fa(RI, nfactors = 2, rotate = "none", fm = "ml")

*# View the factor loadings*

RI\_all\_FA$loadings

##

## Loadings:

## ML2 ML1

## Condition 0.701

## Age

## Gender

## Hand 0.132

## NativeLang

## AcceptRating.RESP

## KnowIdiom.RESP -0.229

## HowOftenUse.RESP 0.872

## HowOftenOthersUse.RESP 0.735

## LikeUsingIdioms.RESP 0.707

## AcceptRating.RT -0.289

## KnowIdiom.RT -0.253

## HowOftenUse.RT -0.532

## HowOftenOthersUse.RT -0.187

## LikeUsingIdioms.RT -0.279

## Idiom\_Type 0.997

##

## ML2 ML1

## SS loadings 2.417 1.495

## Proportion Var 0.151 0.093

## Cumulative Var 0.151 0.245

*# Create a path diagram of the items' factor loadings*

fa.diagram(RI\_all\_FA)

Diagram, schematic

Description automatically generated

*# Loading on EFA model 2*

print(RI\_all\_FA$loadings, cutoff = 0.2)

##

## Loadings:

## ML2 ML1

## Condition 0.701

## Age

## Gender

## Hand

## NativeLang

## AcceptRating.RESP

## KnowIdiom.RESP -0.229

## HowOftenUse.RESP 0.872

## HowOftenOthersUse.RESP 0.735

## LikeUsingIdioms.RESP 0.707

## AcceptRating.RT -0.289

## KnowIdiom.RT -0.253

## HowOftenUse.RT -0.532

## HowOftenOthersUse.RT

## LikeUsingIdioms.RT -0.279

## Idiom\_Type 0.997

##

## ML2 ML1

## SS loadings 2.417 1.495

## Proportion Var 0.151 0.093

## Cumulative Var 0.151 0.245

#### Rotate the Loadings

*# rotated solution*

RI\_all\_FaRot <- fa(RI, nfactors = 2, rotate = "varimax", fm = "ml")

factor.plot(RI\_all\_FA, title = "All factors Unrotated")

Chart, box and whisker chart

Description automatically generated

factor.plot(RI\_all\_FaRot, title = "All factors Rotated")

Chart, scatter chart

Description automatically generated

#### Set up matrix to check Eigen values

# Establish two sets of indices to split the dataset  
N <- nrow(RI)  
indices <- seq(1,N)  
RI\_All\_indices\_EFA <- sample(indices, floor((.5\*N)))  
RI\_All\_indices\_CFA <- indices[!(indices %in% RI\_All\_indices\_EFA)]  
  
# Use those indices to split the dataset into halves for your EFA and CFA  
RI\_All\_EFA <- RI[RI\_All\_indices\_EFA, ]  
RI\_All\_CFA <- RI[RI\_All\_indices\_CFA, ]

#### This shows that there are 6 meaningful factors of 16.

# Calculate the correlation matrix first (Done earlier, but not for split)  
RI\_All\_cor <- cor(RI\_All\_EFA, use = "pairwise.complete.obs")  
  
# Then use that correlation matrix to calculate eigenvalues  
eigenvals <- eigen(RI\_All\_cor)  
  
# Look at the eigenvalues returned  
eigenvals$values

## [1] 3.0654506 1.7087025 1.4024927 1.3730354 1.2084265 1.1548508 0.9558851

## [8] 0.9028731 0.8346936 0.7928097 0.6943735 0.5781750 0.4098849 0.4032459

## [15] 0.2909324 0.2241683

#### Scree plot show 6 significant factors

# Calculate the correlation matrix first  
RI\_All\_cor <- cor(RI\_All\_EFA, use = "pairwise.complete.obs")  
  
# Then use that correlation matrix to create the scree plot  
scree(RI\_All\_cor, factors = FALSE)

Chart, line chart

Description automatically generated

#### Bayesian Exploratory

Bayesian exploratory analysis conducted to determine whether a different analytic method yields similar results. This analysis uses a mixed corrplot in recognition of the differing types of data.

Privstd2 <- scale(RI)  
corrplot.mixed(cor(RI))

Chart, scatter chart

Description automatically generated

#### 

#### Bayesian information criterion

This returns the Bayseian information criterion (BIC) as a test of relative fit. The value with 6 factors is very high. This would lead me to believe that 6 factors are not sufficient to explain the 16 categories of RI as the Eigen values from the FA indicated. The lowest BIC is returned using 9 factors.

# Run each theorized EFA on your dataset  
RI\_All\_theory <- fa(RI\_All\_EFA, nfactors = 9)  
RI\_All\_eigen <- fa(RI\_All\_EFA, nfactors = 6)  
  
*# Compare the BIC values*

RI\_All\_theory$BIC

## [1] -63.25974

RI\_All\_eigen$BIC

## [1] 640.3619

#### Interval Principal Components of Acceptability RT and Known Idiom RT

It is expected that these factors should correlate. The results below show that reaction times for acceptability and known idiom do converge.

# Extract "Acceptability RT" (st) and "KnowIdiom RT" pg subscales  
library(Gifi)  
  
ART <- RI %>% select(c(11))  
KRT <- RI %>% select(c(12))  
AKRT <- data.frame(ART = ART, KRT = KRT)  
# perform a standard PCA with prcomp()  
pcafit <- prcomp(AKRT, scale = TRUE)  
  
  
# Interval RT spline knots for a linear fit  
knotslin <- knotsGifi(AKRT, type = "E")

prlin <- princals(AKRT,   
 ndim = 2,  
 knots = knotslin,   
 degrees = 1,  
 ordinal = TRUE)

#### Interval Principal Components of Acceptability REST and other known REST

I believe that these factors are a measure of something similar but the EFA Eigen value did not show them to quite reach significance. This is a verification of that finding.

The PCA below makes it appear as though these variables do load together.

# Extract "Acceptability RT" (st) and "KnowIdiom RT" pg subscales  
  
AccRSP <- RI %>% select(c(6))  
HyouURSP <- RI %>% select(c(8))  
ARSP <- data.frame(AccRSP = AccRSP, HyouURSP = HyouURSP)  
# perform a standard PCA with prcomp()  
pcafit <- prcomp(ARSP, scale = TRUE)  
  
  
# Interval RT spline knots for a linear fit  
knotslin <- knotsGifi(ARSP, type = "D")  
  
# do optimal scaling with princals() for 2-dimensions  
prlin <- princals(ARSP, ndim = 2,  
 knots = knotslin,   
 degrees = 1,  
 ordinal = TRUE)  
prlin

## Call:

## princals(data = ARSP, ndim = 2, ordinal = TRUE, knots = knotslin,

## degrees = 1)

##

## Loss value: 0.5

## Number of iterations: 3

##

## Eigenvalues: 1.008 0.992

### 

### Princals on Mixed Input Data

This section sets up the data in RI to perform a valid analysis by taking the data type into account.

#### Treat some as ordinal and some as metric

# 1.) Ordinal columns: acceptability rating, how often used response, how often others use response, like use response,   
  
# Add previously defined Acceptability rating (ARSP)  
# Add previously how often used response (HyouURSP)   
  
# Add how others use response (HotherRESP)  
HotherRESP <- RI %>% select(c(9))   
# Adds how must they like using an idiom (LikeRSP)  
LikeRSP <- RI %>% select(c(10))  
  
Ord <- data.frame(AccRSP, HyouURSP, HotherRESP, LikeRSP)  
  
knotsord <- knotsGifi(  
 Ord%>% select(c("AcceptRating.RESP", "HowOftenUse.RESP", "HowOftenOthersUse.RESP", "LikeUsingIdioms.RESP")),  
 type = "D")

#2.) Metric: Numeric columns: Age, all RT responses  
  
Age <- RI %>% select(c(2))   
#AccRT <- RI %>% select(c(11))#redundant  
#KnowRT <- RI %>% select(c(12)) #redundant  
HyouRT <- RI %>% select(c(13))  
HotherRT <- RI %>% select(c(14)) #why row sum  
LikeRT <- RI %>% select(c(15))  
#na <- rowSums(ASTI %>% select(c(3, 6, 8, 12)))  
  
  
Num <- data.frame(Age, ART, KRT, HyouRT, HotherRT, LikeRT)  
  
knotsnum <- knotsGifi(  
 Num %>% select(c("Age", "KnowIdiom.RT", "HowOftenUse.RT", "HowOftenOthersUse.RT", "LikeUsingIdioms.RT")),   
 type = "E")

#3.) Categoric colums: Condition, gender, hand, native language, known idiom,   
  
Cond <- RI %>% select(c(1))   
Gen <- RI %>% select(c(3))  
Hand <- RI %>% select(c(4))   
NatLang <- RI %>% select(c(5))  
Known <- RI %>% select(c(7))   
  
Cat <- data.frame(Cond, Gen, Hand, NatLang, Known)

# combine the two sets of knots into a final list  
knotslist <- c(knotsord, knotslin)

### 

### 2.4 Additional tests

These were not necessarily related to the question regarding frequency addressed by analyzing RESP and RT for acceptability, use, and how well liked an idiom is. They are however interesting and were helpful in learning to set up the above tests.

#### Factor Analysis

The use of non-native English speakers was HIGHLY unusual as native language intuition is relied upon for figurative language judgments even more so than for literal language judgments. The following analyses were part of my code learning process but would be interesting to look into further.

This looks at only the partial response condition

*#Looks at the partial idiom condition responses only*

ConditionF <- RI %>%

filter(Condition == "4")

ConditionF\_RSP <- RI %>%

select(ends\_with("RESP"))

fa(ConditionF\_RSP)

## Factor Analysis using method = minres

## Call: fa(r = ConditionF\_RSP)

## Standardized loadings (pattern matrix) based upon correlation matrix

## MR1 h2 u2 com

## AcceptRating.RESP 0.05 0.002 1.00 1

## KnowIdiom.RESP -0.21 0.045 0.96 1

## HowOftenUse.RESP 0.92 0.848 0.15 1

## HowOftenOthersUse.RESP 0.74 0.550 0.45 1

## LikeUsingIdioms.RESP 0.67 0.444 0.56 1

##

## MR1

## SS loadings 1.89

## Proportion Var 0.38

##

## Mean item complexity = 1

## Test of the hypothesis that 1 factor is sufficient.

##

## df null model = 10 with the objective function = 1.21 with Chi Square = 9554.26

## df of the model are 5 and the objective function was 0.05

##

## The root mean square of the residuals (RMSR) is 0.07

## The df corrected root mean square of the residuals is 0.09

##

## The harmonic n.obs is 7887 with the empirical chi square 689.1 with prob < 1.1e-146

## The total n.obs was 7887 with Likelihood Chi Square = 390.92 with prob < 2.7e-82

##

## Tucker Lewis Index of factoring reliability = 0.919

## RMSEA index = 0.099 and the 90 % confidence intervals are 0.091 0.107

## BIC = 346.06

## Fit based upon off diagonal values = 0.96

## Measures of factor score adequacy

## MR1

## Correlation of (regression) scores with factors 0.94

## Multiple R square of scores with factors 0.88

## Minimum correlation of possible factor scores 0.77

#Looks at English Native speakers’ responses only

SpeakerE <- RI %>%

filter(NativeLang == "1")

English\_RSP <- SpeakerE %>%

select(ends\_with("RESP"))

fa(English\_RSP)

## Factor Analysis using method = minres

## Call: fa(r = English\_RSP)

## Standardized loadings (pattern matrix) based upon correlation matrix

## MR1 h2 u2 com

## AcceptRating.RESP 0.02 0.00059 1.00 1

## KnowIdiom.RESP -0.07 0.00468 1.00 1

## HowOftenUse.RESP 0.91 0.82496 0.18 1

## HowOftenOthersUse.RESP 0.61 0.36890 0.63 1

## LikeUsingIdioms.RESP 0.68 0.46200 0.54 1

##

## MR1

## SS loadings 1.66

## Proportion Var 0.33

##

## Mean item complexity = 1

## Test of the hypothesis that 1 factor is sufficient.

##

## df null model = 10 with the objective function = 0.91 with Chi Square = 4470.25

## df of the model are 5 and the objective function was 0.05

##

## The root mean square of the residuals (RMSR) is 0.07

## The df corrected root mean square of the residuals is 0.09

##

## The harmonic n.obs is 4929 with the empirical chi square 438.51 with prob < 1.5e-92

## The total n.obs was 4929 with Likelihood Chi Square = 232.46 with prob < 3.2e-48

##

## Tucker Lewis Index of factoring reliability = 0.898

## RMSEA index = 0.096 and the 90 % confidence intervals are 0.086 0.107

## BIC = 189.95

## Fit based upon off diagonal values = 0.95

## Measures of factor score adequacy

## MR1

## Correlation of (regression) scores with factors 0.93

## Multiple R square of scores with factors 0.86

## Minimum correlation of possible factor scores 0.72

#4.1.2 This looks only at Chinese speakers’ responses

#Looks at Chinese Native speakers' responses only

SpeakerC <- RI %>%

filter(NativeLang == "2")

Chinese\_RSP <- SpeakerC %>%

select(ends\_with("RESP"))

fa(Chinese\_RSP)

## Factor Analysis using method = minres

## Call: fa(r = Chinese\_RSP)

## Standardized loadings (pattern matrix) based upon correlation matrix

## MR1 h2 u2 com

## AcceptRating.RESP -0.04 0.0013 1.00 1

## KnowIdiom.RESP -0.08 0.0065 0.99 1

## HowOftenUse.RESP 0.91 0.8242 0.18 1

## HowOftenOthersUse.RESP 0.62 0.3879 0.61 1

## LikeUsingIdioms.RESP 0.23 0.0511 0.95 1

##

## MR1

## SS loadings 1.27

## Proportion Var 0.25

##

## Mean item complexity = 1

## Test of the hypothesis that 1 factor is sufficient.

##

## df null model = 10 with the objective function = 0.49 with Chi Square = 857.97

## df of the model are 5 and the objective function was 0.05

##

## The root mean square of the residuals (RMSR) is 0.07

## The df corrected root mean square of the residuals is 0.09

##

## The harmonic n.obs is 1752 with the empirical chi square 154.63 with prob < 1.4e-31

## The total n.obs was 1752 with Likelihood Chi Square = 95.03 with prob < 5.9e-19

##

## Tucker Lewis Index of factoring reliability = 0.788

## RMSEA index = 0.101 and the 90 % confidence intervals are 0.084 0.12

## BIC = 57.68

## Fit based upon off diagonal values = 0.9

## Measures of factor score adequacy

## MR1

## Correlation of (regression) scores with factors 0.92

## Multiple R square of scores with factors 0.84

## Minimum correlation of possible factor scores 0.69

#This looks only at Chinese speakes’ RTs

#Looks at Chinese Native speakers' responses only

SpeakerC <- RI %>%

filter(NativeLang == "2")

Chinese\_RT <- SpeakerC %>%

select(ends\_with("RT"))

fa(Chinese\_RT)

## Factor Analysis using method = minres

## Call: fa(r = Chinese\_RT)

## Standardized loadings (pattern matrix) based upon correlation matrix

## MR1 h2 u2 com

## AcceptRating.RT -0.02 0.00024 0.9998 1

## KnowIdiom.RT -0.07 0.00524 0.9948 1

## HowOftenUse.RT 1.00 0.99851 0.0015 1

## HowOftenOthersUse.RT -0.33 0.10718 0.8928 1

## LikeUsingIdioms.RT 0.09 0.00791 0.9921 1

##

## MR1

## SS loadings 1.12

## Proportion Var 0.22

##

## Mean item complexity = 1

## Test of the hypothesis that 1 factor is sufficient.

##

## df null model = 10 with the objective function = 0.16 with Chi Square = 284.02

## df of the model are 5 and the objective function was 0.03

##

## The root mean square of the residuals (RMSR) is 0.05

## The df corrected root mean square of the residuals is 0.07

##

## The harmonic n.obs is 1752 with the empirical chi square 83.87 with prob < 1.3e-16

## The total n.obs was 1752 with Likelihood Chi Square = 49.62 with prob < 1.7e-09

##

## Tucker Lewis Index of factoring reliability = 0.674

## RMSEA index = 0.071 and the 90 % confidence intervals are 0.054 0.09

## BIC = 12.28

## Fit based upon off diagonal values = 0.84

## Measures of factor score adequacy

## MR1

## Correlation of (regression) scores with factors 1

## Multiple R square of scores with factors 1

## Minimum correlation of possible factor scores 1

summary(Chinese\_RT)

## AcceptRating.RT KnowIdiom.RT HowOftenUse.RT HowOftenOthersUse.RT

## Min. : 696 Min. : 404 Min. :3797 Min. :2946

## 1st Qu.: 5880 1st Qu.: 1569 1st Qu.:5220 1st Qu.:3730

## Median : 7839 Median : 2197 Median :6015 Median :4851

## Mean : 9045 Mean : 2689 Mean :6196 Mean :5087

## 3rd Qu.:10742 3rd Qu.: 3106 3rd Qu.:7105 3rd Qu.:6909

## Max. :73971 Max. :37577 Max. :9789 Max. :8889

## LikeUsingIdioms.RT

## Min. :1726

## 1st Qu.:2366

## Median :3636

## Mean :3813

## 3rd Qu.:4521

## Max. :8349

#Looks at English Native speakers’ RT only

SpeakerE <- RI %>%

filter(NativeLang == "1")

English\_RT <- SpeakerE %>%

select(ends\_with("RT"))

fa(English\_RT)

## Factor Analysis using method = minres

## Call: fa(r = English\_RT)

## Standardized loadings (pattern matrix) based upon correlation matrix

## MR1 h2 u2 com

## AcceptRating.RT 0.22 0.049 0.95 1

## KnowIdiom.RT 0.19 0.037 0.96 1

## HowOftenUse.RT 0.47 0.225 0.78 1

## HowOftenOthersUse.RT 0.85 0.728 0.27 1

## LikeUsingIdioms.RT 0.80 0.632 0.37 1

##

## MR1

## SS loadings 1.67

## Proportion Var 0.33

##

## Mean item complexity = 1

## Test of the hypothesis that 1 factor is sufficient.

##

## df null model = 10 with the objective function = 0.92 with Chi Square = 4529.4

## df of the model are 5 and the objective function was 0.02

##

## The root mean square of the residuals (RMSR) is 0.03

## The df corrected root mean square of the residuals is 0.05

##

## The harmonic n.obs is 4929 with the empirical chi square 110.75 with prob < 2.8e-22

## The total n.obs was 4929 with Likelihood Chi Square = 79.32 with prob < 1.2e-15

##

## Tucker Lewis Index of factoring reliability = 0.967

## RMSEA index = 0.055 and the 90 % confidence intervals are 0.045 0.066

## BIC = 36.8

## Fit based upon off diagonal values = 0.99

## Measures of factor score adequacy

## MR1

## Correlation of (regression) scores with factors 0.91

## Multiple R square of scores with factors 0.83

## Minimum correlation of possible factor scores 0.65

summary(English\_RT)

## AcceptRating.RT KnowIdiom.RT HowOftenUse.RT HowOftenOthersUse.RT

## Min. : 81 Min. : 588 Min. :1823 Min. :1880

## 1st Qu.: 4173 1st Qu.: 1182 1st Qu.:3223 1st Qu.:2864

## Median : 5489 Median : 1477 Median :4338 Median :3384

## Mean : 6166 Mean : 1726 Mean :4594 Mean :3796

## 3rd Qu.: 7259 3rd Qu.: 1908 3rd Qu.:5508 3rd Qu.:4040

## Max. :65666 Max. :13366 Max. :9680 Max. :8560

## LikeUsingIdioms.RT

## Min. :1591

## 1st Qu.:2001

## Median :2417

## Mean :2803

## 3rd Qu.:3355

## Max. :6162

More variation is seen between Chinese and English speakers in RT than RESP but there are loading differences in both.

## Analysis Conclusion

Based on this analysis, it appears as though acceptability, how often one uses an idiom, how often one perceives others use an idiom, and how well liked a given idiom is may be indicative of the effects referred to as idiomatic frequency. This serves as an interesting starting point to test additional likely formative responses that may aid in the creation of a reliable quantification of factors that should be included in “idiom frequency”. Further investigation into variance patterning by condition could add more weight to this idea. This data contradicts some assumptions made by Geeraert et al. 2016 regarding frequency and confirms personal suspicions regarding underlying experiment assumptions.

# 3. Data Set Evaluation

The final element of this project is a brief evaluation of the data set with respect to one of the FAIR principles. FAIR referees to findable, accessible, interpretable, and reusable. To satisfy the first and second principles, data should be easy to find, openly available, and easy to access access from a data base that is clearly explained. The data used in this analysis was not freely available in an open-source repository. However, it is exceedingly rare for supplementary materials of any kind to be published in my field. Instead, the paper stated that any research-related inquiries about the paper or data should be directed to the corresponding author. I emailed her to explain my project and to ask if she would be wiling to share any of her data. To my surprise, she replied to my request within a few days and was more than happy to share her raw data along with the R code she used in her analyses.

To satisfy the third and fourth principles, data should be systematically structured, should provide a clear explanation of concepts and representations so that data can be reused, and should be reusable for future research by including things such as documentation, standards for data quality, attribution guidelines, and citation guidelines. While Dr. Geeraert shared her R code with me, I was not able to get her code to run on my computer. In this sense, the code itself was not immediately reusable. However, it was useful to see how her analysis was set up as I will be doing a similar type of analysis in the very near future. More importantly, her conditions were clearly defined and her stimuli were reasonably controlled for. As such, stimuli used in her original experiment can be used in other studies. Idiom comprehension work suffers from a dearth of unreliable replication, often due to uncontrolled for properties of idioms and differently operationalized constructs between research groups. Recently, some have published norming studies, releasing stimuli that can be reused between experiments to increase comparability. While that was not the primary goal of Geeraert et al. 2017, they are clearly open to sharing data to remedy this methodological flaw. Because this data was shared with me in private, I am not able to upload the raw data analyzed in this project. However, I hope that enough documentation was provided so as to explain the variables I created as well as to walk through the steps of my analysis. Additionally, Dr. Geeraert asked that I share my findings with her since I was analyzing her data in a new way. While my analysis will likely be of little use to her, it promotes an open science approach by sharing data that may be of use at a later date.

# References

1. Bobrow, S., Bell, S. (1973). On catching on to idiomatic expressions. Memory and Cognition, 1, 343–346.
2. Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. Journal of Memory and Language, 27, 668-683.
3. Geeraert, K., Newman, J. and Baayen, R.H. (2017). Idiom Variation: Experimental Data and a Blueprint of a Computational Model. Topics in Cognitive Science, 9, 653-669. DOI: 10.1111/tops.12263.
4. Hockett, C. F. (1958). A Course in Modern Linguistics. New York: Macmillan.
5. Katz, J. (1973). Compositionality, idiomaticity, and lexical substitution. In Anderson, S., Kiparsky, P. (Eds.). A Fest-schrift for Morris Halle, 357-376. Holt, New York.
6. Mair, P. (2018). Modern Psychometrics with R (R. Gentleman, K. Hornik, & G. Parmigiani, Eds.). Springer International Publishing AG.
7. Marlies, E.C. (1995). You Don’t Die Immediately When You Kick an Empty Bucket: A Processing View on Semantic and Syntactic Characteristics of idioms. In M. Everaet & E.J. V.D. Linden & A. Schenk and R. Schreuder (Eds.), Idioms: Structural and Psychological Perspectives, 283-299.
8. Mel’cuk, I. (1995). Phrasemes in language and phraseology in linguistics. In Everaert et al. (Eds.), Idioms, (167-232). Hillsdale N.J.: Lawrence Erlbaum Associates.
9. Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. Language, 70, 491-534.
10. Penttila, E. (2010). Prototype-based taxonomy of idiomatic expressions. Applications of Cognitive Linguistics, 14, 145–162. 2010
11. Pulman, S.G. (1993). Recognition and Interpretation of Idioms. In Cristina Cacciari & Patrizia Tabossi (Eds.), Idioms: Processing, Structure, and Interpretation, 249-270. NJ: Lawrance Erlbaum Associates.
12. Swinney, D., Cutler, A. (1979). The access and processing of idiomatic expressions. Journal of Verbal Learning and Verbal Behavior, 18, 523–534.
13. Weinreich, U. (1969). Problems in the analysis of idioms. Substance and structure of language, 23-82. University of California Press.