LIWCEgoR

Kevin Lanning

Friday, February 06, 2015

# Introduction

**Ego level** is conceptualized both as a single construct or dimension of individual differences in personality and as a series of qualitatively distinct milestones whose potential achievement occurs in a strict sequence over the course of maturation. This implies both that predictors of Ego Level may be non-linear as well as linear (Einstein & Lanning).

One other implication is that it may be worthwhile to examine the LIWC correlates of Ego Level at particular cutting points - that is, the LIWC categories which distinguish speech at adjacent ego levels. For example, the Conformist stage can be considered as an equilibrium, or as a local minimum from which "escape" (i.e., movement to the next stage) is difficult, running counter to the powerful social incentives of group loyalty and simple reciprocity. Loevinger (19xx) speculated that one avenue away from the gravitational pull of simple conformism might be the (insert quote here). This suggests that individuals at the Self-aware stage, in comparison with those at the just-prior Conformist stage, would be characterized by greater \_\_\_\_\_\_\_\_\_\_\_\_ and lesser \_\_\_\_\_\_\_\_\_.

**LIWC** (text from Daniel's thesis).

## Overview of studies

Treat exemplars and participant data as two separate studies.

# Study 1

# Introduction

Describe manual, cite paper in 1990s which describes SCT as the best validated projective measure.

# Methods

**Data cleaning** scanning and cleaning of SCT manual, discuss question mark problem,...

**LIWC categories** are not mutually exclusive. Dictionaries for superordinate categories (e.g., negative emotion) do not correspond precisely with the sum of subordinate dictionaries (dictionaries for anxiety, anger, and sadness). Consequently, no comprehensive relationship between LIWC language use categories and a criterion such as ego level can be examined without including overlapping categories. To help better understand relations between speech categories and Ego Level, we consider both (a) the broadest, logically independent, superordinate LIWC categories as well as (b) all LIWC categories.

library (psych)  
library(data.table)  
# partitioning of dataset into id, super, and subordinate LIWC categories  
source <- ("C:/Users/the/Dropbox/Shared Lopez Pauletti Lanning/Coded Exemplars/Exemplar LIWC Results.csv")  
LIWC.0 <- (read.csv(source))  
#attach(LIWC.0)  
idnames <- c("Item","Level","Segment")  
super <- c( "WC", "Sixltr", "Dic", "funct","verb","social","affect","cogmech","percept","bio","relativ","assent", "nonfl","filler","AllPct")  
LIWCIdnames.0 <- LIWC.0[idnames]  
LIWCSuperord.0 <- LIWC.0[super]  
LIWCSubord.0 <- subset(LIWC.0[,!(names(LIWC.0) %in% super)],  
 select = -(c(Item, Level, Segment, WPS)))

Scores on LIWC categories are typically percentages of words within a particular category (e.g., words of six or more letters) divided by the number of words in a given corpus (Word count). Because the 'corpi' under examination consist of responses to a set of standard prompts (e.g., "My mother..."), we will instead examine the properties of counts rather than percentages. These values are more directly interpretable, and allow examination of the contribution of fluency or word count as a separate predictor (ref to Loevinger argument for not adjusting).

# descriptive statistics for raw LIWc measures (percentages)  
head(describe(LIWCSuperord.0),3L)

## vars n mean sd median trimmed mad min max range skew  
## WC 1 7659 7.23 4.37 7.00 6.76 4.45 1 65 64 2.05  
## Sixltr 2 7659 19.44 19.40 16.67 16.61 24.70 0 100 100 1.49  
## Dic 3 7659 91.25 14.89 100.00 94.52 0.00 0 100 100 -2.59  
## kurtosis se  
## WC 10.68 0.05  
## Sixltr 3.31 0.22  
## Dic 9.30 0.17

head (describe(LIWCSubord.0),3L)

## vars n mean sd median trimmed mad min max range skew  
## pronoun 1 7659 16.42 14.90 16.67 14.86 19.23 0 100 100 0.81  
## ppron 2 7659 12.03 12.65 11.11 10.33 16.47 0 100 100 1.04  
## i 3 7659 5.03 10.25 0.00 2.54 0.00 0 100 100 2.42  
## kurtosis se  
## pronoun 1.03 0.17  
## ppron 1.65 0.14  
## i 7.66 0.12

LIWCSuperordc.0 <-  
 as.data.frame((LIWCSuperord.0) \* .01\* LIWCSuperord.0$WC)  
LIWCSubordc.0 <-  
 as.data.frame((LIWCSubord.0) \* .01\* LIWCSuperord.0$WC)  
LIWCSuperordc.0$WC <- LIWCSuperord.0$WC   
remove(LIWC.0)  
# Entries should be within rounding of integers   
head(LIWCSuperordc.0,3L)

## WC Sixltr Dic funct verb social affect cogmech percept bio relativ  
## 1 7 0 7 5.0001 1.9999 1.9999 1.0003 1.9999 0 0.0000 0  
## 2 4 0 4 2.0000 1.0000 1.0000 1.0000 0.0000 0 0.0000 0  
## 3 3 0 3 2.0001 0.9999 0.9999 0.0000 0.0000 0 0.9999 0  
## assent nonfl filler AllPct  
## 1 0 0 0 1.0003  
## 2 0 0 0 0.0000  
## 3 0 0 0 0.0000

head(LIWCSubordc.0,3L)

## pronoun ppron i we you shehe they ipron article auxverb past present  
## 1 1.9999 1.9999 0 0 0 1.9999 0.0000 0 0 1.0003 0 1.9999  
## 2 0.0000 0.0000 0 0 0 0.0000 0.0000 0 1 1.0000 0 1.0000  
## 3 0.9999 0.9999 0 0 0 0.0000 0.9999 0 0 0.9999 0 0.9999  
## future adverb preps conj negate quant number swear family friend  
## 1 0 0 1.0003 1.0003 1.0003 0 0 0 0 0  
## 2 0 0 0.0000 0.0000 0.0000 0 0 0 0 0  
## 3 0 0 0.0000 0.0000 0.0000 0 0 0 0 0  
## humans posemo negemo anx anger sad insight cause discrep tentat certain  
## 1 0 1.0003 0 0 0 0 0 0 1.0003 1.0003 0  
## 2 0 0.0000 1 0 0 0 0 0 0.0000 0.0000 0  
## 3 0 0.0000 0 0 0 0 0 0 0.0000 0.0000 0  
## inhib incl excl see hear feel body health sexual ingest motion space  
## 1 0 0 1.0003 0 0 0 0 0.0000 0 0 0 0  
## 2 0 0 0.0000 0 0 0 0 0.0000 0 0 0 0  
## 3 0 0 0.0000 0 0 0 0 0.9999 0 0 0 0  
## time work achieve leisure home money relig death Period Comma Colon  
## 1 0 0 0 1.0003 0 0 0 0 0 0 0  
## 2 0 0 0 1.0000 0 0 0 0 0 0 0  
## 3 0 0 0 0.0000 0 0 0 0 0 0 0  
## SemiC QMark Exclam Dash Quote Apostro Parenth OtherP  
## 1 0 0 0 0 0 1.0003 0 0  
## 2 0 0 0 0 0 0.0000 0 0  
## 3 0 0 0 0 0 0.0000 0 0

LIWCSuperordc.0 <- round(LIWCSuperordc.0,0)  
LIWCSubordc.0 <- round(LIWCSubordc.0,0)

# Results

The mean Ego Level in the exemplar data is 5.44, between the Self-Aware (5) and the Conscientious (6) levels and near to the midpoint of the range from Level 2 (Impulsive) to 9 (Integrated). \* We'll want to compare this with the values for the participant data.\*

The average length of the 7659 sentence completions is 7.23 (sd = 4.37, median = 7). These snippets of text are short, and consequently the counts for most LIWC categories for most utterances are 0. For 7 of the 14 of the superordinate categories, and for 59 of the 64 subordinate categories, the median count was 0, and the distribution of scores is necessarily skewed.

describe(LIWCIdnames.0)

## vars n mean sd median trimmed mad min max range  
## Item\* 1 7659 19.17 10.30 19 19.19 13.34 1 37 36  
## Level 2 7659 5.44 1.28 6 5.48 1.48 2 9 7  
## Segment 3 7659 3830.00 2211.11 3830 3830.00 2839.18 1 7659 7658  
## skew kurtosis se  
## Item\* -0.02 -1.13 0.12  
## Level -0.28 0.00 0.01  
## Segment 0.00 -1.20 25.27

describe(LIWCSuperordc.0)

## vars n mean sd median trimmed mad min max range skew  
## WC 1 7659 7.23 4.37 7 6.76 4.45 1 65 64 2.05  
## Sixltr 2 7659 1.26 1.23 1 1.09 1.48 0 12 12 1.72  
## Dic 3 7659 6.66 4.15 6 6.24 2.97 0 63 63 1.93  
## funct 4 7659 4.26 3.00 4 3.94 2.97 0 40 40 1.67  
## verb 5 7659 1.34 1.17 1 1.21 1.48 0 16 16 1.40  
## social 6 7659 1.14 1.25 1 0.95 1.48 0 15 15 1.64  
## affect 7 7659 0.67 0.82 1 0.54 1.48 0 8 8 1.53  
## cogmech 8 7659 1.34 1.44 1 1.12 1.48 0 21 21 1.94  
## percept 9 7659 0.15 0.41 0 0.05 0.00 0 5 5 2.99  
## bio 10 7659 0.15 0.40 0 0.04 0.00 0 4 4 2.76  
## relativ 11 7659 0.66 0.95 0 0.48 0.00 0 8 8 1.81  
## assent 12 7659 0.00 0.06 0 0.00 0.00 0 1 1 15.88  
## nonfl 13 7659 0.01 0.11 0 0.00 0.00 0 2 2 9.72  
## filler 14 7659 0.01 0.11 0 0.00 0.00 0 2 2 8.84  
## AllPct 15 7659 0.51 0.97 0 0.28 0.00 0 9 9 3.02  
## kurtosis se  
## WC 10.68 0.05  
## Sixltr 6.12 0.01  
## Dic 10.33 0.05  
## funct 7.31 0.03  
## verb 5.48 0.01  
## social 5.60 0.01  
## affect 4.23 0.01  
## cogmech 8.92 0.02  
## percept 11.47 0.00  
## bio 8.30 0.00  
## relativ 4.27 0.01  
## assent 250.24 0.00  
## nonfl 96.56 0.00  
## filler 78.99 0.00  
## AllPct 13.12 0.01

describe(LIWCSubordc.0)

## vars n mean sd median trimmed mad min max range skew  
## pronoun 1 7659 1.22 1.22 1 1.05 1.48 0 13 13 1.56  
## ppron 2 7659 0.86 0.93 1 0.74 1.48 0 11 11 1.62  
## i 3 7659 0.35 0.71 0 0.19 0.00 0 10 10 2.75  
## we 4 7659 0.03 0.18 0 0.00 0.00 0 3 3 7.88  
## you 5 7659 0.03 0.18 0 0.00 0.00 0 3 3 8.10  
## shehe 6 7659 0.32 0.65 0 0.18 0.00 0 9 9 2.59  
## they 7 7659 0.14 0.42 0 0.01 0.00 0 4 4 3.45  
## ipron 8 7659 0.36 0.63 0 0.23 0.00 0 4 4 1.92  
## article 9 7659 0.38 0.61 0 0.28 0.00 0 4 4 1.62  
## auxverb 10 7659 0.89 0.92 1 0.78 1.48 0 11 11 1.35  
## past 11 7659 0.17 0.45 0 0.05 0.00 0 4 4 3.05  
## present 12 7659 0.92 0.97 1 0.79 1.48 0 9 9 1.34  
## future 13 7659 0.07 0.27 0 0.00 0.00 0 3 3 3.80  
## adverb 14 7659 0.28 0.56 0 0.17 0.00 0 5 5 2.21  
## preps 15 7659 0.90 0.99 1 0.76 1.48 0 11 11 1.54  
## conj 16 7659 0.42 0.71 0 0.28 0.00 0 7 7 2.07  
## negate 17 7659 0.18 0.40 0 0.09 0.00 0 3 3 2.13  
## quant 18 7659 0.22 0.48 0 0.12 0.00 0 6 6 2.43  
## number 19 7659 0.03 0.18 0 0.00 0.00 0 3 3 6.60  
## swear 20 7659 0.00 0.06 0 0.00 0.00 0 2 2 18.82  
## family 21 7659 0.09 0.32 0 0.00 0.00 0 4 4 3.92  
## friend 22 7659 0.02 0.14 0 0.00 0.00 0 2 2 7.98  
## humans 23 7659 0.22 0.45 0 0.13 0.00 0 4 4 1.99  
## posemo 24 7659 0.45 0.70 0 0.32 0.00 0 6 6 1.85  
## negemo 25 7659 0.21 0.48 0 0.10 0.00 0 5 5 2.62  
## anx 26 7659 0.05 0.23 0 0.00 0.00 0 3 3 4.52  
## anger 27 7659 0.06 0.25 0 0.00 0.00 0 3 3 4.91  
## sad 28 7659 0.04 0.20 0 0.00 0.00 0 2 2 5.21  
## insight 29 7659 0.24 0.51 0 0.13 0.00 0 6 6 2.41  
## cause 30 7659 0.12 0.37 0 0.01 0.00 0 5 5 3.47  
## discrep 31 7659 0.18 0.44 0 0.07 0.00 0 5 5 2.85  
## tentat 32 7659 0.24 0.53 0 0.12 0.00 0 5 5 2.58  
## certain 33 7659 0.11 0.33 0 0.00 0.00 0 3 3 3.06  
## inhib 34 7659 0.06 0.25 0 0.00 0.00 0 3 3 4.37  
## incl 35 7659 0.28 0.57 0 0.17 0.00 0 5 5 2.30  
## excl 36 7659 0.25 0.56 0 0.12 0.00 0 5 5 2.69  
## see 37 7659 0.03 0.19 0 0.00 0.00 0 3 3 6.85  
## hear 38 7659 0.04 0.20 0 0.00 0.00 0 4 4 6.74  
## feel 39 7659 0.08 0.28 0 0.00 0.00 0 2 2 3.66  
## body 40 7659 0.02 0.16 0 0.00 0.00 0 2 2 7.41  
## health 41 7659 0.07 0.27 0 0.00 0.00 0 3 3 3.95  
## sexual 42 7659 0.05 0.23 0 0.00 0.00 0 3 3 4.71  
## ingest 43 7659 0.01 0.11 0 0.00 0.00 0 2 2 10.87  
## motion 44 7659 0.09 0.30 0 0.00 0.00 0 3 3 3.39  
## space 45 7659 0.30 0.58 0 0.18 0.00 0 6 6 2.25  
## time 46 7659 0.27 0.56 0 0.15 0.00 0 5 5 2.41  
## work 47 7659 0.15 0.41 0 0.04 0.00 0 4 4 3.14  
## achieve 48 7659 0.23 0.50 0 0.13 0.00 0 4 4 2.33  
## leisure 49 7659 0.06 0.26 0 0.00 0.00 0 3 3 4.59  
## home 50 7659 0.05 0.22 0 0.00 0.00 0 3 3 5.07  
## money 51 7659 0.04 0.21 0 0.00 0.00 0 2 2 5.84  
## relig 52 7659 0.02 0.14 0 0.00 0.00 0 2 2 7.62  
## death 53 7659 0.01 0.09 0 0.00 0.00 0 2 2 14.04  
## Period 54 7659 0.03 0.21 0 0.00 0.00 0 6 6 11.17  
## Comma 55 7659 0.17 0.54 0 0.03 0.00 0 7 7 4.42  
## Colon 56 7659 0.00 0.04 0 0.00 0.00 0 1 1 24.21  
## SemiC 57 7659 0.01 0.09 0 0.00 0.00 0 2 2 12.41  
## QMark 58 7659 0.00 0.09 0 0.00 0.00 0 5 5 31.19  
## Exclam 59 7659 0.01 0.11 0 0.00 0.00 0 3 3 12.59  
## Dash 60 7659 0.03 0.17 0 0.00 0.00 0 2 2 6.99  
## Quote 61 7659 0.03 0.18 0 0.00 0.00 0 3 3 8.04  
## Apostro 62 7659 0.11 0.35 0 0.01 0.00 0 4 4 3.29  
## Parenth 63 7659 0.03 0.18 0 0.00 0.00 0 3 3 6.07  
## OtherP 64 7659 0.05 0.24 0 0.00 0.00 0 3 3 5.32  
## kurtosis se  
## pronoun 5.07 0.01  
## ppron 6.52 0.01  
## i 12.27 0.01  
## we 70.55 0.00  
## you 76.17 0.00  
## shehe 10.68 0.01  
## they 13.48 0.00  
## ipron 3.99 0.01  
## article 2.88 0.01  
## auxverb 4.20 0.01  
## past 10.83 0.01  
## present 3.01 0.01  
## future 14.77 0.00  
## adverb 5.69 0.01  
## preps 5.17 0.01  
## conj 5.99 0.01  
## negate 4.00 0.00  
## quant 8.10 0.01  
## number 53.73 0.00  
## swear 385.89 0.00  
## family 19.21 0.00  
## friend 67.94 0.00  
## humans 4.04 0.01  
## posemo 4.98 0.01  
## negemo 9.11 0.01  
## anx 21.68 0.00  
## anger 27.49 0.00  
## sad 27.56 0.00  
## insight 7.62 0.01  
## cause 15.91 0.00  
## discrep 10.68 0.01  
## tentat 8.14 0.01  
## certain 9.33 0.00  
## inhib 21.10 0.00  
## incl 6.77 0.01  
## excl 8.98 0.01  
## see 57.24 0.00  
## hear 59.64 0.00  
## feel 13.32 0.00  
## body 60.20 0.00  
## health 16.30 0.00  
## sexual 24.00 0.00  
## ingest 130.85 0.00  
## motion 11.66 0.00  
## space 6.53 0.01  
## time 7.24 0.01  
## work 11.68 0.00  
## achieve 6.26 0.01  
## leisure 24.48 0.00  
## home 28.24 0.00  
## money 36.96 0.00  
## relig 60.94 0.00  
## death 219.62 0.00  
## Period 182.71 0.00  
## Comma 27.50 0.01  
## Colon 584.00 0.00  
## SemiC 161.17 0.00  
## QMark 1355.35 0.00  
## Exclam 194.20 0.00  
## Dash 53.30 0.00  
## Quote 79.92 0.00  
## Apostro 13.08 0.00  
## Parenth 41.43 0.00  
## OtherP 34.23 0.00

To partially address this skew, we transformed LIWC counts by taking their square roots. This transformation is, we believe, conceptually as well as psychometrically sound: One can make the argument that the psychophysics of at least some LIWC categories is not a linear function, but shows some marginal decline. That is, within a given sentence completion, the difference between 0 and 1 counts of LIWC categories such as question marks, swear words, or expressions of negative affect is arguably more meaningful than the difference between counts of 1 and 2. Transformed counts may better capture the meaning of LIWC category scores. (Because of the preponderance of zero values in these counts, a square-root rather than log transformation is appropriate here; Tukey (198x))

LIWCSuperordRootCount.0 <- sqrt(LIWCSuperordc.0)  
LIWCSubordRootCount.0 <- sqrt(LIWCSubordc.0)

## Assessing the shape of the function predicting Ego Level from individual LIWC categories

In order to assess possible non-linearity in relationships between LIWC Categories and Ego Level, a k-fold analysis is used.

In order to assess the shape (Linear, Quadratic, or Cubic) of the function predicting SCT scores from LIWC categories, I undertook a k-fold cross validation, as below.

# install.packages("cvTools")  
library (cvTools)

## Loading required package: lattice  
## Loading required package: robustbase  
##   
## Attaching package: 'robustbase'  
##   
## The following object is masked from 'package:psych':  
##   
## cushny

# Initialize results frame with column labels. Values will be appended below  
results = c ("x", "r", "Cases", "Best model",  
 "Linear RTMSPE", "Intercept", "b.linear", "linearR2",   
 "Quadrt RTMSPE", "Intercept", "b.linear", "b.quad", "quadR2",  
 "Cubic RTMSPE", "Intercept", "b.linear", "b.quad", "b.cubic", "cube R2")  
  
# I ran the two sets of variables separately  
  
#xdata <- LIWCSuperordRootCount.0  
xdata <- LIWCSubordRootCount.0  
  
# Ego level  
ydata <- LIWCIdnames.0[,2]  
  
#Next line is Dummy value for knitR; real line follows  
for (i in 1:1)  
# for (i in 1:ncol(xdata))  
 {  
 xlabel <- colnames(xdata[i])  
 xvar <- xdata[,i]  
 yvar <- ydata  
 seed <- 999  
 workdata <- as.data.frame(cbind(xvar,yvar))  
 folds <- cvFolds(nrow(workdata), K = 10, R = 1)  
 names(workdata) <- c("x","y")  
   
 # Set up functions. Using LMrob does not work for some highly skewed variables, hence LM  
 c1 <- call("lm", formula = y ~ poly(x,1, raw=T))  
 c2 <- call("lm", formula = y ~ poly(x,2, raw=T))  
 c3 <- call("lm", formula = y ~ poly(x,3, raw=T))  
   
 # k-fold regression  
 r1 <- cvFit(c1, data = workdata, y = workdata$y, folds = folds, cost = rtmspe, costArgs = list(trim = 0.1))  
 r2 <- cvFit(c2, data = workdata, y = workdata$y, folds = folds, cost = rtmspe, costArgs = list(trim = 0.1))  
 r3 <- cvFit(c3, data = workdata, y = workdata$y, folds = folds, cost = rtmspe, costArgs = list(trim = 0.1))  
   
 # choice of best model  
 bestmod <- cvSelect(Lin = r1, Qua = r2, Cub = r3)  
   
 # regression coefficients (weights, R2)  
 l1<-summary(lm(formula = y ~ poly(x,1, raw=T),data = workdata))  
 l2<-summary(lm(formula = y ~ poly(x,2, raw=T),data = workdata))  
 l3<-summary(lm(formula = y ~ poly(x,3, raw=T),data = workdata))  
  
 # and the simple correlation  
 corxy <- cor (xvar,yvar)  
   
 resultsValues = c(xlabel,corxy[1],bestmod$n, bestmod$best[[1]],  
 bestmod$cv[1,2], l1$coefficients[[1]], l1$coefficients[[2]], l1$r.squared,  
 bestmod$cv[2,2], l2$coefficients[[1]], l2$coefficients[[2]], l2$coefficients[[3]],   
 l2$r.squared,  
 bestmod$cv[3,2], l3$coefficients[[1]], l3$coefficients[[2]], l3$coefficients[[3]],  
 l3$coefficients[[4]], l3$r.squared)  
 results <- cbind (results,resultsValues)  
 }  
#write.csv(results, file = "results1.csv")  
#write.csv(results, file = "results2.csv")

Results have been compiled and transposed into a separate Excel spreadsheet.

## Item effects

We should look at these, but may not be able to analyze these in the exemplar data (which are relatively sparse).

## Multiple regressions

K-fold, possibly LASSO to look at whole dimension (ego level). K-fold, possibly LASSO to look at distinctions between adjacent stages (ego level).

...

# Study 2

# Introduction

Description of data

Several differences between these data and the exemplar data used in Study 1 should be noted. First, the distribution of Ego Level in real data can be expected to be narrower than in the exemplars, as the manual is written to be roughly equidiscriminating at all Ego Levels.

Second, in the real data there are person as well as item effects. Here, a multi-level approach is used in which we first examine the effects of LIWC categories on responses to the (whole) 18-item measure, then consider differences between stems.

# Methods

Answer is to combine 18 items into a single corpus (Daniel).

Set up data for Level II analysis (does the r between ego level and LIWC vary as a function of stem); Rachel.

# Results

See Exemplars section above.

## Analyses of respondent data using other methods (including raw word counts)

The raw word count analyses should use LASSO regressions or something similar to control for the number of predictors.