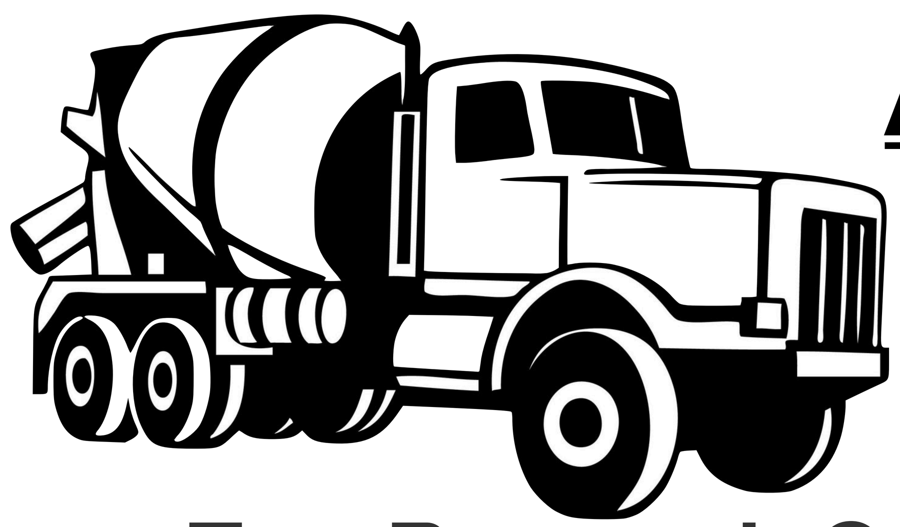


A Concrete Example of Construct Construction in Natural Language

Michael Yeomans, Imperial College London

https://zoom.us/j/97712135696

Two Research Questions

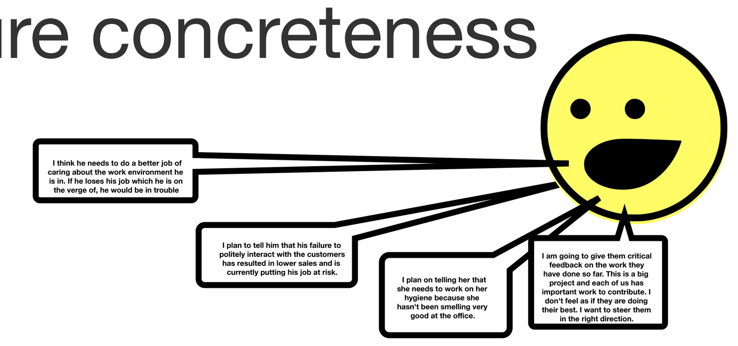
Abstract:

How do we turn words into numbers?

(Cronbach & Meehl, 1955; John & Benet- Martinez, 2000; Flake, Pek & Hehman, 2017; Fried & Flake, 2018)

Concrete:

How well can we measure concreteness in natural language?



-1.9 | -0.7 | 1.37 | 1.51

One Option: Humans

Plusses

What we've always been doing

More accurate than algorithms for complex tasks

Minuses

High marginal cost of labor

Not reliable

Not transparent

Another Option: Text Analysis Algorithms

Plus: There is an existing algorithm for concreteness

Minus: There are *eight* existing algorithms for concreteness

Big Minus: Language contains an *infinite number* of researcher degrees of freedom

Study 1: Advice Data

Borrowed from other research teams

One Ground Truth Measure:

Specificity - annotated by humans

Dataset Name	Context	Sample Size	Word Count	Source	Inter-Rater Agreement
Workplace Feedback	Annual 360 Reviews in a food processing firm	1334	20 (20)	Blunden, Green & Gino, 2018	0.82
Teacher Feedback	Parent-to-teacher letters for middle school students	304	36 (19)	Rogers & Kraft, 2015	0.89
Letter Advice	mTurkers giving advice for mistake-filled cover letter	951	32 (22)	Yoon, Blunden, Kristal & Whillans, 2020	0.92
Life Goals	mTurkers giving advice on how to live a good life	301	36 (25)	Zhang & North, 2020	0.63
Personal Feedback	mTurkers recalling giving recent personal feedback	171	36 (21)	Blunden, Green & Gino, 2018	0.86
Task Tips	Lab participants gave advice for games (e.g. darts, boggle)	228	38 (25)	Levari, Wilson & Gilbert, 2020	0.69

Study 2: Plan-Making Data

Collected in HarvardX pre-course surveys

Two Ground Truth Measures:

Specificity - annotated by humans

Distance - randomly assigned (week- vs. course-long plans)

Course Name	Sample Size	Word Count mean (sd)
American Government (HKS)	591	52.3 (36.5)
Contract Law (HLS)	322	50.3 (37.5)
Masterpieces of World Literature	470	46.4 (36.5)
Principles of Biochemistry	301	53.5 (34)
Data Science: R Basics	494	45.8 (32.8)
Using Python for Research	2003	38.5 (31.2)
Science & Cooking: From Haute Cuisine to Soft Matter Science	991	46.2 (38.1)

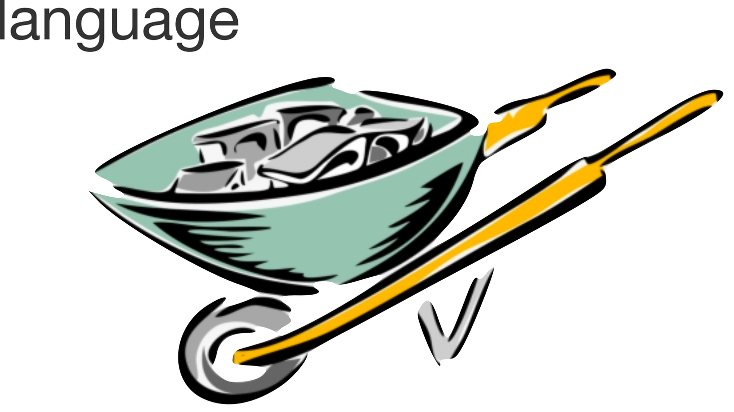
Summary of Results from Previous Models

Type of Measure	Name of Measure	Source	Measurement Validity: Advice	Measurement Validity: Plan Distance	Measurement Validity: Plan Specificity	Measurement Validity: Describing	Reproducibility
Word-Level Dictionary	mTurk Ratings	Brysbaert, Warriner & Kuperman, 2014	Low	Low	Low	Low	Medium
	Original MRC	Coltheart, 1981	Low	Low	Very Low	Medium	Medium
	Bootstrap MRC	Paetzold & Specia, 2016	Low	Low	Low	Low	Medium
Broad Categorical Scoring	Immediacy	Pennebaker & King, 1999	Zero	Very Low	Zero	Medium	Low
	Larrimore- LIWC	Larrimore et al., 2011	Very Low	Very Low	Very Low	Zero	Low
	Pan-LIWC	Pan et al., 2018	Zero	Very Low	Very Low	Zero	Low
	Original LCM	Seih, Beier & Pennebaker, 2017	Zero	Very Low	Zero	Medium	Low
	Syntax LCM	Johnson-Grey et al., 2019	Zero	Zero	Very Low	Low	High
	DICTION	Hart, 2001	Very Low	Zero	Zero	Very Low	Very Low
Machine Learning	**doc2concrete**	Yeomans, 2020	**Medium**	**Medium**	**Medium**	**Low**	**Very High**

Zero = < .03 Very Low = .03 - .1 Low = .1 - .2 Medium = .2 - .4 High = .4 - .6 Very High = >.6

Takeaways

- Off-the-shelf measures routinely fail
- Quality is correlated with transparency
- Quality is inversely correlated with price
- Expect domain-specificity as a rule
- Description text is simpler than natural language



Our Solution

"Mega-Analysis"

Compare many measures (m=12) across many contexts (k=17) with large samples (N=9,780)

The Results

- Most existing measures have no validity in our data
- A few existing measures have some validity
- New domain-specific measures perform better

- All analyses reproducible in *doc2concrete*

Get it Right: Build your own Model!

Simple Recipe for Machine Learning

Collect Ground Truth:

Train human annotators (ideally 2+, for reliability)

Collect annotations in-domain (no less than 500)

Extract features:

All 1,2,3-word sequences ("n-grams")

Extra features: Brysbaert & Paetzold scores

Estimate model:

Predict annotations using features

LASSO algorithm - regression-like

Evaluate Accuracy:

In-domain: nested cross-validation

Out-of-domain: transfer learning

In- vs. Out-of-Domain

Training Dataset	Test Dataset: Advice	Test Dataset: Plan Distance	Test Dataset: Plan Specificity	Test Dataset: Describing
Advice	**.228 [.195, .260]**	.004 [-.024, .031]	.258 [.232, .283]	-.113 [-.166, -.059]
Plan Distance	.022 [-.012, .056]	**.339 [.315, .363]**	.026 [-.001, .053]	-.012 [-.066, .042]
Plan Specificity	.191 [.158, .224]	.038 [.011, .065]	**.733 [.720, .745]**	-.032 [-.086, .022]
Describing	.119 [.085, .152]	.012 [-.015, .039]	.417 [.394, .439]	**.092 [.038, .145]**
Best Previous	.155 [.122, .188]	.047 [.020, .075]	.438 [.416, .460]	.363 [.315, .409]

How Many Annotations?