The Dark Side of Likert-type Scales: Implications of the Midscale Disagreement Problem

Dimitri Paisios, Nathalie Huet, Elodie Labeye

THE DARK SIDE OF LIKERT-TYPE SCALES

IMPLICATIONS OF THE MIDSCALE DISAGREEMENT PROBLEM

DIMITRI PAISIOS, NATHALIE HUET & ELODIE LABEYE
CLLE LAB (CNRS UMR 5263), UNIVERSITY OF TOULOUSE

STIMULI ARE OFTEN NORMED WITH LIKERT-TYPE SCALES

HOW FEARFUL IS THIS FACE?

NOT AT ALL — EXTREMELY

MA ET AL. (2015)

TO WHAT EXTENT DOES THIS IMAGE APPEAR TO BE VISUALY COMPLEX?

VERY SIMPLE — VERY COMPLEX

BRODEUR ET AL. (2010)

HOW EASY IS IT FOR THE HUMAN BODY TO PHYSICALLY INTERACT WITH THIS OBJECT?

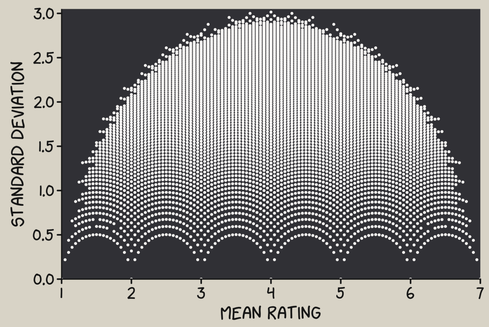
hammer

LOW — MEDIUM — HIGH

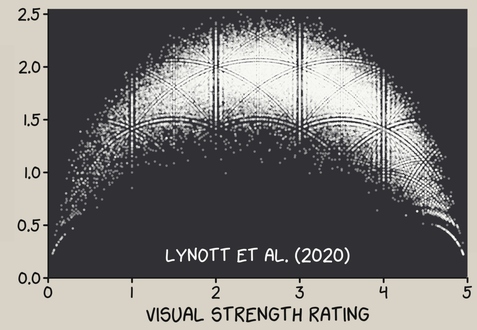
PEXMAN ET AL. (2018)

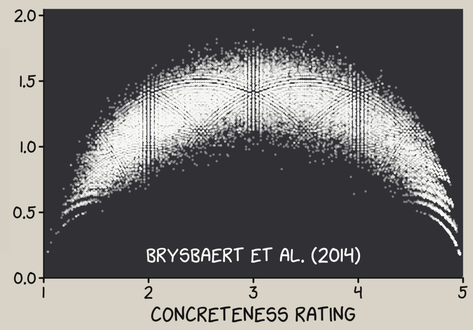
BUT THESE "NORMS" ARE JUST AVERAGES ... AND AVERAGES CAN GO VERY WRONG

THEORETICALLY POSSIBLE MEANS & STANDARD DEVIATIONS



POLLOCK (2018) OBSERVED THAT RATINGS GENERALLY DISPLAY A CONCAVE RELATIONSHIP WITH THEIR STANDARD DEVIATIONS





LET'S TAKE A CLOSER LOOK TO BODY-OBJECT INTERACTION (BOI) RATINGS

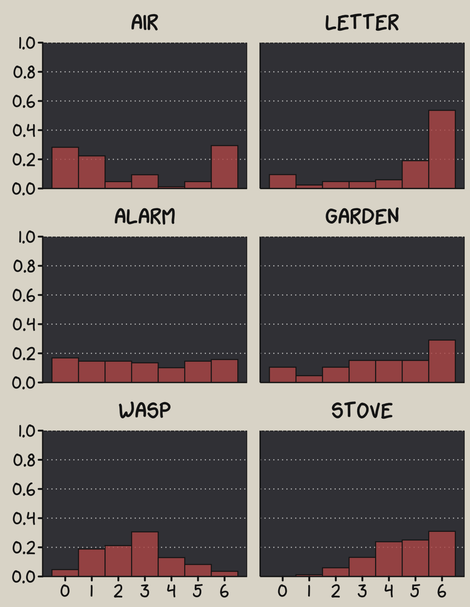
DATABASES TYPICALLY DON'T PROVIDE ITEM-LEVEL RATINGS

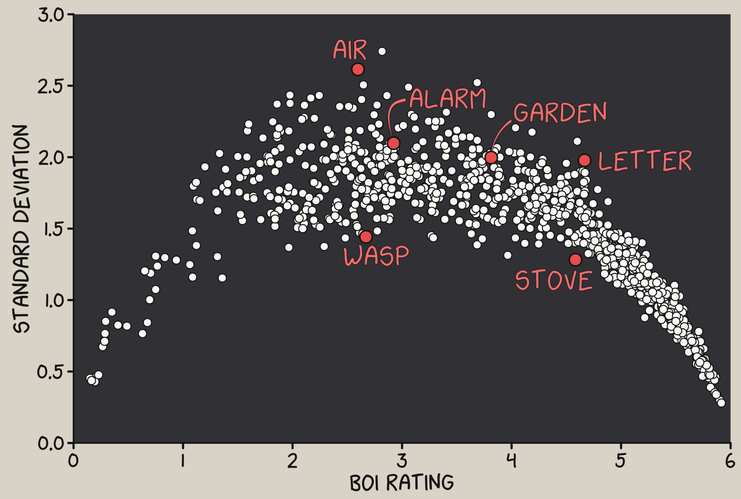
WE'VE COLLECTED OUR OWN BODY-OBJECT INTERACTION (BOI) RATINGS TO UNDERSTAND WHAT IS HAPPENING IN THE MIDDLE OF THE SCALE

RATE THE EASE WITH WHICH THE HUMAN BODY CAN PHYSICALLY AND DIRECTLY INTERACT WITH WHAT EACH WORD REPRESENTS

IMPOSSIBLE | VERY DIFFICULT | DIFFICULT | RATHER DIFFICULT | RATHER EASY | EASY | VERY EASY
0 | 1 | 2 | 3 | 4 | 5 | 6

1019 WORDS, M = 86 OBS/WORD [75, 96] — PAISIOS ET AL. (2023)

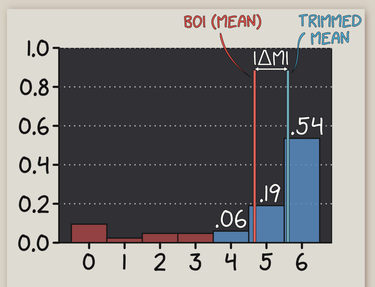




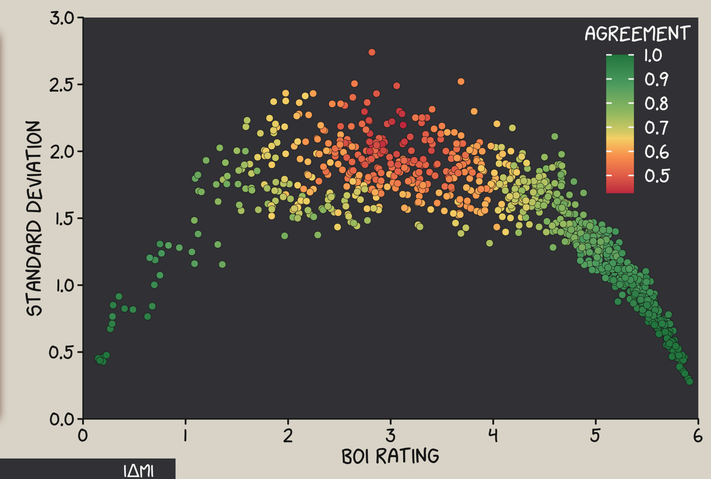
THE MIDDLE OF THE SCALE SEEMS TO CAPTURE A LOT OF DISAGREEMENT, WITH AVERAGE RATINGS NOT REPRESENTATIVE OF THE UNDERLYING RESPONSE DISTRIBUTIONS

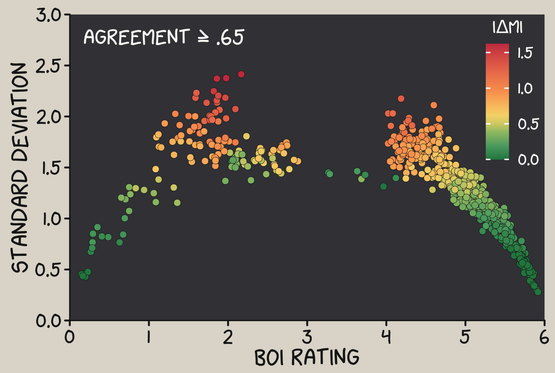
... AND ZOOM BACK FOR THE BIGGER PICTURE

AS MOST STUDIES REPORT MEANS AND STANDARD DEVIATIONS, WE WANTED A "MAP" OF THE PLOT FOR BETTER READABILITY AND TO INTERPRET SIMILAR DATA IN THE LITERATURE



WE DEFINE AGREEMENT AS THE HIGHEST PROPORTION OF RESPONSES ON ANY THREE CONSECUTIVE RATING OPTIONS



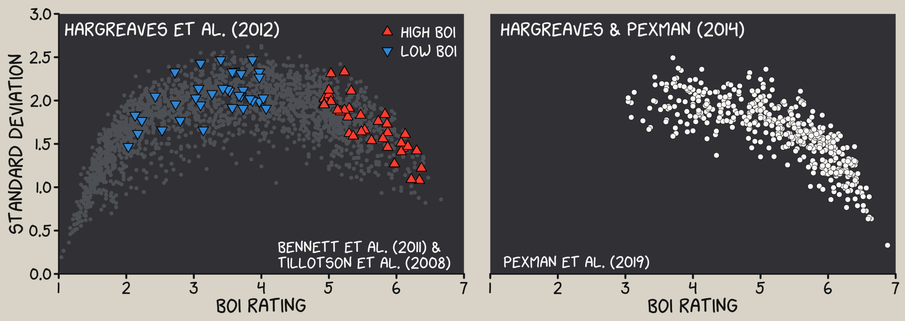


MOST RATINGS IN ABOUT THE MIDDLE THIRD OF THE SCALE ARE ARTEFACTS OF DISAGREEMENT, NOT CONSENSUAL JUDGEMENTS

FOR THE REST OF THE ITEMS, MOST RESPONSES ARE CLUSTERED AROUND THE EDGES & HIGHER STANDARD DEVIATIONS ROUGHLY CORRESPOND TO INCREASINGLY BIASED RATINGS

WHAT DOES THIS ALL MEAN ?!

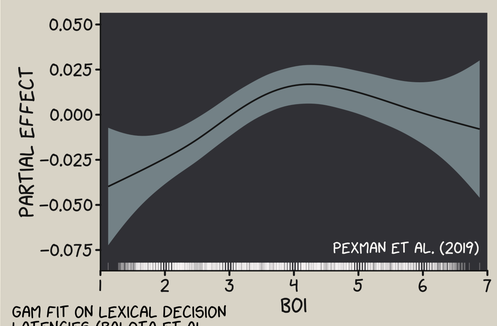
1. THE MIDDLE OF THE SCALE DOESN'T CAPTURE THE DIMENSION OF INTEREST



SUCH STUDIES ARE JUST COMPARING ITEMS THAT GENERATE HIGH DISAGREEMENT WITH ITEMS RATED HIGH

NO CONCLUSIONS ARE POSSIBLE ON THE VARIABLE'S EFFECT

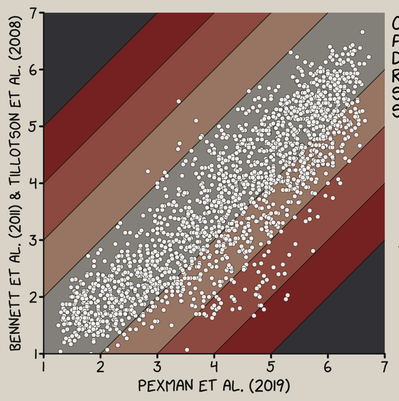
2. WHATEVER GENERATES DISAGREEMENT CAN ALSO AFFECT TASK PERFORMANCE



GAM FIT ON LEXICAL DECISION LATENCIES (BALOTA ET AL., 2007) WITH SEVERAL OTHER PREDICTORS

THERE IS A CLEAR MIDSCALE EFFECT & ASSUMING LINEARITY LEADS TO BIASED ESTIMATES

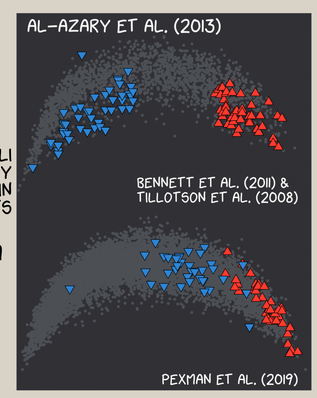
3. DISAGREEMENTS CAN LEAD TO MEASURMENT ERROR ... AND WE HAVE NO CLUE ABOUT THE REQUIRED SAMPLE SIZE



COMPARING THE RATINGS FROM TWO COMPLEMENTARY DATASETS (Y) WITH A MORE RECENT ONE (X) SHOWS SIGNIFICANT VARIATION FOR SOME ITEMS

IN EXPERIMENTS, STIMULI GENERALLY HAVE VERY DIFFERENT RATINGS IN DIFFERENT DATASETS

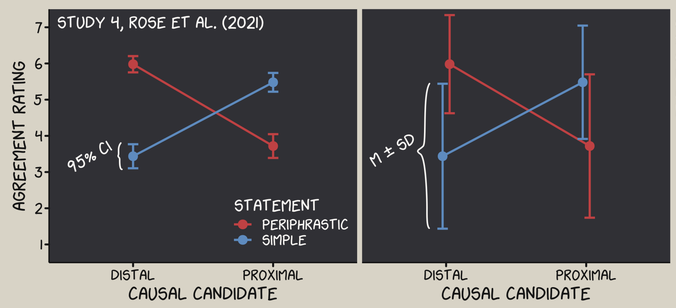
THE NUMBER OF OBS/ITEM IS HIGHLY VARIABLE ACROSS STUDIES & GETTING LOWER. NOBODY DISCUSSES HOW MANY WE NEED FOR RELIABLE RATINGS



LIKERT-TYPE RATINGS ARE ALSO USED AS DEPENDANT VARIABLES & ANALYSED WITH METRIC MODELS LIDDELL & KRUSCHKE (2018)

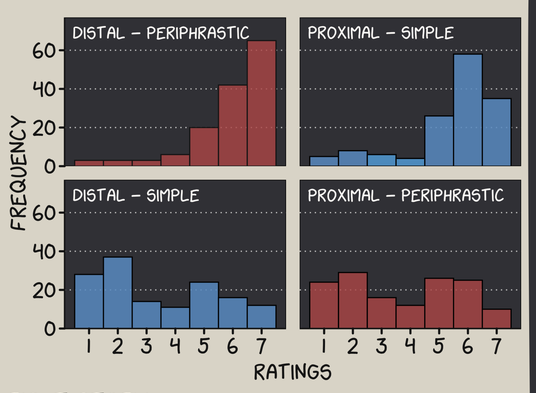
LET'S LOOK AT AN EXAMPLE FROM ROSE ET AL.'S (2021) STUDIES ON CAUSAL JUDGMENT

THE AUTHORS WERE INTERESTED IN HOW MUCH PARTICIPANTS AGREE THAT AN AGENT IS THE CAUSE OF AN EVENT DEPENDING ON HOW THE QUESTION IS FRAMED



CONFIDENCE INTERVALS (AND STANDARD ERRORS MORE GENERALLY) CAN CONCEAL DISAGREEMENTS AND GIVE A FALSE SENSE OF CONFIDENCE IN THE RESULTS

THE RAW DATA SHOWS THAT RATINGS WERE NOT OVERALL LOWER IN THE TWO CONDITIONS WITH MIDSCALE AVERAGES – PARTICIPANTS DISAGREED MORE ABOUT THEIR JUDGEMENTS



AVERAGES IN THE FACE OF DISAGREEMENT DO NOT TELL THE RIGHT STORY

SOME REMARKS & CONCLUSIONS

- LOOK AT YOUR DATA (!)
  WE OVERRELY ON STATISTICS WITHOUT TAKING THE TIME TO UNDERSTAND OUR DATA
- THERE ARE NO METHODOLOGICAL GUIDELINES FOR RATINGS
- WE DON'T REALLY KNOW HOW TO INTERPRET RELIABILITY METRICS (ICC, SPLIT-HALF CORRELATION, CRONBACH'S ALPHA, ...) AND HOW RELIABLE THEY ARE
- SOME RECENT LARGE-SCALE DATASETS ARE, AT BEST, VERY NOISY DUE TO A LOW SAMPLE SIZE (~10 OBS/ITEM)
- THE RESULTS OF A LARGE NUMBER OF STUDIES (ESPECIALLY fMRI) ARE QUESTIONABLE DUE TO UNRELIABLE RATINGS AND INADEQUATE STIMULUS SAMPLING & STATISTICAL MODELING

REFERENCES

AL-AZARY, H., YU, T., & MCRAE, K. (2022). CAN YOU TOUCH THE N400? THE INTERACTIVE EFFECTS OF BODY–OBJECT INTERACTION AND TASK DEMANDS ON N400 AMPLITUDES AND DECISION LATENCIES. *BRAIN AND LANGUAGE*, *231*, 105147. HTTPS://DOI.ORG/10.1016/J.BANDL.2022.105147

BALOTA, D. A., YAP, M. J., HUTCHISON, K. A., CORTESE, M. J., KESSLER, B., LOFTIS, B., NEELY, J. H., NELSON, D. L., SIMPSON, G. B., & TREIMAN, R. (2007). THE ENGLISH LEXICON PROJECT. *BEHAVIOR RESEARCH METHODS*, *39*(3), 445–459. HTTPS://DOI.ORG/10.3758/BF03193014

BENNETT, S. D. R., BURNETT, A. N., SIAKALUK, P. D., & PEXMAN, P. M. (2011). IMAGEABILITY AND BODY–OBJECT INTERACTION RATINGS FOR 599 MULTISYLLABIC NOUNS. *BEHAVIOR RESEARCH METHODS*, *43*(4), 1100–1109. HTTPS://DOI.ORG/10.3758/S13428-011-0117-5

BRODEUR, M. B., DIONNE-DOSTIE, E., MONTREUIL, T., & LEPAGE, M. (2010). THE BANK OF STANDARDIZED STIMULI (BOSS), A NEW SET OF 480 NORMATIVE PHOTOS OF OBJECTS TO BE USED AS VISUAL STIMULI IN COGNITIVE RESEARCH. *PLOS ONE*, *5*(5), E10773. HTTPS://DOI.ORG/10.1371/JOURNAL.PONE.0010773

BRYSBAERT, M., WARRINER, A. B., & KUPERMAN, V. (2014). CONCRETENESS RATINGS FOR 40 THOUSAND GENERALLY KNOWN ENGLISH WORD LEMMAS. *BEHAVIOR RESEARCH METHODS*, *46*(3), 904–911. HTTPS://DOI.ORG/10.3758/S13428-013-0403-5

HARGREAVES, I. S., LEONARD, G. A., PEXMAN, P. M., PITTMAN, D. J., SIAKALUK, P. D., & GOODYEAR, B. G. (2012). THE NEURAL CORRELATES OF THE BODY–OBJECT INTERACTION EFFECT IN SEMANTIC PROCESSING. *FRONTIERS IN HUMAN NEUROSCIENCE*, *6*. HTTPS://DOI.ORG/10.3389/FNHUM.2012.00022

HARGREAVES, I. S., & PEXMAN, P. M. (2014). GET RICH QUICK: THE SIGNAL TO RESPOND PROCEDURE REVEALS THE TIME COURSE OF SEMANTIC RICHNESS EFFECTS DURING VISUAL WORD RECOGNITION. *COGNITION*, *131*(2), 216–242. HTTPS://DOI.ORG/10.1016/J.COGNITION.2014.01.001

LIDDELL, T. M., & KRUSCHKE, J. K. (2018). ANALYZING ORDINAL DATA WITH METRIC MODELS: WHAT COULD POSSIBLY GO WRONG? *JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY*, *79*, 328–348. HTTPS://DOI.ORG/10.1016/J.JESP.2018.08.009

LYNOTT, D., CONNELL, L., BRYSBAERT, M., BRAND, J., & CARNEY, J. (2020). THE LANCASTER SENSORIMOTOR NORMS: MULTIDIMENSIONAL MEASURES OF PERCEPTUAL AND ACTION STRENGTH FOR 40,000 ENGLISH WORDS. *BEHAVIOR RESEARCH METHODS*, *52*(3), 1271–1291. HTTPS://DOI.ORG/10.3758/S13428-019-01316-Z

MA, D. S., CORRELL, J., & WITTENBRINK, B. (2015). THE CHICAGO FACE DATABASE: A FREE STIMULUS SET OF FACES AND NORMING DATA. *BEHAVIOR RESEARCH METHODS*, *47*(4), 1122–1135. HTTPS://DOI.ORG/10.3758/S13428-014-0532-5

PAISIOS, D., HUET, N., & LABEYE, E. (2023). ADDRESSING THE ELEPHANT IN THE MIDDLE: IMPLICATIONS OF THE MIDSCALE DISAGREEMENT PROBLEM THROUGH THE LENS OF BODY–OBJECT INTERACTION RATINGS. *COLLABRA: PSYCHOLOGY*, *9*(1), 84564. HTTPS://DOI.ORG/10.1525/COLLABRA.84564

PEXMAN, P. M., MURAKI, E., SIDHU, D. M., SIAKALUK, P. D., & YAP, M. J. (2019). QUANTIFYING SENSORIMOTOR EXPERIENCE: BODY–OBJECT INTERACTION RATINGS FOR MORE THAN 9,000 ENGLISH WORDS. *BEHAVIOR RESEARCH METHODS*, *51*(2), 453–466. HTTPS://DOI.ORG/10.3758/S13428-018-1171-Z

POLLOCK, L. (2018). STATISTICAL AND METHODOLOGICAL PROBLEMS WITH CONCRETENESS AND OTHER SEMANTIC VARIABLES: A LIST MEMORY EXPERIMENT CASE STUDY. *BEHAVIOR RESEARCH METHODS*, *50*(3), 1198–1216. HTTPS://DOI.ORG/10.3758/S13428-017-0938-Y

ROSE, D., SIEVERS, E., & NICHOLS, S. (2021). CAUSE AND BURN. *COGNITION*, *207*, 104517. HTTPS://DOI.ORG/10.1016/J.COGNITION.2020.104517

TILLOTSON, S. M., SIAKALUK, P. D., & PEXMAN, P. M. (2008). BODY–OBJECT INTERACTION RATINGS FOR 1,618 MONOSYLLABIC NOUNS. *BEHAVIOR RESEARCH METHODS*, *40*(4), 1075–1078. HTTPS://DOI.ORG/10.3758/BRM.40.4.1075

cnrs

CLLE COGNITION LANGUES LANGAGE ERGONOMIE (UMR 5263)

Université de Toulouse