Metric Calibration of Psychological Instruments in Social Psychology

Etienne LeBel & Bertram Gawronski

The University of Western Ontario, London, Canada

INTRODUCTION

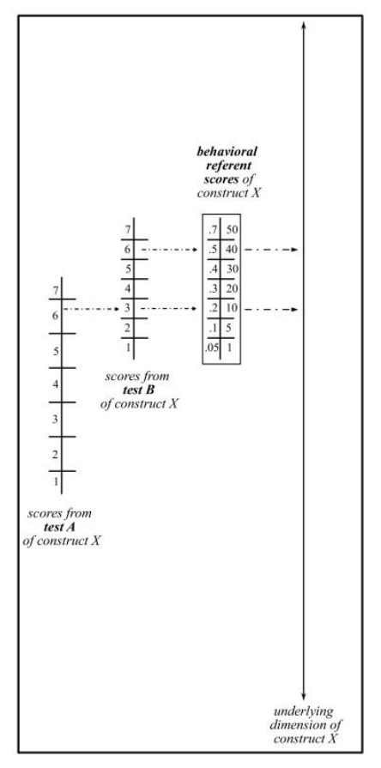
Goal: Argue that it is both ***feasible*** and ***useful*** to reduce the metric arbitrariness of psychological instruments used in basic research.

Definitions

***Metric**:* unit of measurement quantifying the amount of something.

Arbitrary metric: when it is empirically unknown where a given score locates an individual on the underlying psychological dimension (Blanton & Jaccard, 2006a, 2006b).

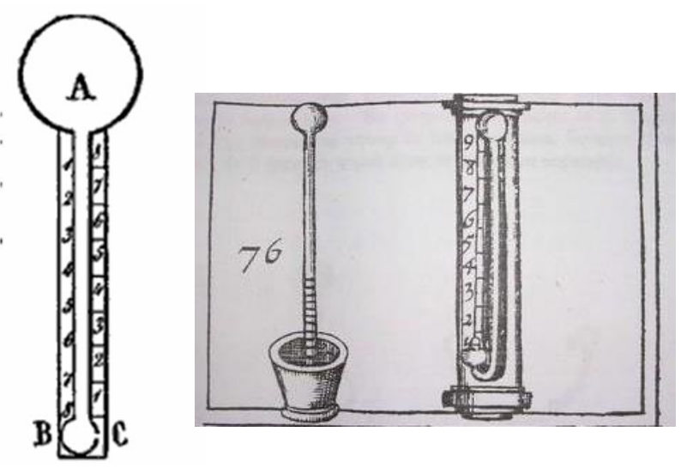
Virtually all instruments in psychology have an arbitrary metric.


behavioral referent scores of construct X
scores from test B of construct X
scores from test A of construct X
underlying dimension of construct X

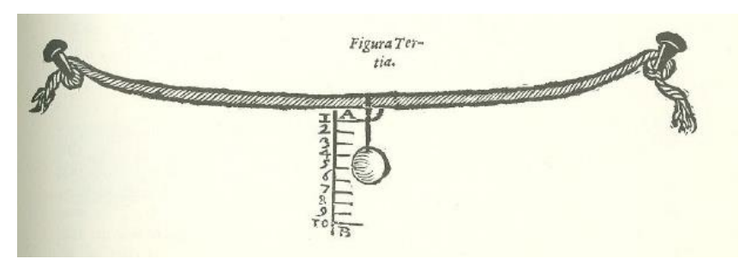
Background Inspirations

Development of Instruments in the Natural Sciences

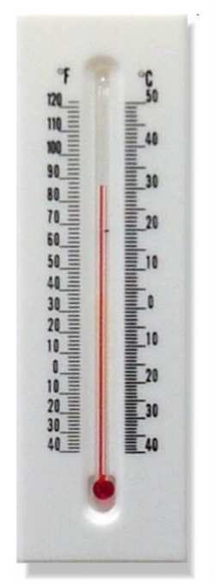
Early thermoscopes (i.e., thermometers) and hygrometers had scales with arbitrary metrics; however, eventually meaningful metrics were developed by calibrating instruments to relevant fixed points.



Early thermoscopes using scale with arbitrary metric (1611-1613).



Santorio's early string hygrometer using a scale with arbitrary metric (circa 1612).

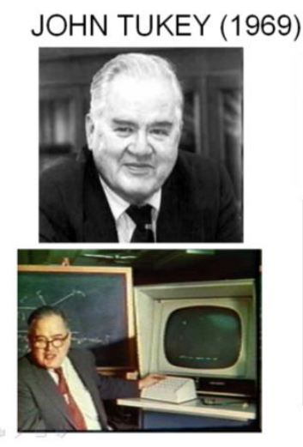


Daniel Fahrenheit proposed Fahrenheit scale (1724) and Anders Celsius proposed Celsius scale (1742), both calibrating to the same freezing and boiling points of water as fixed points.

Past psychology giants

Several prominent psychologists have uttered statements broadly consistent with the idea that reducing the metric arbitrariness of our instruments would benefit our science.

JOHN TUKEY (1969)



ANALYZING DATA:
SANCTIFICATION OR DETECTIVE WORK?
JOHN W. TUKEY
Princeton University and Bell Telephone Laboratories

"The physical sciences have learned much by storing up amounts, not just directions. If, for example, elasticity had been confined to "When you pull on it, it gets longer!" Hooke's law, the elastic limit, plasticity, and many other important topics could *not* have appeared" (emphasis added) (p. 86).

"…being so disinterested in our variables that we do not care about their units can hardly be desirable" (Tukey, 1969, p. 89).

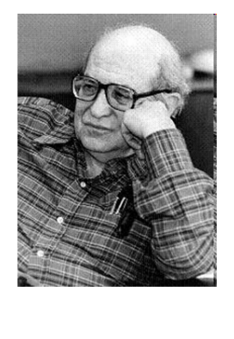
PAUL MEEHL (1978)

Journal of Consulting and Clinical Psychology 1978, Vol. 46, 806-834. #113
Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology
Paul E. Meehl
University of Minnesota

"the more dangerous tests [a theory] has survived, the better corroborated it is" (p. 817)

"…a theory that makes precise predictions and correctly picks out *narrow intervals* or *point values* out of the range of experimental possibilities is a much stronger theory" (p. 818, emphasis in original).

JACOB COHEN (1994)



The Earth Is Round ($p < .05$)
Jacob Cohen

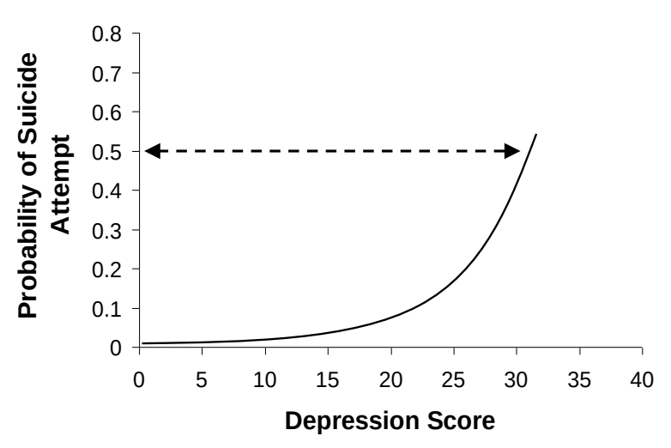
"But if all we learn from a research is that A is larger than B ($p < .01$), we have not learned very much. And this is typically all we learn" (p. 1001)

LEE SECHREST (1996)

"**Psychologists *cannot* claim to have high-quality measures if they do not have full knowledge of their [behavioral] implications**. We believe that knowledge, understanding, and progress in the science of psychology would be furthered greatly by concerted efforts to calibrate psychological measures…" (p. 1071).

General strategy to reduce metric arbitrariness

1. Develop consensus among researchers about which particular behaviors places an individual at the very high (or low) end of the theoretical continuum of the underlying construct
2. Map observed test scores to these agreed-upon theoretically-meaningful unambiguous behaviors, which serve as behavioral fixed points. Behaviors can either be:
   i. noteworthy differences in behavior (e.g., absence or presence of behavior) or
   ii. gradation of a behavior (e.g., behavioral counts)
3. Test scores gain meaning with respect to behavioral reference point
   (& then can translate scale into more intuitive metric, e.g., -10° to +10° degrees rather than 1 to 7)



Characteristics of ideal behavioral reference point:

- theoretically relevant
- interpretationally meaningful
- unambiguous (construct-wise)
- objective
- precisely measurable

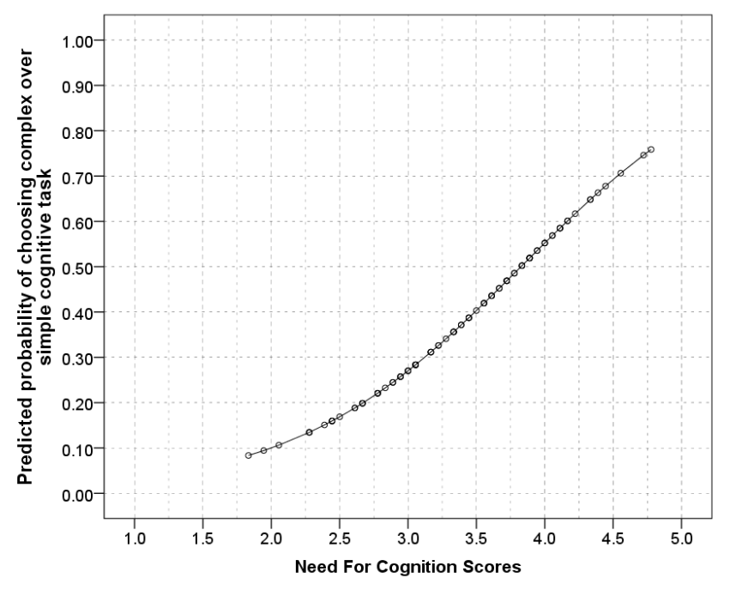
***And must consider interpretational context

EMPIRICAL DEMONSTRATIONS

Study 1

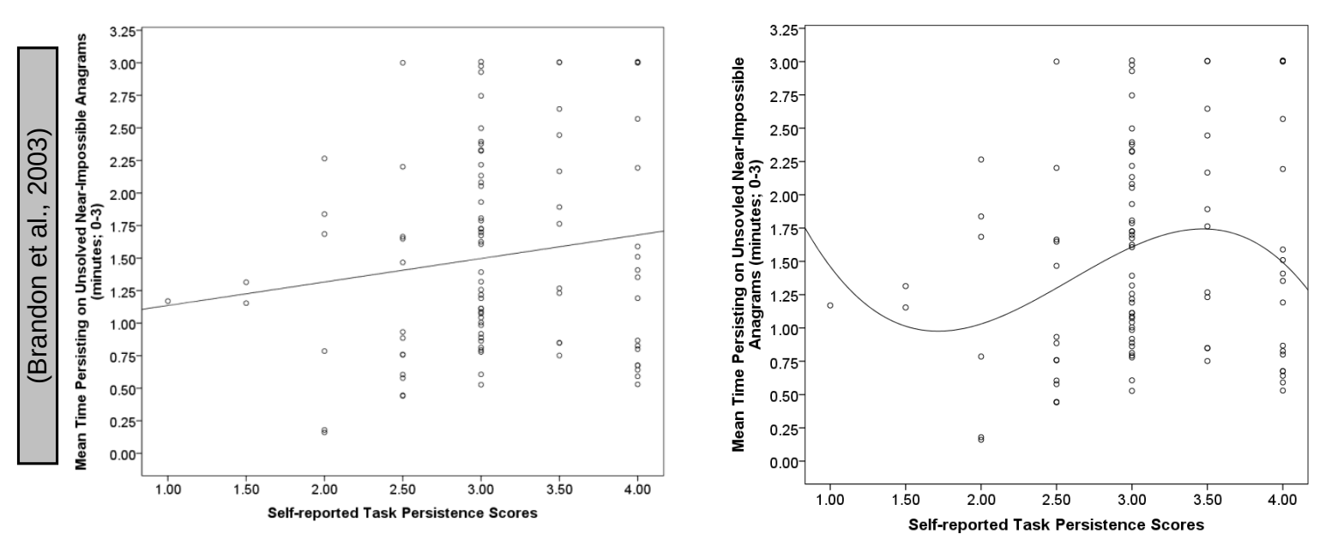
N = 94 (69 females, 25 males; mean age = 18.5, SD = 2.2, range = 17 to 30), UWO undergraduates participated for course credit

Need for Cognition



Example item: "*Thinking is not my idea of fun.*" (reverse-coded)
(5-point scale: *extremely uncharacteristic-extremely characteristic*)
(Cacioppo et al., 1984)

Task Persistence

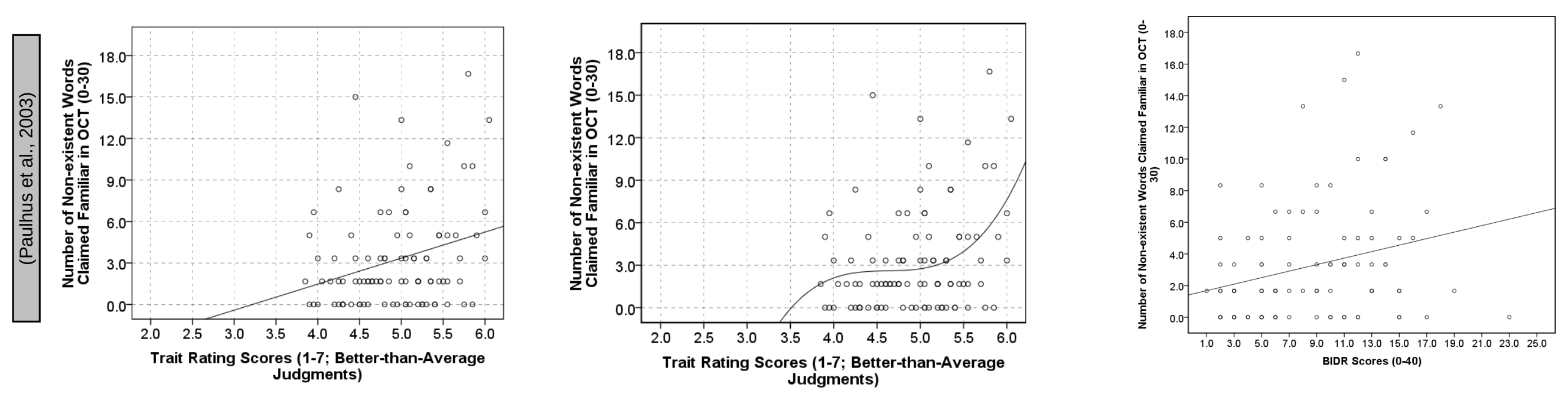


Example item: "*I will keep trying the same thing over again even when I have not had success the first time.*"
(4-point scale: *very untrue, not at all like me-very true, very much like me*) (Steinberg et al., 2007)

Study 2

N = 97 (50 females, 47 males; mean age = 18.9, SD = 1.3, range = 17 to 25), UWO undergraduates participated for course credit

Self-enhancement



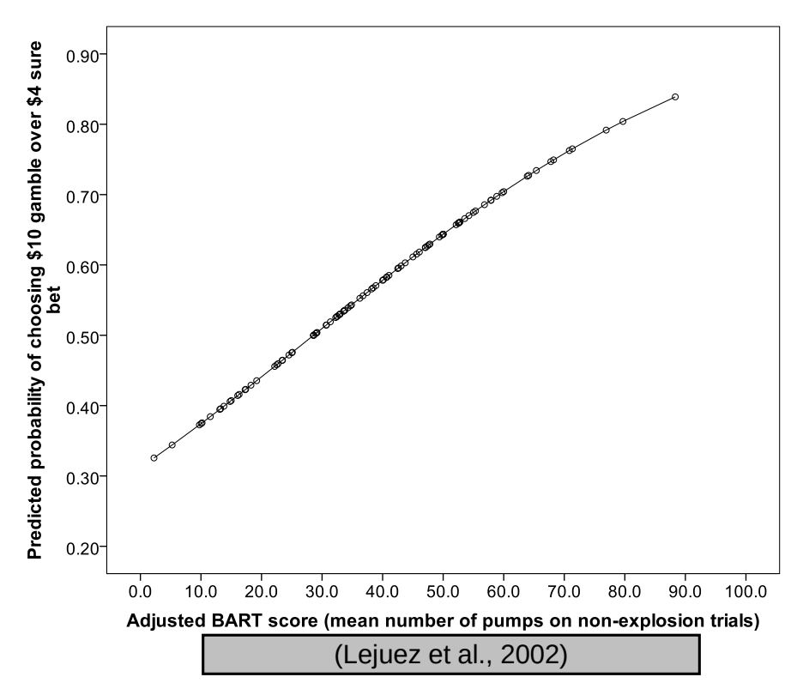
Example item: To what extent does the following trait describe you: *dependable*
(9-point scale: *1=much worse than the average UWO student of my age and gender*,
4=as well as the average UWO student, *7=much better than the average UWO student*) (Alicke et al., 1995)

Example item: "*I never regret my decisions.*"
(7-point scale: 1=*not true*, 4=*somewhat true*, 7=*very true;*
Scoring: add 1 point for each "6" or "7") (Paulhus, 1984)

Study 3

N = 99 (39 females, 58 males; mean age = 24.5, SD = 5.5, range = 17 to 46), UWO undergraduates paid $5 (CDN) + BART earnings

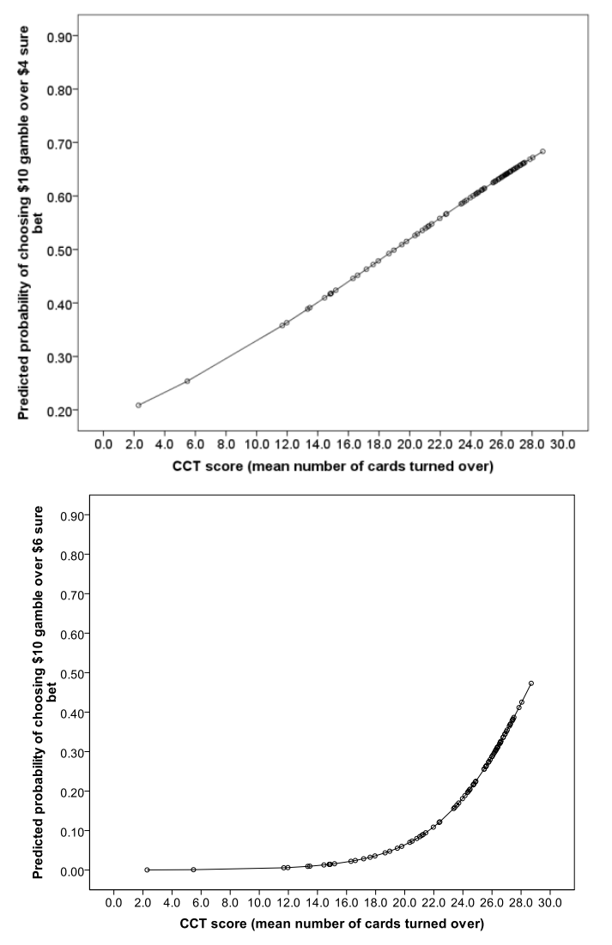
Risk-taking



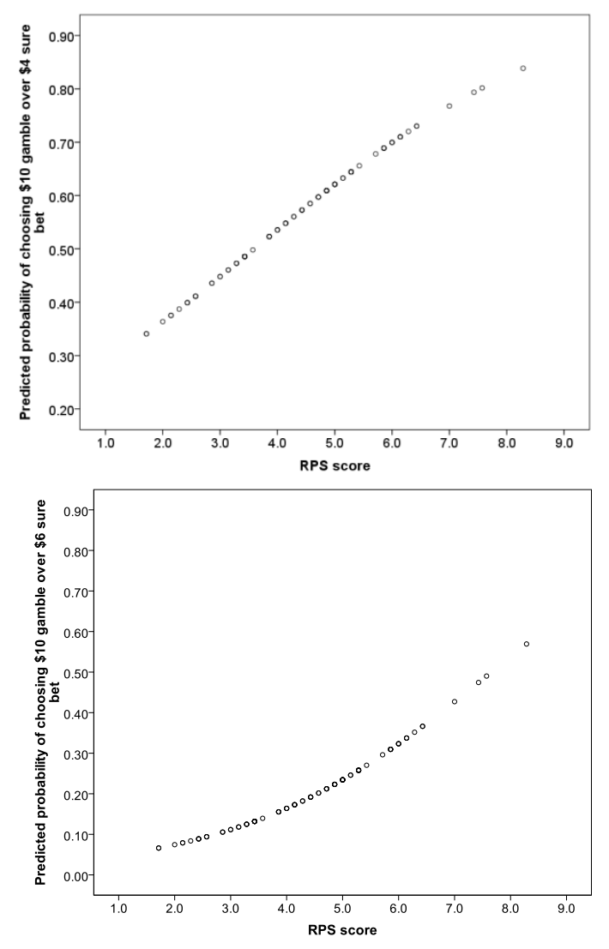
(Lejuez et al., 2002)

Behavioral referent choice: "*Which lottery option do you prefer?*"
(2 of 100 Ps actually receive monetary outcome of choice) (Hsee & Weber, 1999)

Option A: $6 for certain **Option B**: Flip a coin. Receive $10 if heads, receive $0 if tails.



(Figner et al., 2009)



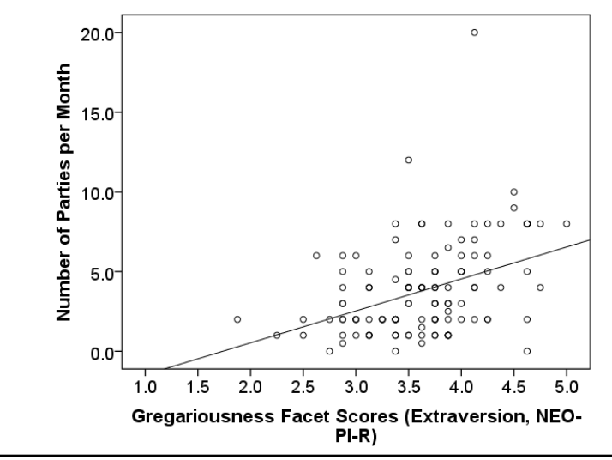
Example item: "*I prefer to avoid risks.*" (9-point scale: *totally disagree-totally agree*)
(Meertens & Lion, 2008)

Other Analyses

Sample 1: Re-analysis of Hong & Paunonen (2009)

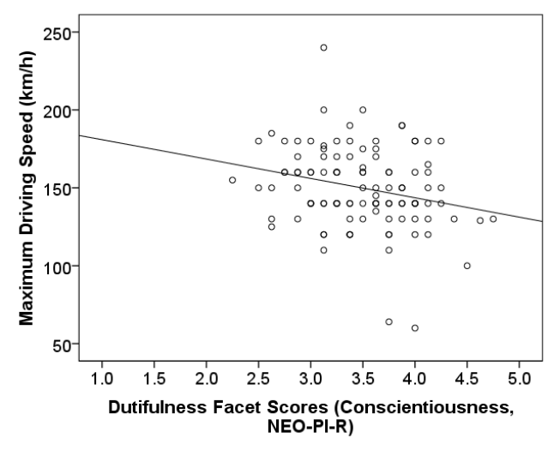
N = 124 (82 females, 42 males; mean age = 18.8, SD = .7), UWO undergraduates participated for course credit

Extraversion



Example (paraphrased) item: "*I enjoy the company of many other people.*"
(5-point scale: *strongly disagree-strongly agree*) (Costa & McCrae, 1992)

Conscientiousness

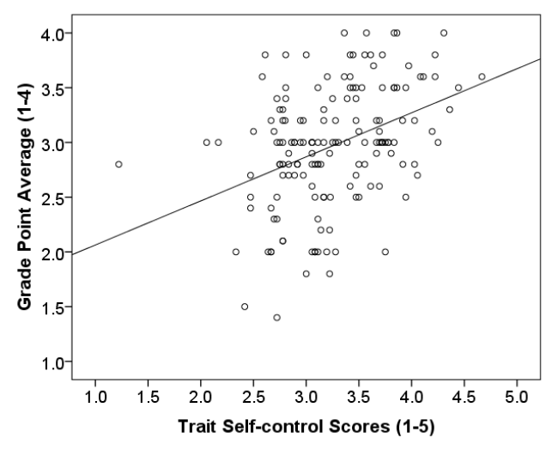


Example (paraphrased) item: "*I try to behave within my moral principles.*"
(5-point scale: *strongly disagree-strongly agree*) (Costa & McCrae, 1992

Sample 2: Re-analysis of Tangney, Baumeister, & Boone (2004)

N = 157 (113 females, 44 males; mean age = 20.0, SD = 5.0, range = 18 to 55), undergraduates from large East Coast US university for course credit

Self-control



Example item: "*I am good at resisting temptation.*"
(5-point scale: *not at all-very much*) (Tangney et al., 2004)

GENERAL DISCUSSION

Summary of Proposed & Demonstrated Benefits

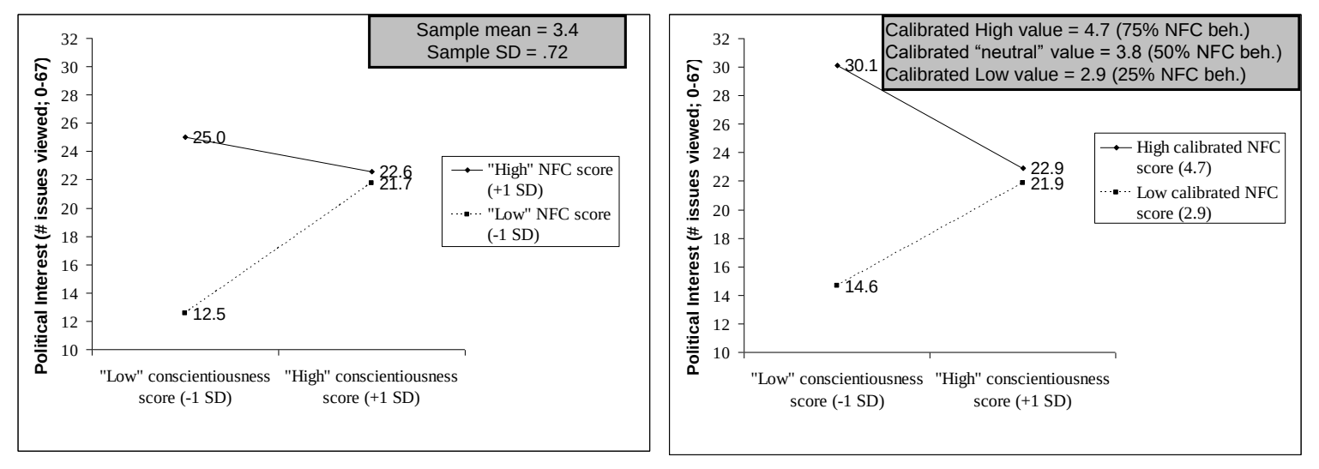
Calibrated non-arbitrary metrics could be *useful* in the following ways:

1. Help in the interpretation of data

a. Enhance the interpretability of statistical effects

Example: **Study 1 NFC**

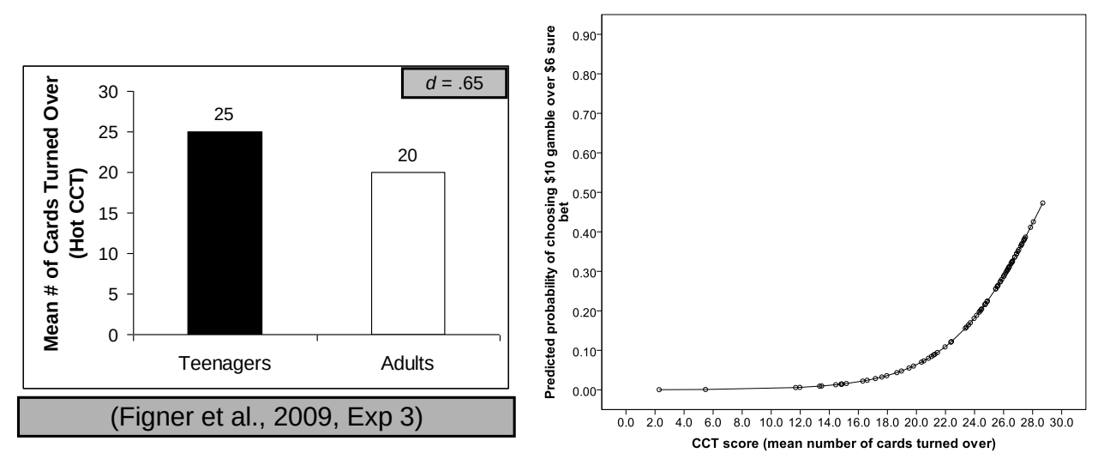
MMR re-analyses of O'Hara et al. (2009)



b. Facilitate the extraction of more information from data patterns

Example: **Study 3 CCT**

Enhance interpretation of mean difference at different locations on the scale; experimental effects found at different ranges in CCT metric would mean something different psychologically

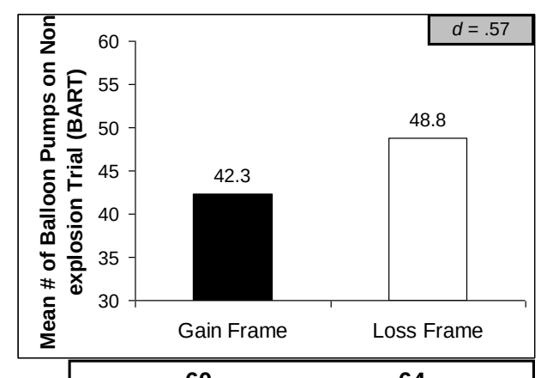


(Figner et al., 2009, Exp 3)

c. Overcome limitations of null hypothesis significance testing (NHST)

Example: **Study 3 BART**

Re-interpret Benjamin & Robbins (2007)



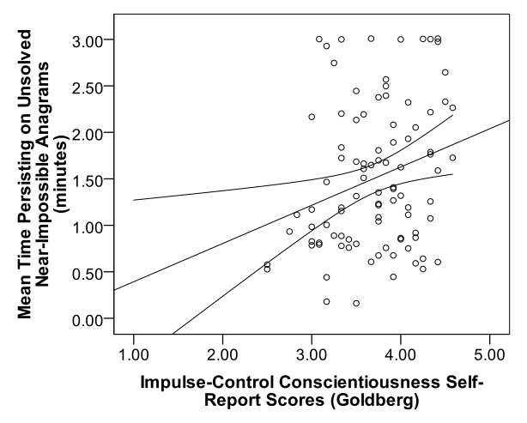
.60 **.64**
Corresponding predicted prob. of choosing risky $10 gamble over $4 safe bet

2. Facilitate construct validity research

a. Construct illumination: calibrating measure can shed more light on a construct

Example:

(Study 1 conscientiousness== task persistence)



b. Help with construct definition and construct theory: calibrating measure may help clarify conceptual ambiguities (e.g., whether construct definition too broad or narrow)

Example: Study 1 conscientiousness

Failed to find metric linkages between four different conscientiousness facets and meaningful conscientiousness behavior (# of errors found in essay task)

c. Behavioral reference points could provide measurement benchmark for improving measures (and/or detecting problems)

Example: **Study 1 task-persistence self-report**

3. Contribute to theoretical development

a. Aid (and allow) theoretical debates involving absolute claims

Example: **Study 2 self-enhancement**

b. Allow for more precise theorizing in our scientific language

Example: "…high-SE individual possess self-doubts and insecurities…"

Unsubstantiated claims and potentially misleading, given they are based on scores with non-calibrated metrics; this impedes accurate theorizing and interferes with theory development.

c. Allow (or provide platform) for quantitative testing of theories (Meehl, 1978)

First step for point value predictions is to make our metrics meaningful (i.e., non-arbitrary)

4. Facilitate general accumulation of knowledge

a. Metric calibration findings are valuable information in their own right

b. Metric calibration approach as guiding framework for cataloguing the quantity/magnitude of psychological effects

c. Could also facilitate phenomenon-based research (Rozin, 2001)

Limitations/Caveats

- Preliminary demonstrations: Calibration studies requires larger targeted samples
- Consensus required for behavioral reference points
- Conceptual hurdles to overcome (e.g., multiple reference points, features of ideal beh. fixed point)

Future directions

- Experimental approach to metric calibration
- Within-subjects approach using state-space models (Commandeur & Koopman, 2007)
- Richer methodology for behavioral reference points (e.g., eye-tracking, Microsoft SenseCam, EAR, observational studies)

References:
Alice, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average-effect. *Journal of Personality and Social Psychology, 68*, 804-825.
Benjamin, A. M., & Robbins, S. J. (2007). The role of framing effects in performance on the Balloon Analogue Risk Task (BART). *Personality and Individual Differences, 43,* 221-230.
Blanton, H., & Jaccard, J. (2006a). Arbitrary metrics in psychology. *American Psychologist, 61*, 27-41.
Blanton, H., & Jaccard, J. (2006b). Arbitrary metrics redux. *American Psychologist, 61*, 62-71.
Brandon, T. H., Herzog, T. A., Juliano, L. M., Irvin, J. E., Lazev, A. B., & Simmons, V. (2003). Pretreatment task persistence predicts smoking cessation outcome. *Journal of Abnormal Psychology, 112*, 448−456.
Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306-307.
Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.
Commandeur, J. J. F., & Koopman, S. J. (2007). *An introduction to state space time series analysis.* Oxford: Oxford University Press.
Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.
Ditre, J. W., & Brandon, T. H. (2008). Does self-reported task persistence predict performance on behavioral measures of task persistence and distress tolerance? Paper presented at the meeting of the Society for Research on Nicotine and Tobacco, Portland, OR, USA.
Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: Age differences in risk taking in the Columbia Card Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 709-730.
Hong, R. Y., & Paunonen, S. V. (2009). Personality traits and health-risk behaviours in university students. *European Journal of Personality, 23,* 675-696.
Hsee, C. K. & Weber, E. U. (1999). Cross-national differences in risk preferences and lay predictions for the differences. *Journal of Behavioral Decision Making, 12*, 165-179.
Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8*, 75–84.
Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806-834.
Meertens, R. M., & Lion, R. (2008). Measuring an individual's tendency to take risks: The risk propensity scale. *Journal of Applied Social Psychology, 38*, 1506-1520.
O'Hara, R. E., Walter, M. I., & Christopher, A. N. (2009). Need for cognition and conscientiousness as predictors of political interest and voting strategy. *Journal of Applied Social Psychology, 39,* 1397-1416.
Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609.
Paulhus, D.L., Harms, P.D., Bruce, M.N., & Lysy, D.C. (2003). The over-claiming technique: Measuring bias independent of accuracy. *Journal of Personality and Social Psychology, 84*, 681-693.
Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5*, 2–14.
Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *American Psychologist, 51*, 1065-1071.
Sedikides, C., Gaertner, L. & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology, 84*, 60–70.
Steinberg, M. L., Krejci, J. A., Collett, K., Brandon, T. H., Ziedonis, D.M., & Chen, K. (2007). Relationship between self-reported task persistence and history of quitting smoking, plans for quitting smoking, and current smoking status in adolescents. *Addictive Behaviors, 32*, 1451−1460.
Tangney, J. P., Baumeister, R. F., & Boone, A. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality,72,* 271-324.
Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83-91.

Etienne LeBel & Bertram Gawronski
Department of Psychology
The University of Western Ontario
London ON, Canada
elebel@uwo.ca – bgawrons@uwo.ca