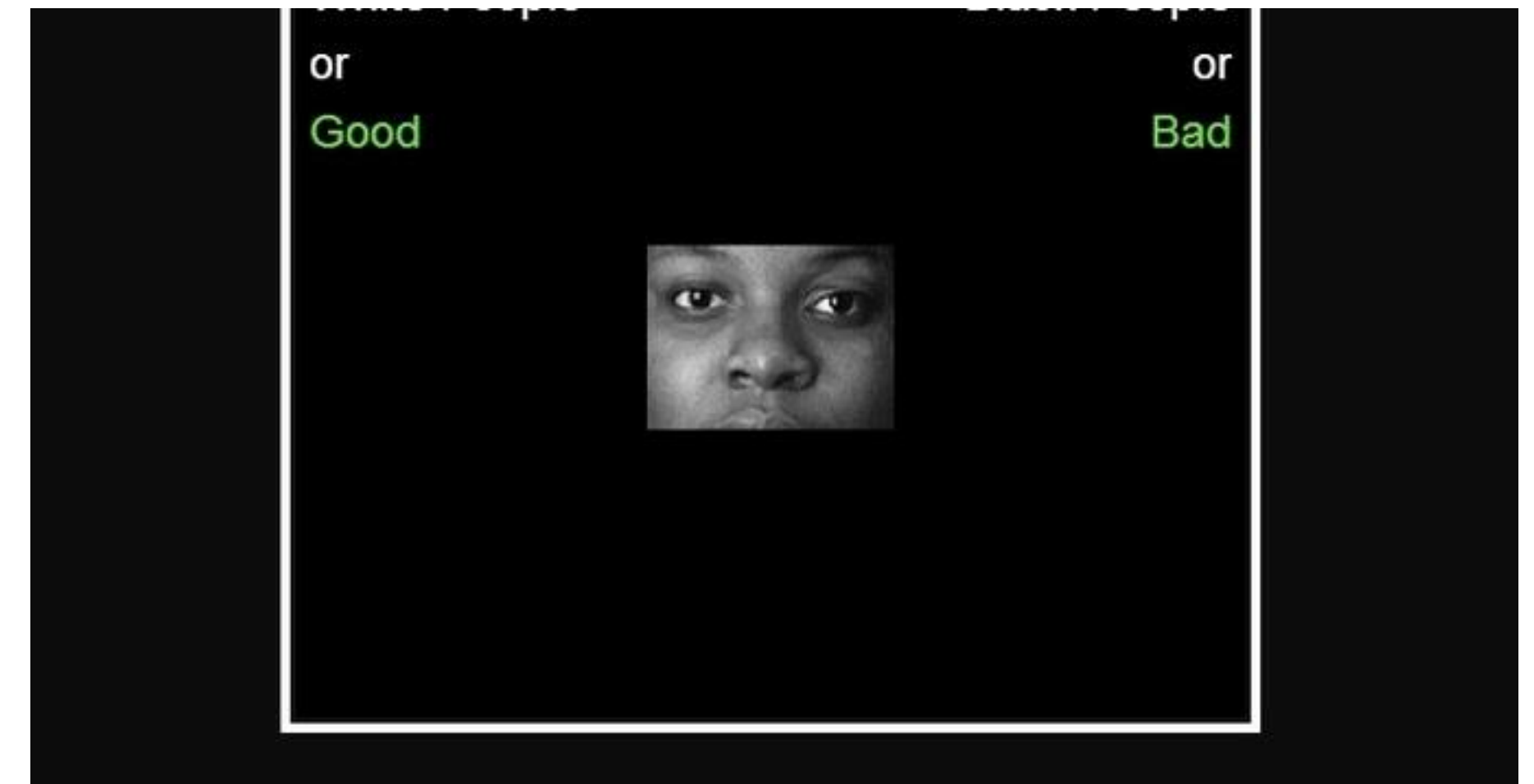
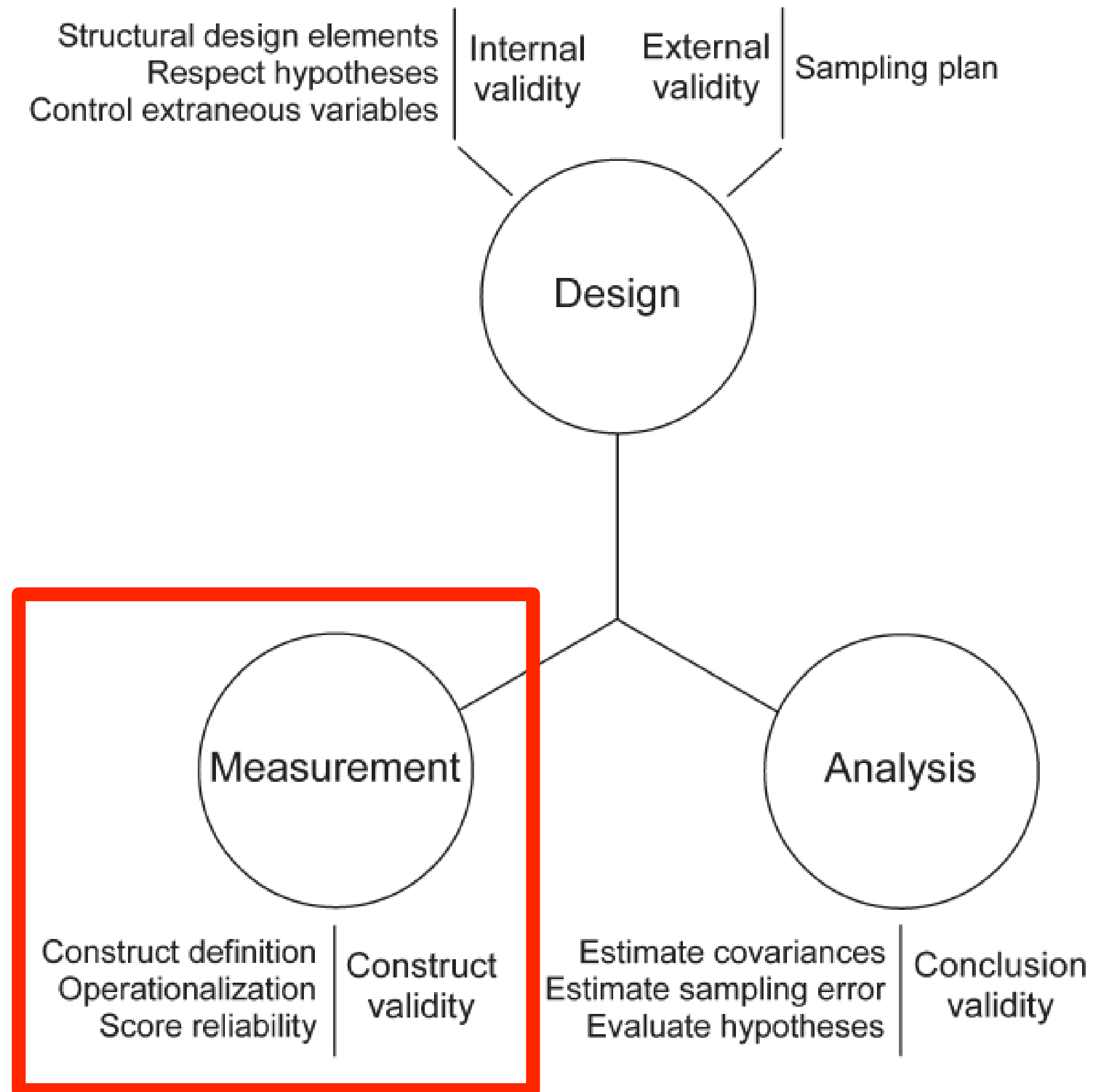


The IAT crisis within the credibility crisis



Marco Tullio Liuzza, Ph.D
University of Padua

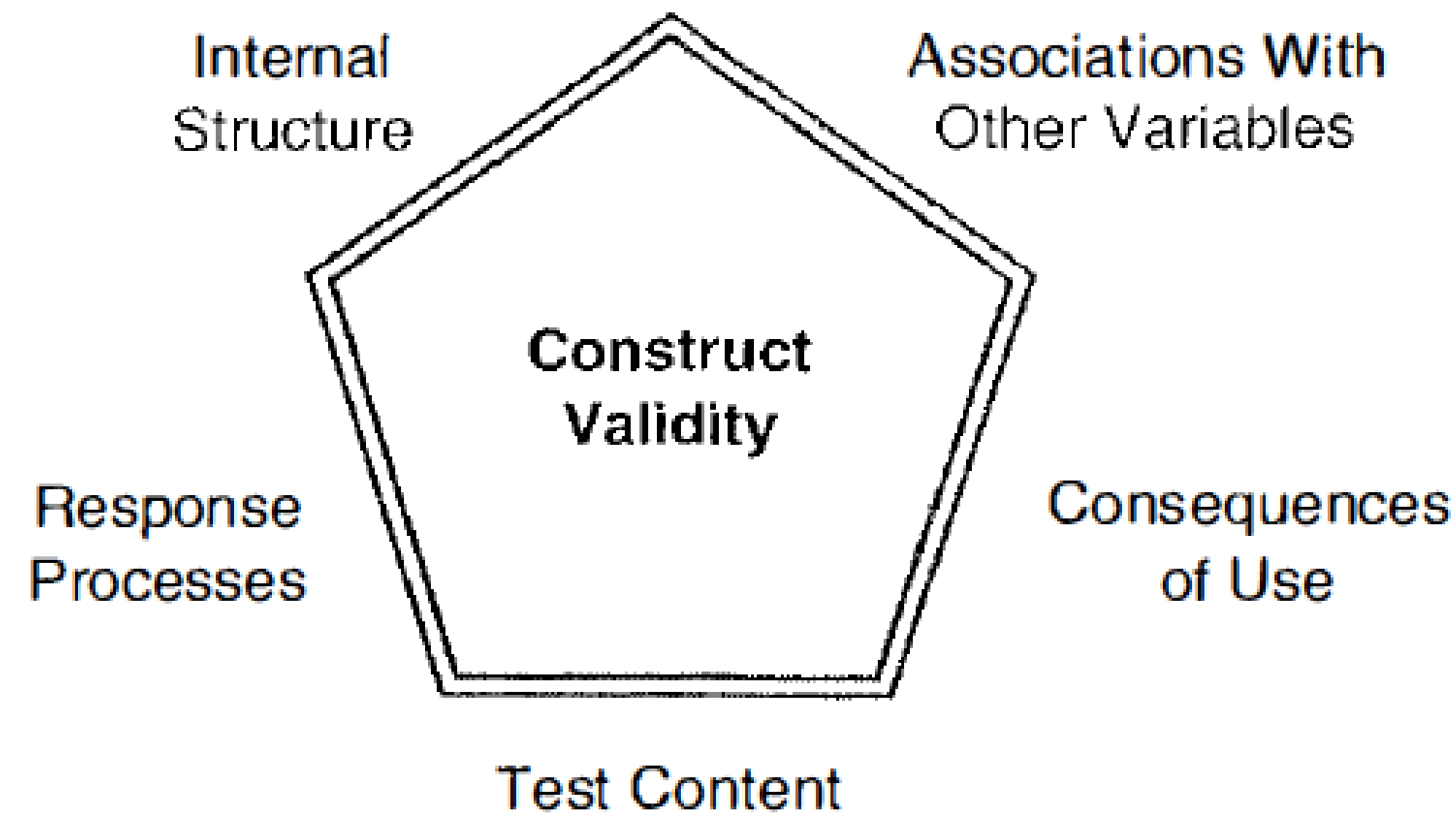
Research Trinity



- Most of published studies provide little information about validity (Flake, Pek, Hehman, 2017) and *v-hacking* (Hussey and Hughes, 2018).
- *Measurement schmeasurement: wide questionable measurement practices in Psychology* (Flake and Fried, 2019)

Validity crisis in Psychology definitions

- "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses" of **a test** (AERA, APA, and NCME, 1999, in Furr, R. M., 2017. *Psychometrics: an introduction*. Sage Publications, p. 9.)



Validity crisis in Psychology

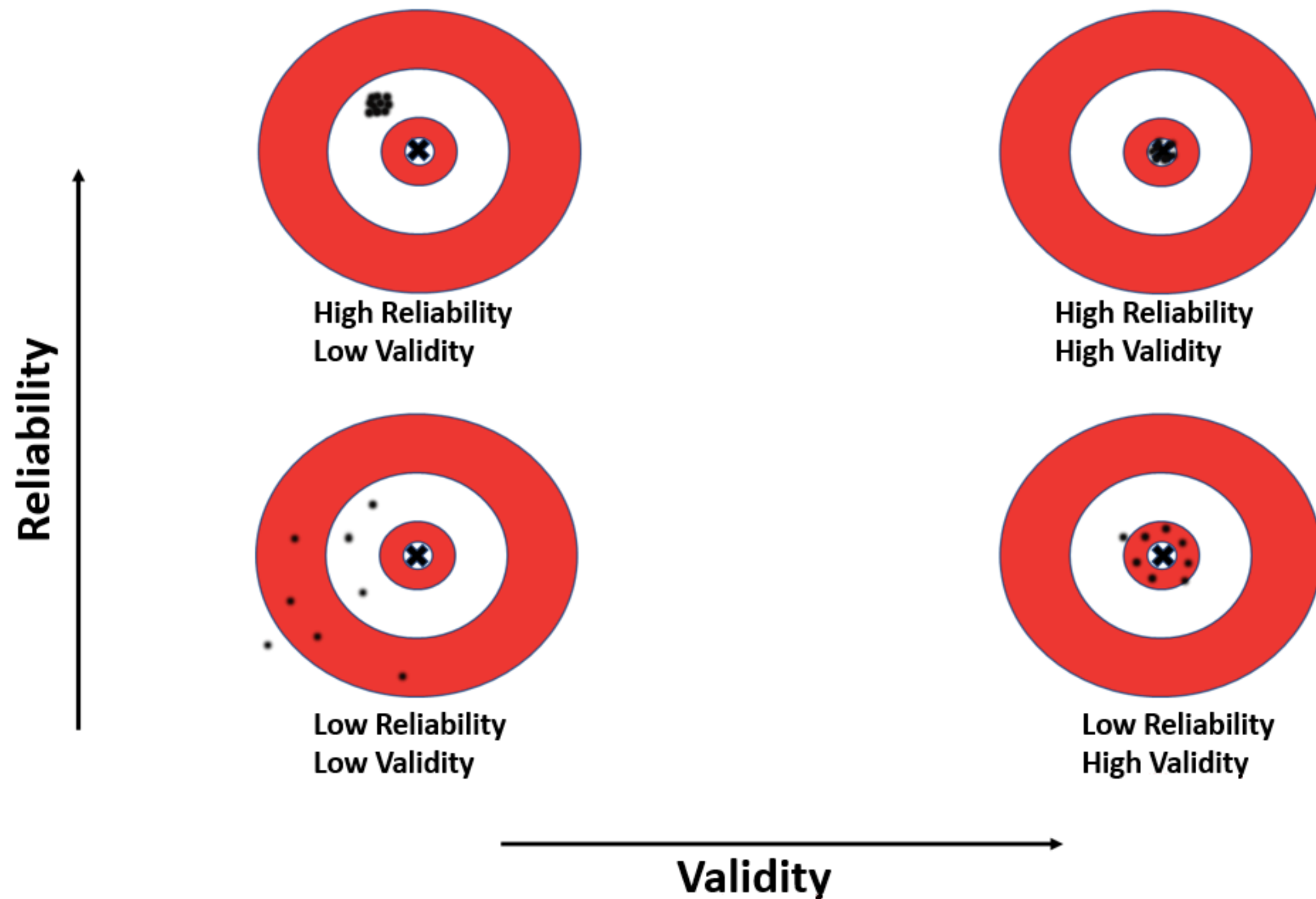
Table 1. Examples of Validity Evidence and Resources for Each Phase of Construct Validation.

Phase	Validity Evidence	Description
Substantive	Literature review and construct conceptualization	Identifying depth and breadth of construct (Gehlbach & Brinkworth, 2011)
	Item development and scaling selection	Expert review (Gehlbach & Brinkworth, 2011)
	Content relevance and representativeness	Item mapping (Dawis, 1987), focus groups, and cognitive interviewing (i.e., think aloud; Willis, 2004), investigate construct under representation or irrelevancy (i.e., content validity; Sireci, 1998)
Structural	Item analysis	Response distributions, item–total correlations, and difficulty
	Factor analysis	Exploratory and confirmatory analyses including structural equation models and item response theory
	Reliability	Coefficients: α and ω (McDonald, 1999); interitem correlations, test–retest (McCrae, Kurtz, Yamagata, & Terracciano, 2011), dependability (Chmielewski & Watson, 2009)
	Measurement invariance (i.e., differential item functioning) testing	Multiple group factor analysis, item response theory, and differential item functioning tests (Millsap, 2011)
External	Convergent and discriminant	Correlations between other scales meant to capture similar and different constructs, multitrait-multimethod matrix analyses (Campbell & Fiske, 1959)
	Predictive/criterion	Regressions on criterion variables of import
	Known groups	Detecting differences between groups known to differ on construct

Note. Table draws from a collection of seminal works and texts on validation and measurement more broadly including Benson (1998), Clark and Watson (1995), Crocker and Algina (2006), Loevinger (1957), Strauss and Smith (2009), and Raykov and Marcoulides (2011).

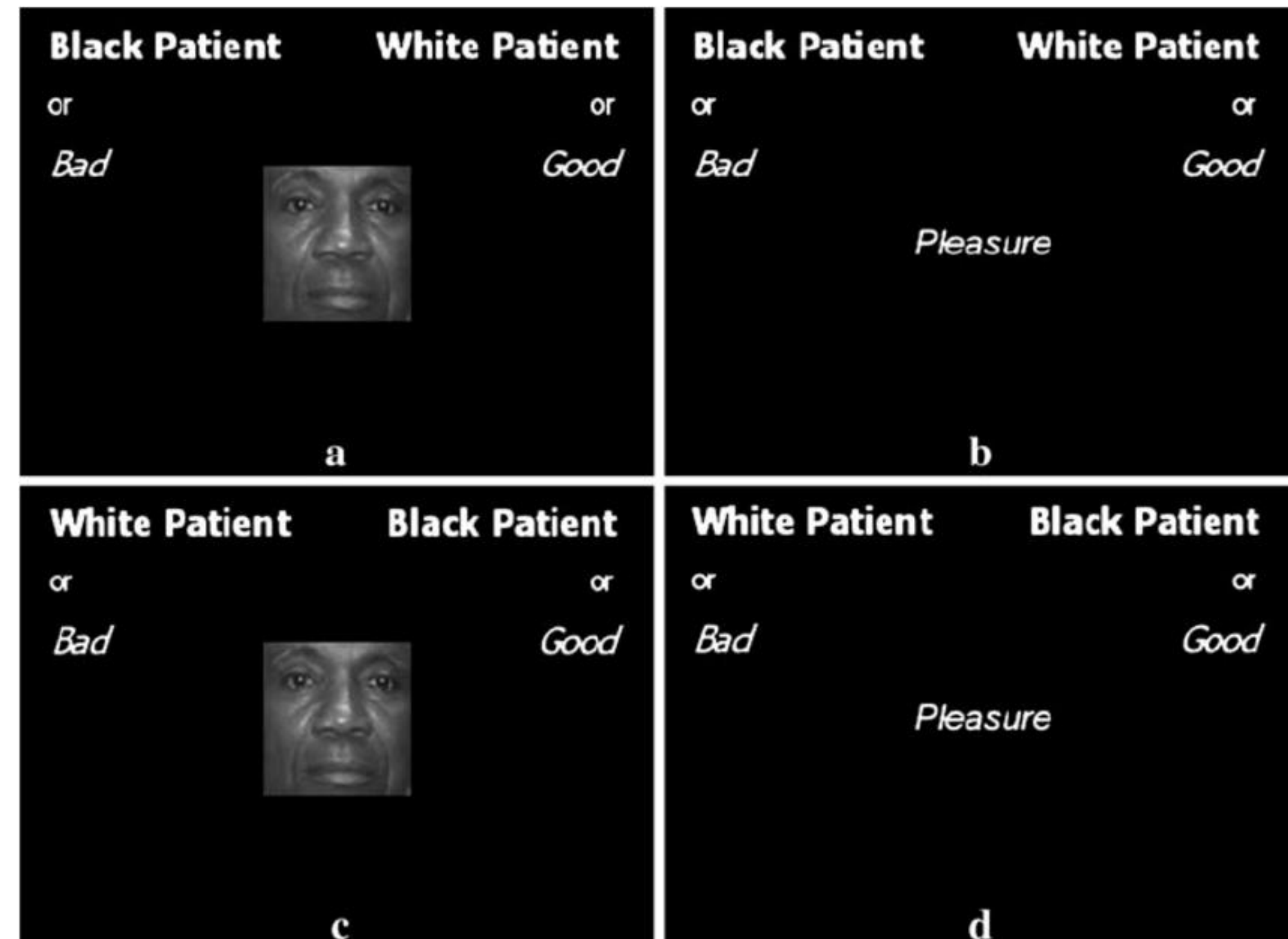
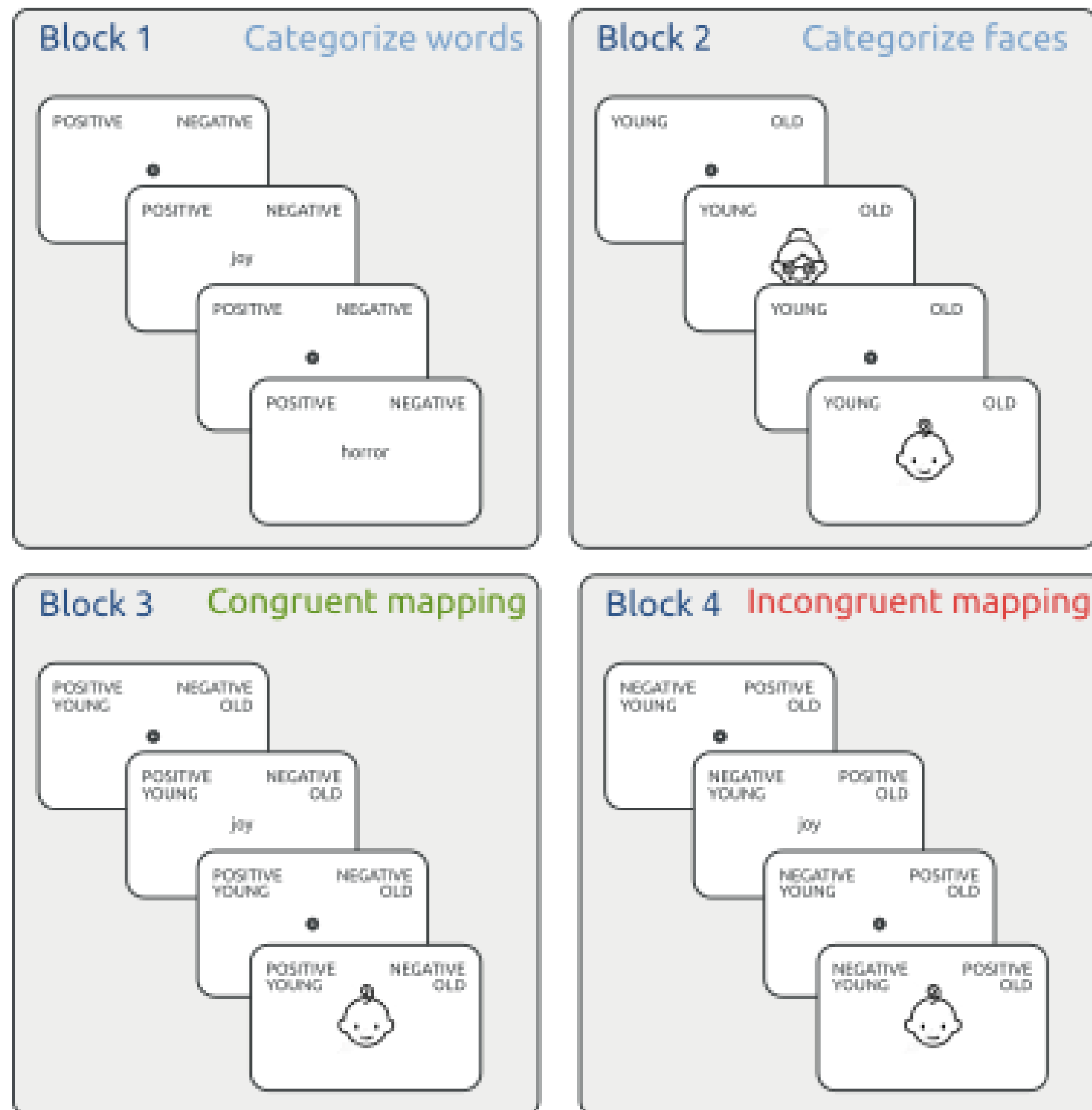
Flake et al. 2017

Reliability sets the upper limit of validity



- Temporal Stability
- Internal consistency

IAT, a Method in Search of a construct



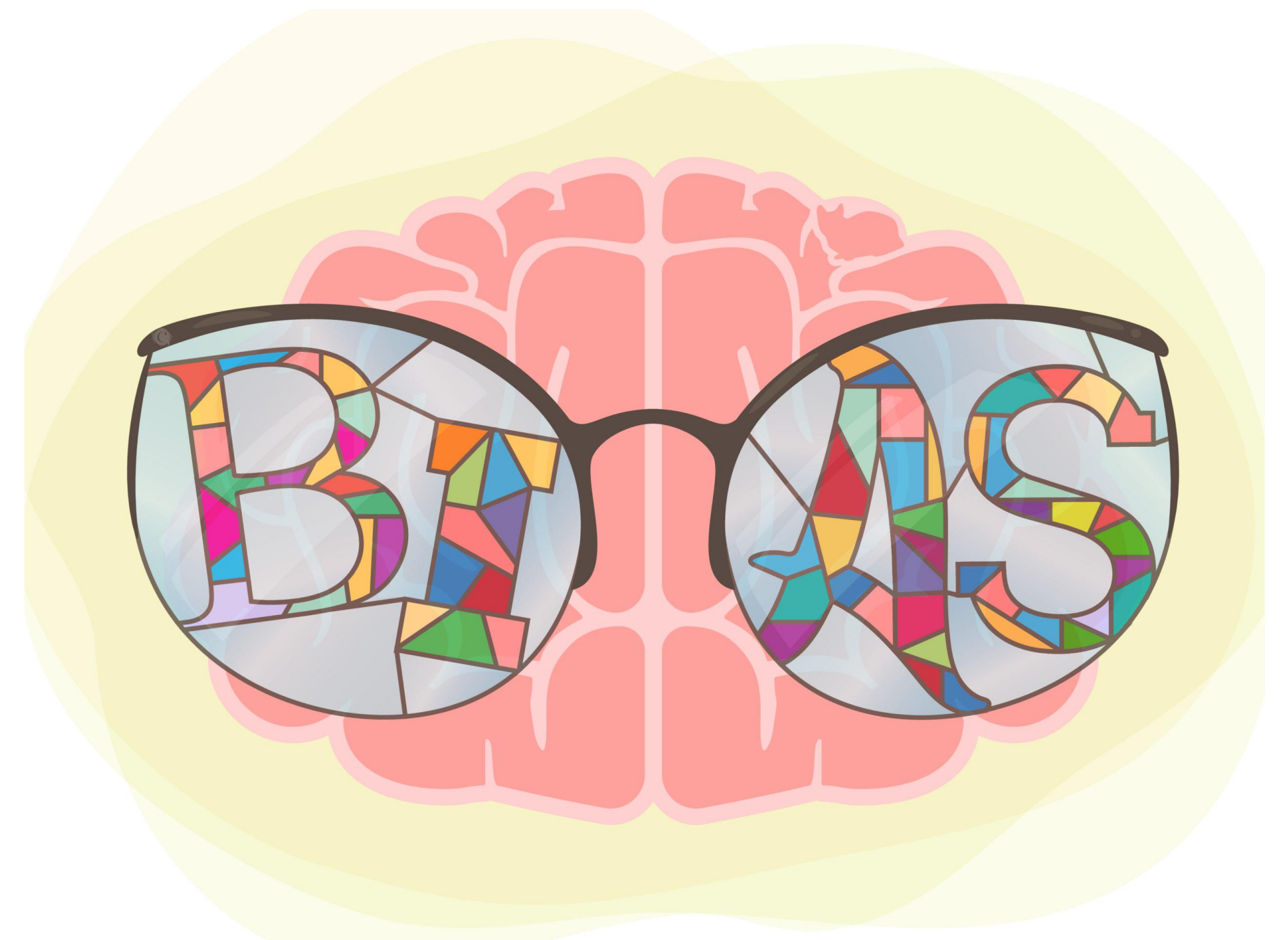
IAT, a Method in Search of a construct

Implicit attitudes

Not a single formal theory

But can be treated as a body of thought

Defined by research that discusses behaviour emitted within so-called implicit measures



Credit: Dr. Ian Hussey, The best theory is a flawed one, 14/02/2022 to RIOT Science Club

IAT, a Method in Search of a construct

Key claims

1. Implicit measures measure unconscious associations
2. Uniquely predictive of overt behaviour
3. Reliable & valid measures
4. Acquired via unconscious experience

Fazio & Olson (2003), Greenwald et al. (1995), Olson & Fazio (2001)

Credit: Dr. Ian Hussey, The best theory is a flawed one, 14/02/2022 to RIOT Science Club

IAT, a Method in Search of a construct

“Implicit measures
measure **unconscious** associations”

Do implicit measures measure the unconscious?

- No: “Implicit \neq Unconscious”
- Greenwald & Banaji (2017), Greenwald & Lai (2020)
- So much confusion about what ‘implicit’ does refer to
- Corneille & Hutter (2021) Implicit? What do you mean?

Credit: Dr. Ian Hussey, The best theory is a flawed one, 14/02/2022 to RIOT Science Club

IAT, a Method in Search of a construct

**“Implicit measures
measure unconscious associations”**

“Unavailable to self-report or introspection”? Greenwald & Lai (2020)

- No, they can accurately report them
- Hahn et al. (2014)

Expressed without conscious awareness?

- No, effects highly correlated with level of awareness
- Hughes, Hussey & Cummins (2021)

Acquired without conscious awareness?

- No

Credit: Dr. Ian Hussey, The best theory is a flawed one, 14/02/2022 to RIOT Science Club

IAT, a Method in Search of a construct

**“Implicit measures
measure unconscious associations”**

What is it we’re even trying to measure?

- Purely associative mental representations (Fazio, Bargh)
- Purely propositional mental representations (De Houwer)
- Dual process representations (Gawronski)
- Who cares: focus on predictions (Greenwald, Nosek)

Credit: Dr. Ian Hussey, The best theory is a flawed one, 14/02/2022 to RIOT Science Club

IAT, a Method in Search of a construct

- IAT reliably shows group differences
- But is it sufficient to claim that measures individual differences?

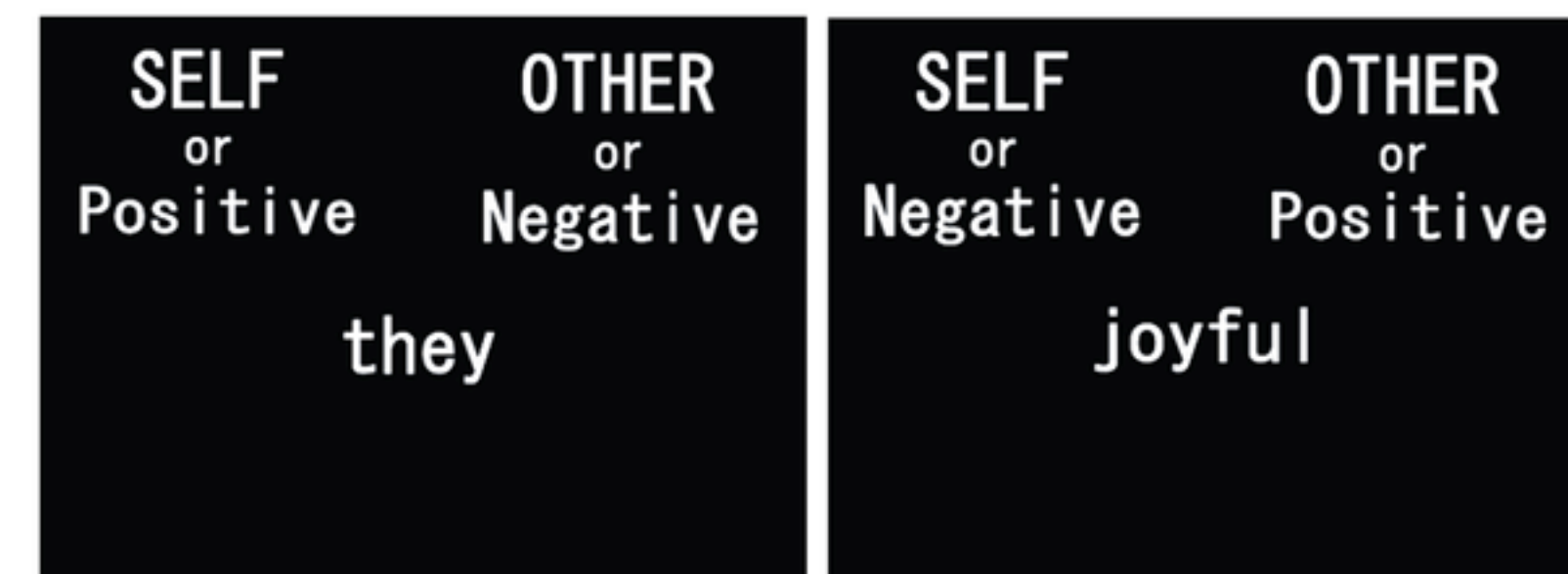
FOUR SCENARIOS

1. No stable attributes that influence performance on the IAT (Payne et al. 2017)
2. IAT is a poor measure of stable attributes (Falk & Heine, 2017)
3. IAT is good measure of stable attributes, but it is just as good as explicit measures (Samayoa and Fazio, 2017)
4. IAT is good measure of stable attributes THAT CANNOT be tested by explicit measures (Greenwald et al., 1998)

The example of low correlation between implicit and explicit SE

THREE SCENARIOS

- 1.They measured different constructs -> but this should be measured by a MTMM approach
- 2.The Rosenberg SE has low validity (very unlikely)
- 3.The IAT has low validity



Congruent (C) condition

Incongruent (IC) condition

A CFA for a dual model construct validity

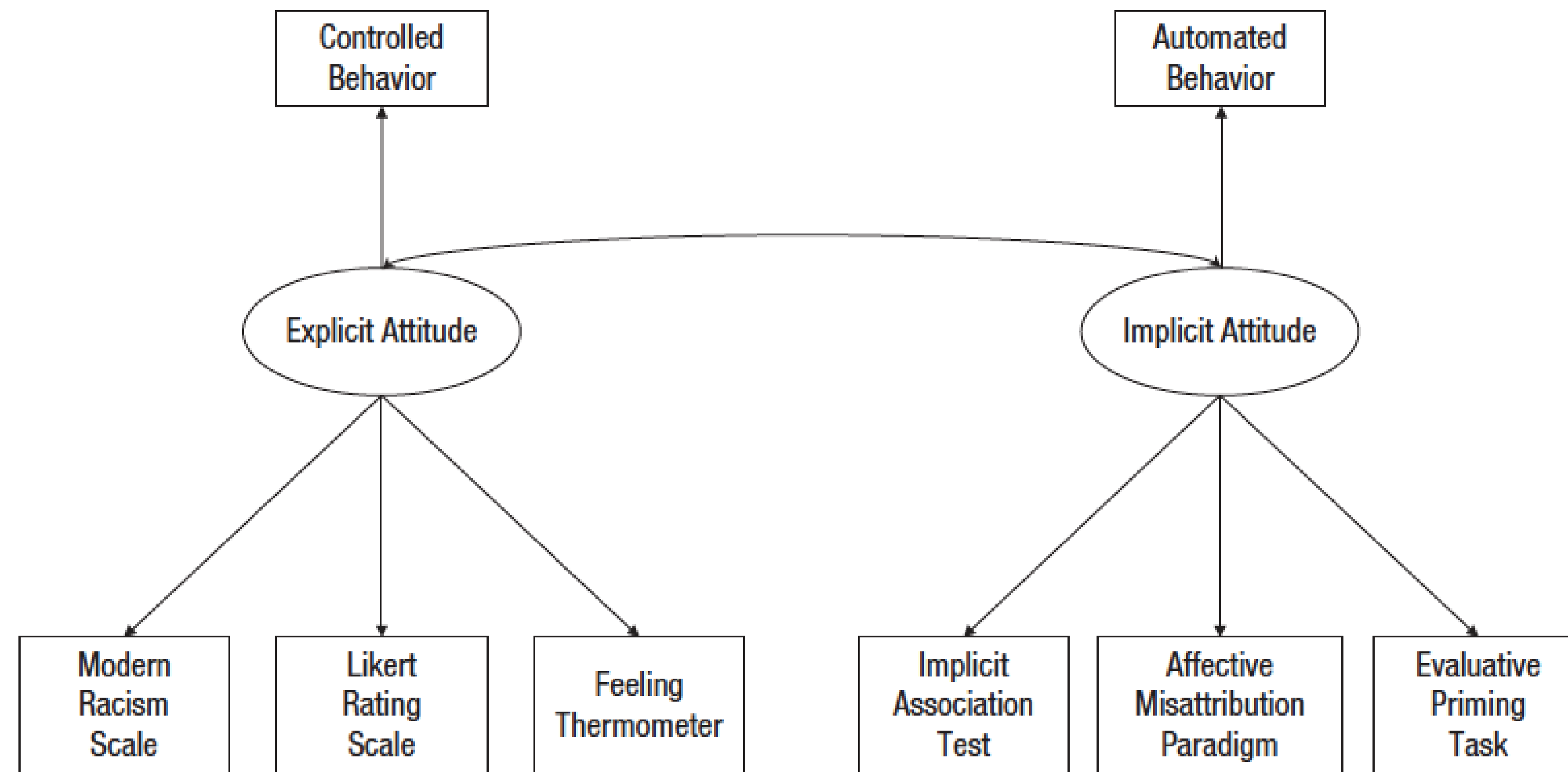
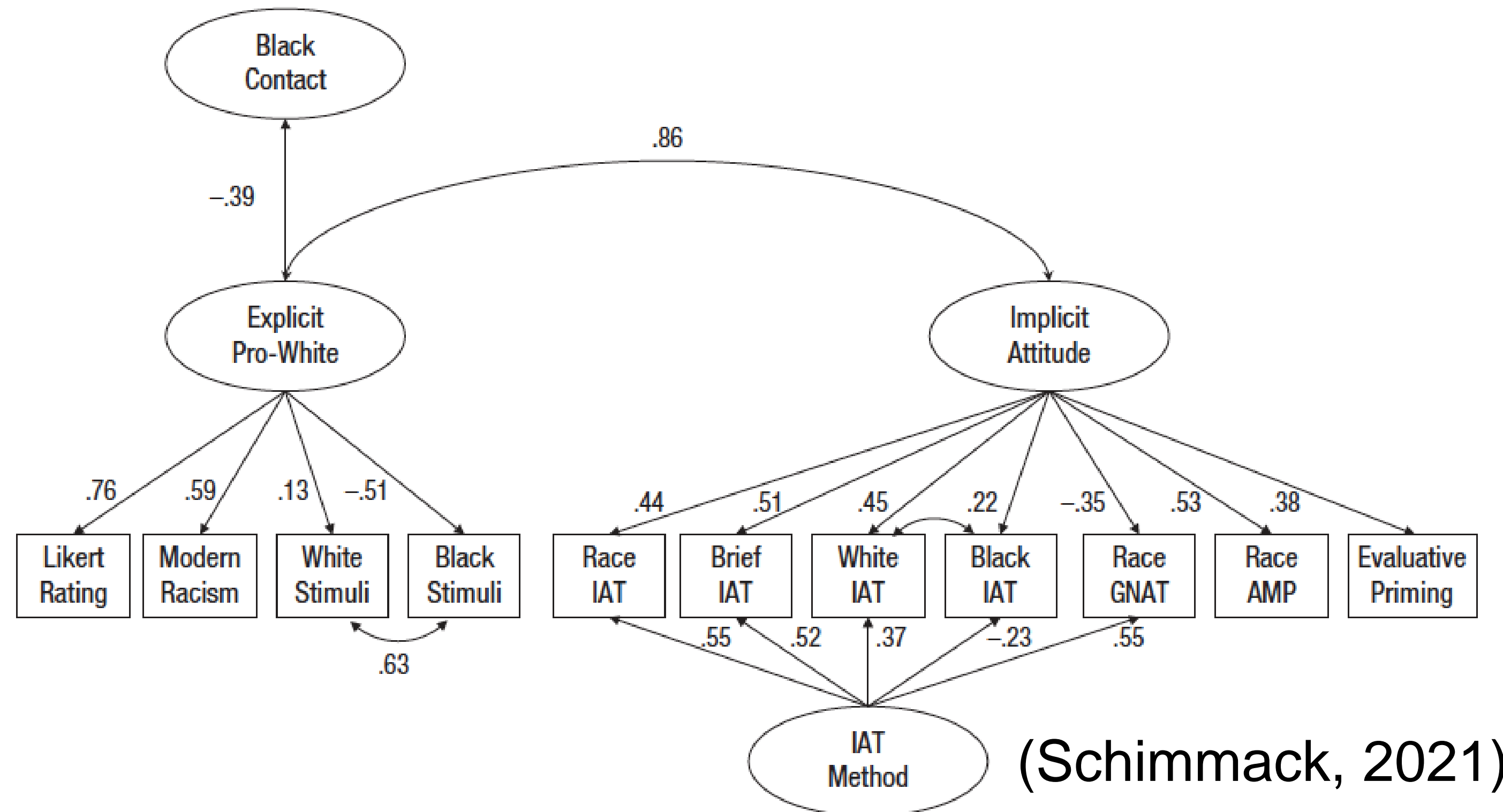


Fig. 1. Hypothetical dual-attitude model.

A CFA for a dual model construct validity



Noteworthy findings

- Very high correlation btw the LVs
- Explicit measures predict contact

Fig. 3. Reanalysis of Bar-Anan and Vianello (2018) racial attitude measures. IAT = implication association test; AMP = affective misattribution paradigm; GNAT = go/no-go association task.

A CFA for a dual model construct validity

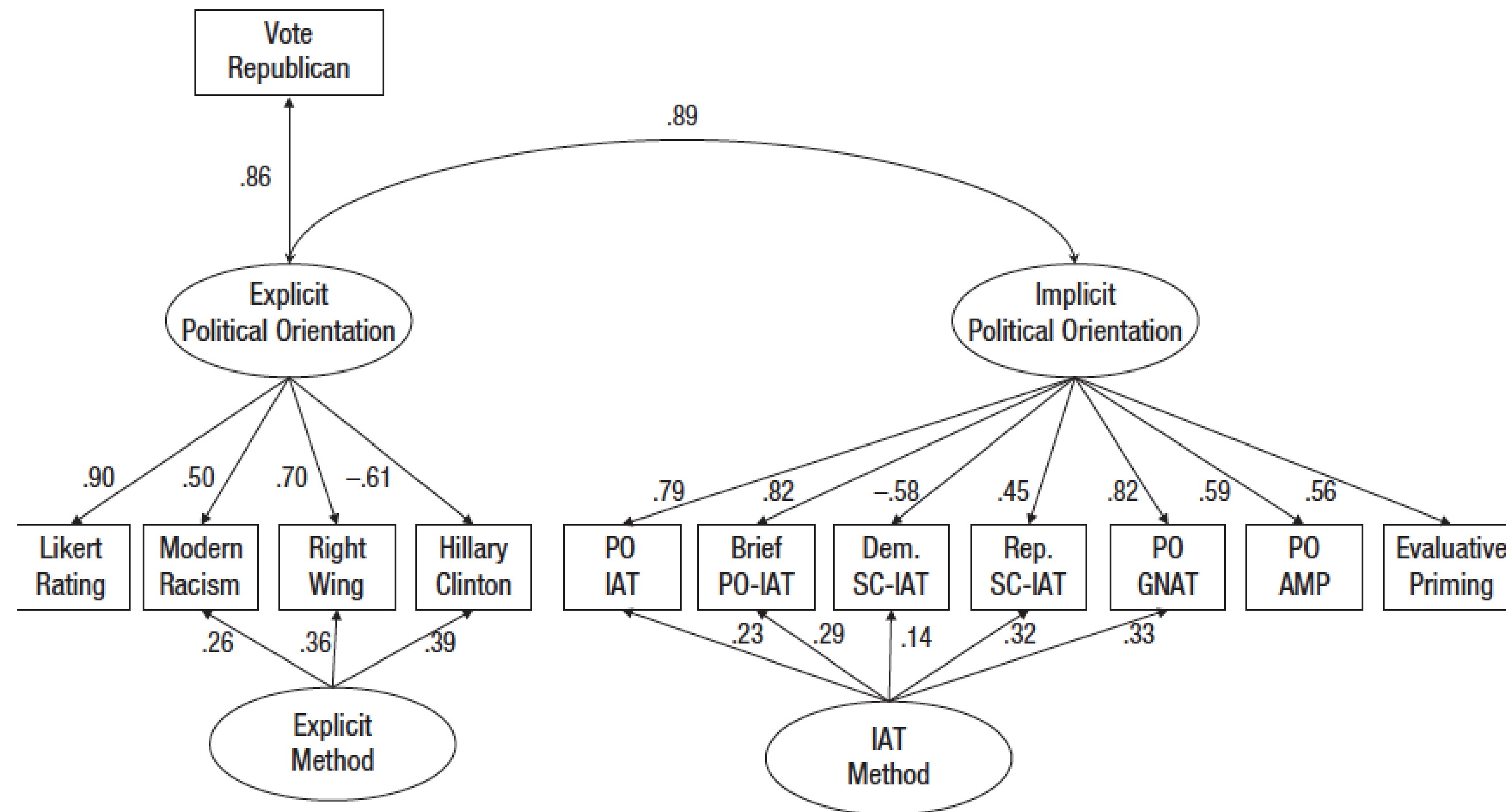


Fig. 8. Analysis of Bar-Anan and Vianello (2018). IAT = implicit association test; AMP = affective misattribution paradigm; GNAT = go/no-go association task; SC = single category.

Noteworthy findings

- Very high correlation btw the LVs
- Explicit measures predict Vote

(Schimmack, 2021)

A CFA for a unidimensional model construct validity

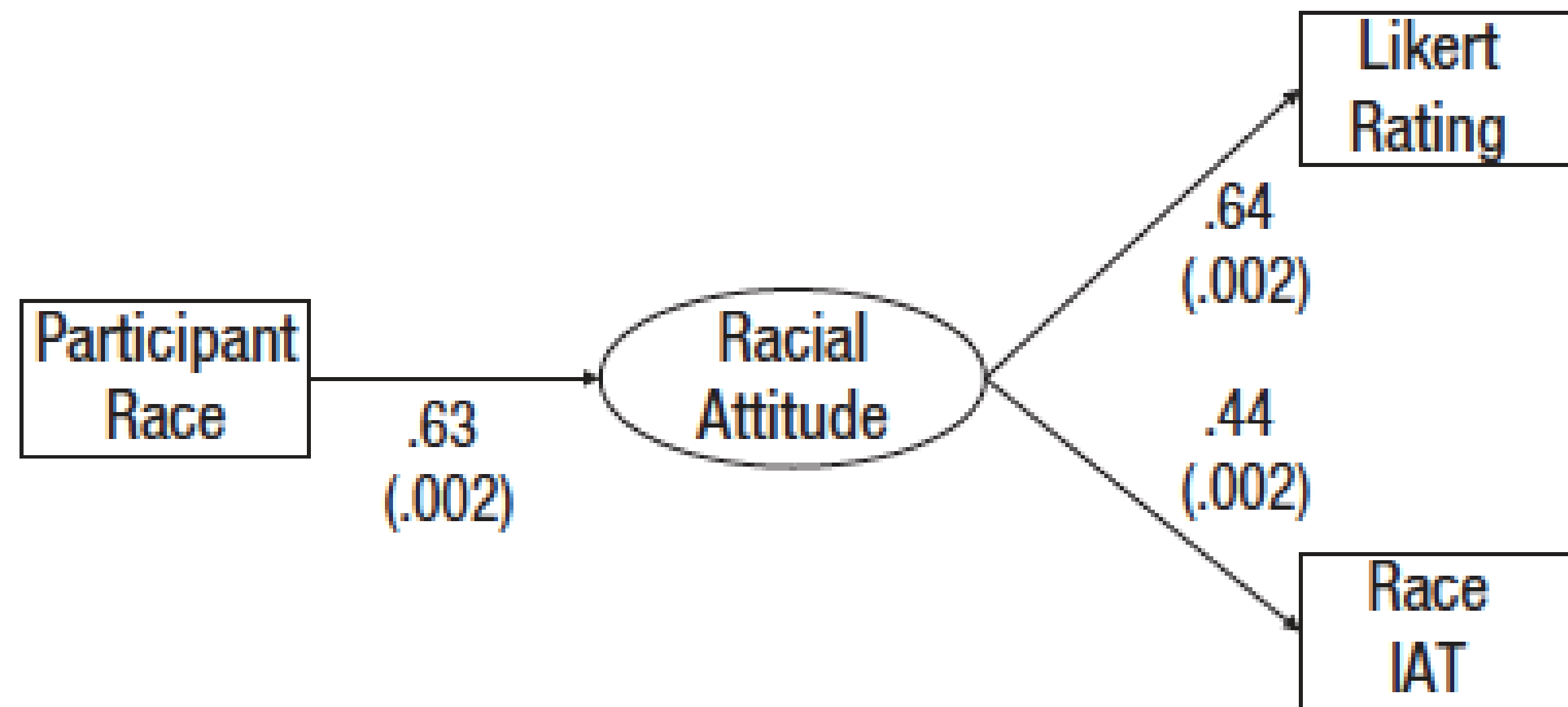


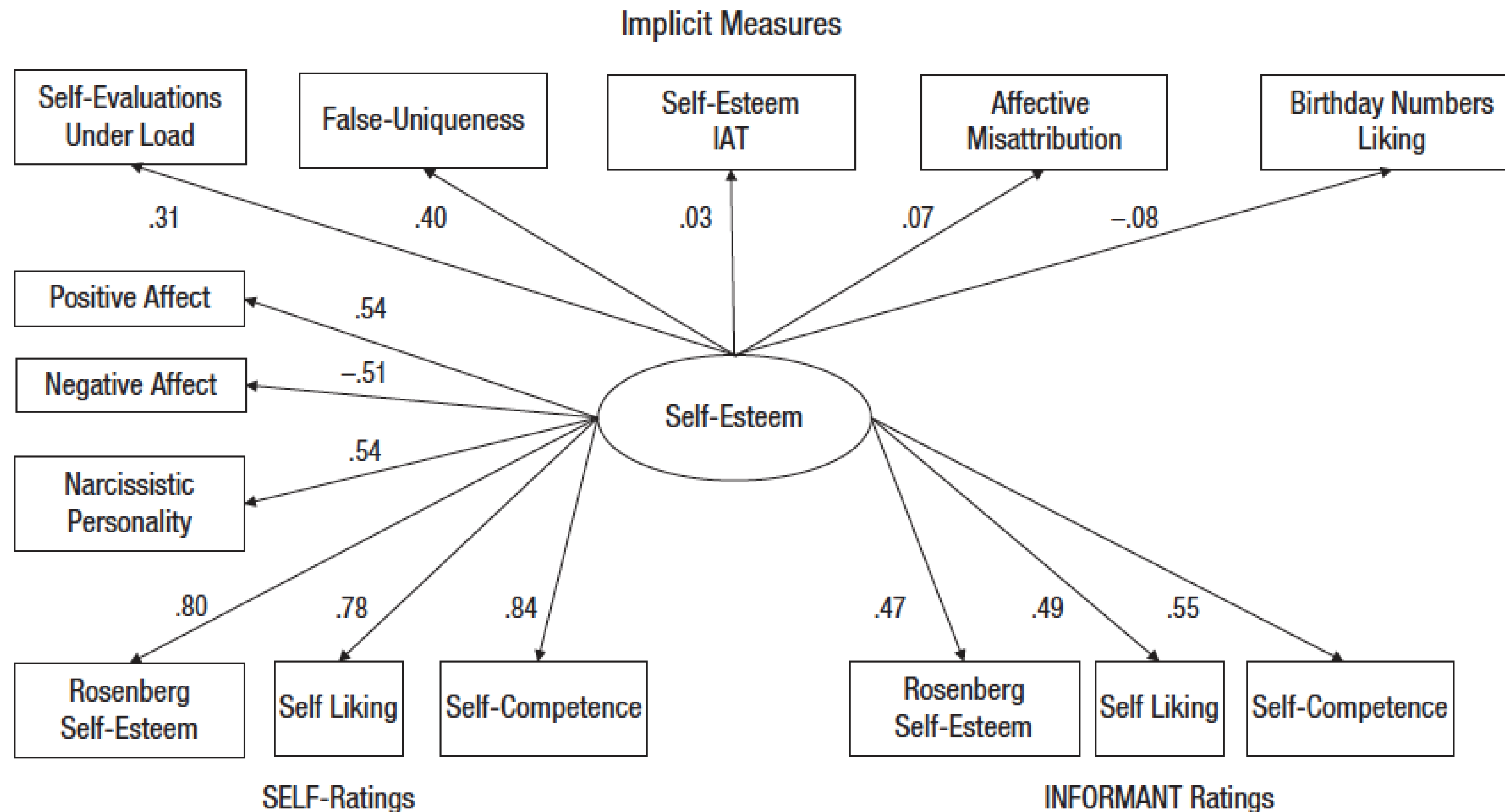
Fig. 5. Analysis of Axt (2018). IAT = implicit association test. Values in parentheses are standard errors.

Noteworthy findings

- Likert ratings loaded *more* onto the racial attitude LV

(Schimmack, 2021)

A CFA for a unidimensional model construct validity



Noteworthy findings

- Explicit measures loaded *more* onto the Self Esteem LV
- Implicit measures loadings are *not* even statistically significant

(Schimmack, 2021)

Fig. 6. Analysis of Falk, Heine, Takemura, Zhang, and Hsu (2015). IAT = implicit association test.

Not all IATs are created equal



- SE IAT is not valid
- Race IAT is 20% valid
- Political IAT is 40% valid

(Schimmack, 2021)



What about criterion and incremental validity?

Predictive of overt behaviour?

- Independently:

- Not very usefully (Kurdi et al., 2019)

“debates about whether implicit cognition and behavior are related to each other are unlikely to offer any meaningful conclusions.”

r 90% prediction interval = [-0.14, .32]

Credit: Dr. Ian Hussey, The best theory is a flawed one, 14/02/2022 to RIOT Science Club

What about criterion and incremental validity?

Journal of Personality and Social Psychology

Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies

Frederick L. Oswald, Gregory Mitchell, Hart Blanton, James Jaccard, and Philip E. Tetlock
Online First Publication, June 17, 2013. doi: 10.1037/a0032734

Table 1
Meta-Analysis of Implicit-Criterion Correlations (ICCs): Overall and by Subgroups

Criterion	<i>k</i> (<i>s</i> ; <i>N</i> _{total})	$\hat{\rho}$ [95% CI]	$\hat{\tau}$	<i>M</i>	<i>SD</i>
All effects: Overall	298 (86; 17,470)	.14 [.10, .19]	.17	.12	.24
Interpersonal behavior	11 (6; 796)	.14 [.03, .26]	.12	.21	.15
Person perception	138 (46; 7,371)	.13 [.07, .18]	.13	.10	.21
Policy preference	21 (9; 4,677)	.13 [.07, .19]	.03	.14	.09
Microbehavior _a	96 (21; 3,879)	.07 [−.03, .18]	.19	.10	.24
Response time	6 (5; 300)	.19 [.02, .36]	.27	.31	.28
Brain activity _a	26 (8; 447)	.42 [.11, .73]	.68 ^b	.26	.40
Black vs. White groups: Overall	206 (63; 9,899)	.15 [.09, .21]	.19	.13	.26
Interpersonal behavior	10 (5; 691)	.14 [.01, .28]	.14	.22	.16
Person perception	75 (30; 3,564)	.13 [.08, .19]	.12	.09	.22
Policy preference _a	8 (5; 1,855)	.10 [.02, .19]	.05	.09	.10
Microbehavior _b	87 (18; 3,162)	.07 [−.06, .19]	.22	.10	.25
Response time ^a	6 (5; 300)	.19 [.02, .37]	.27	.31	.28
Brain activity _{a,b}	20 (8; 327)	.43 [.12, .73]	.67 ^b	.30	.42
Ethnic minority vs. majority groups: Overall	92 (24; 7,571)	.12 [.06, .19]	.12	.12	.18
Interpersonal behavior _a	1 (1; 105)	.19 [°]	—	.19	^c
Person perception	63 (16; 3,807)	.11 [−.01, .23]	.15	.11	.19
Policy preference	13 (4; 2,822)	.16 [.08, .25]	.00	.17	.07
Microbehavior	9 (3; 717)	.11 [−.09, .31]	.14	.11	.19
Response time	—	—	—	—	—
Brain activity _a	6 (1; 120)	.11 [°]	—	.11	.27

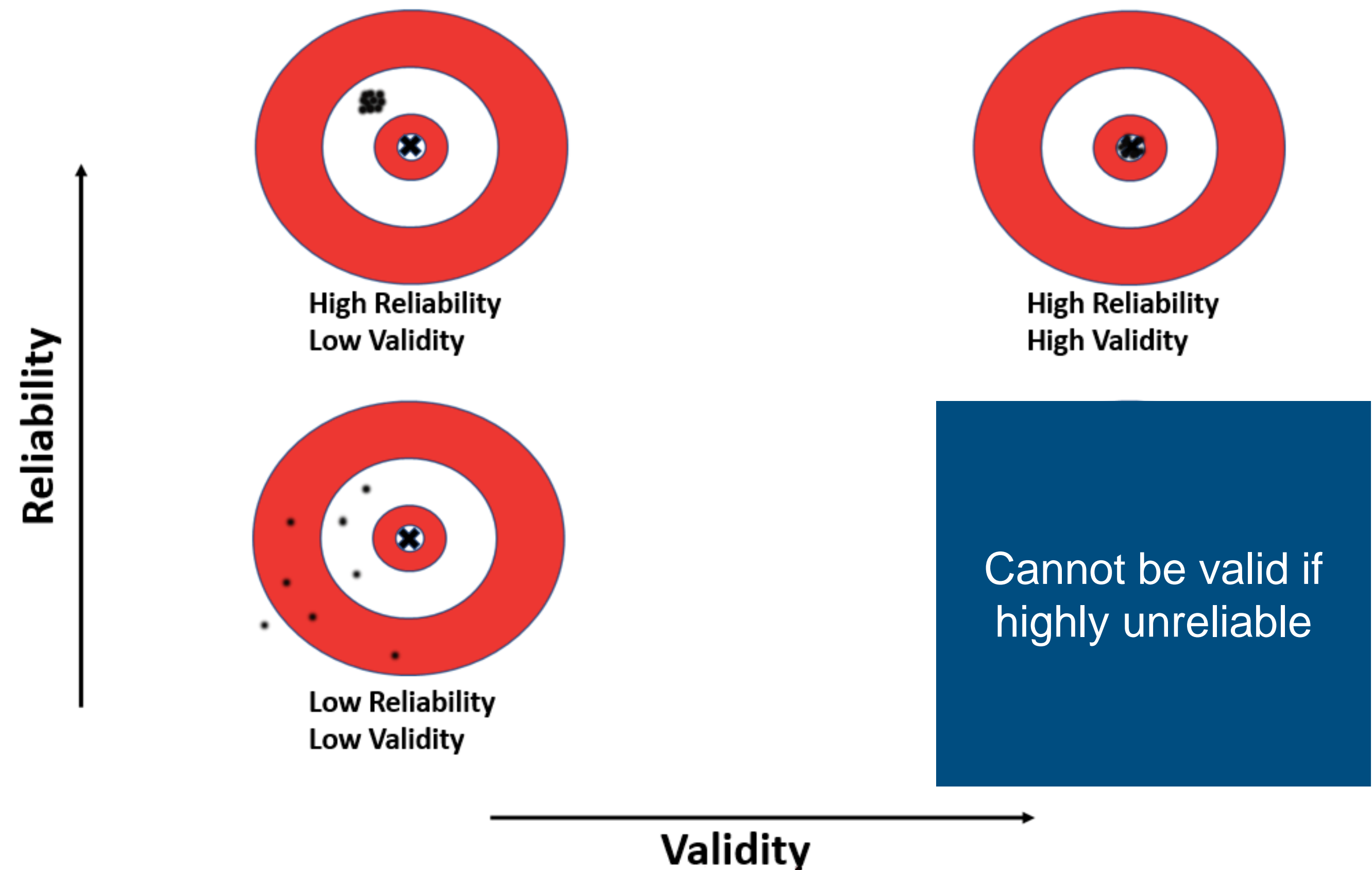
Noteworthy findings

- Small effects (.15 at best)
- Big heterogeneity (taus bigger than ESs)

What about test- retest reliability?

Good measurement properties

- No.
- Very poor test-retest:
few studies, $r = .30$ (Nosek et al., 2005)



Credit: Dr. Ian Hussey, The best theory is a flawed one, 14/02/2022 to RIOT Science Club

- 1) Overall, the validity of the IAT is low, and varies as a function of the domain
- 2) The domains in which the IAT is valid it largely overlaps with explicit measures
- 3) The IAT does not seem to be a reliable measure of individual differences
- 4) The IAT might still be of some use for group studies
- 5) We should not dismiss explicit measures
- 6) More psychometric training is needed...

Thanks!

