Implicit Association Tests

Stimuli Validation from Participant Responses

Sally A.M. Hogenboom
Leendert van Maanen (UU)
Katrin Schulz (UvA)

November 17, 2022

Table of contents I

Table of contents II

Doelstellingen

Doelstellingen

1. Input voor de discussie van *preregistered report* [1]
2. A *note of caution* voor IAT onderzoek(ers)

[1] Embargoed Registered Report, British Journal of Social Psychology, https://osf.io/hqe2r

Introductie

Er was eens...

- De behoefte om *attitudes* te meten.

Er was eens...

- De behoefte om *attitudes* te meten.
- *Implicit Association Tests* (IAT)

Er was eens...

- De behoefte om *attitudes* te meten.
- *Implicit Association Tests* (IAT)
  - 'The golden standard'

Er was eens...

- De behoefte om *attitudes* te meten.
- *Implicit Association Tests* (IAT)
  - 'The golden standard'
  - 16 nieuwe artikelen per maand!

Er was eens...

- De behoefte om *attitudes* te meten.
- *Implicit Association Tests* (IAT)
  - 'The golden standard'
  - 16 nieuwe artikelen per maand!
  - Vermindering van de *social desirability bias*

Hoe maak je een *goede* IAT over een *nieuw* onderwerp?

Handleidingen

(1) *"The Implicit Association Test at age 20: What is known and what is not known about implicit bias"* - Greenwald et al., 2020 [2]

(2) *Best research practices for using the Implicit Association Test* - Greenwald et al., 2021 [3]

[2] https://psyarxiv.com/bf97c/

[3] https://link.springer.com/article/10.3758/s13428-021-01624-3

Tipje van de Sluier

Exemplars that just one of a small group of pilot subjects finds difficult to classify [RT $<$ 800 ms & Error $<$ 10%] are safely discarded without further consideration

- *Greenwald et al., (2021)*[4]

[4] https://link.springer.com/article/10.3758/s13428-021-01624-3

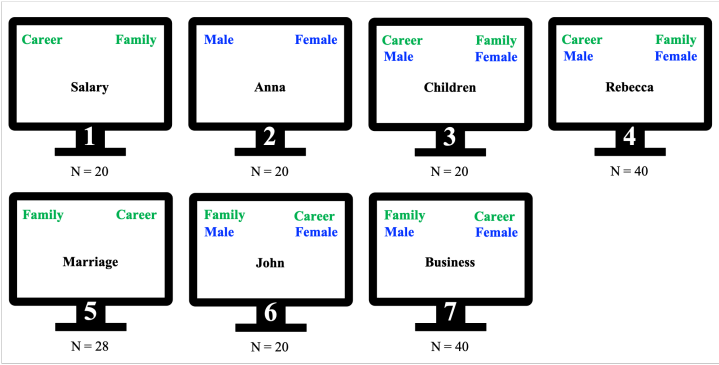
In welke mate voldoen de huidige IAT stimuli aan deze criteria?

1. Greenwald et al., (2021) geven **geen** onderbouwing voor deze criteria
2. De RT verdelingen van de Gender-Career IAT lieten in eerdere analyses grote verschillen zien.

Implicit Association Test

Design

2 x 2 Categoriën - met N exemplars per Categorie



Bias

- $RT_{incongruent}$ vs. $RT_{congruent}$
- D-score (i.e., verschilscore)
- Inferentie op basis van Categorieën - maar classificatie van Exemplars.

Categorieën/Exemplars moeten dus met aandacht gekozen worden - maar waar gaat het 'fout'?

Het probleem

Door stimulus- en particpant karakteristieken kunnen *cross-category associations* ontstaan: onverwachtte en/of meervoudige associaties

Voorbeeld 1: Stimulus Karakteristieken

1. Gender-Career IAT (Men/Women/Career/Family)
2. Gender-Criminality IAT (Men/Women/Criminal/Innocent)

Category: *Male*

Exemplar: *Jack*

Voorbeeld 1: Stimulus Karakteristieken

1. Gender-Career IAT (Men/Women/Career/Family)
2. Gender-Criminality IAT (Men/Women/Criminal/Innocent)

Category: *Male*

Exemplar: *Jack*

Cross-Category Associations:

- Gender-Career IAT: n.v.t.
- Gender-Criminality IAT: Jack the Ripper == Man & Crimineel

Voorbeeld 2: Participant Karakteristieken

Garimella et al. (2017)[5]

1. Gender: Male / Female
2. Country: U.S.A / India

Stimulus: *Bath*

[5] Garimella, A., Banea, C., & Mihalcea, R. (2017). Demographic-aware word associations. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2285–2295. https://doi.org/10.18653/v1/D17-1242

Voorbeeld 2: Participant Karakteristieken

Garimella et al. (2017)[5]

1. Gender: Male / Female
2. Country: U.S.A / India

Stimulus: *Bath*

Cross-Category Associations:

- Male: U.S.A = *Water* // India = *Water*
- Female: U.S.A = *Bubbles* // India = *Soap*

[5] Garimella, A., Banea, C., & Mihalcea, R. (2017). Demographic-aware word associations. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2285–2295. https://doi.org/10.18653/v1/D17-1242

Problematische *cross-category associations* kunnen dus ontstaan wanneer je de IAT-stimuli of proefpersoon populatie aanpast. Elke aanpassing zou dus moeten vragen om (her)nieuw(d)e stimulus validatie.

... maar hoe?

De criteria van Greenwald et al. (2021) zijn momenteel de enige expliciet geformuleerde criteria op *stimulus* niveau.

Greenwald et al., (section A8, 2021)

A8. Exemplar stimuli for target and attribute categories are best selected by pilot testing using the category classification tasks planned for the IAT [...] ease of classification [..]. Subjects for ***pilot testing*** *should come from the* ***intended research subject population****. [...] The useful data will come from* ***Blocks 1 and 2*** *of the standard procedure (see Appendix A). Pilot subjects should be able to categorize* ***all stimuli*** *in these two blocks* ***rapidly*** *(average latency in the range of 600–* ***800*** *ms for most* ***young adult*** *subjects) and with* ***low error*** *rates (* ***less than 10%*** *). Exemplars that just one of a small group of pilot subjects finds difficult to classify are safely discarded without further consideration. [...].* [6]

[6] https://link.springer.com/article/10.3758/s13428-021-01624-3

Onderzoeksdoelen

- Evaluatie van validatie criteria dat exemplars snel (RT $<$ 800 ms) en accuraat ($<$ 10% error) gecategoriseerd worden.

Geplande Analyses

1. Toepassing op 15 individuele IATs
2. Exploreren van context gevoeligheid; validiteit van stimuli die in meerdere IATs gebruikt worden.
3. Exploreren effect van stimulus type; verschilt validiteit voor images, nouns, adjectives, en names?

Methode

Data

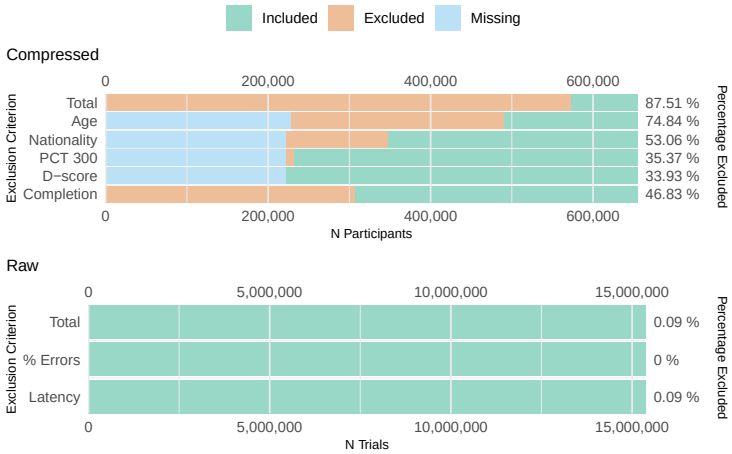
- 15 IATs beschikbaar via Project Implicit [7]
- Age; Arab; Asian; Disability; Gender-Career; Gender-Science; Native-American; President; Race; Religion; Sexuality; Skin; Transgender; Weapons; Weight

[7] https://osf.io/y9hiq/

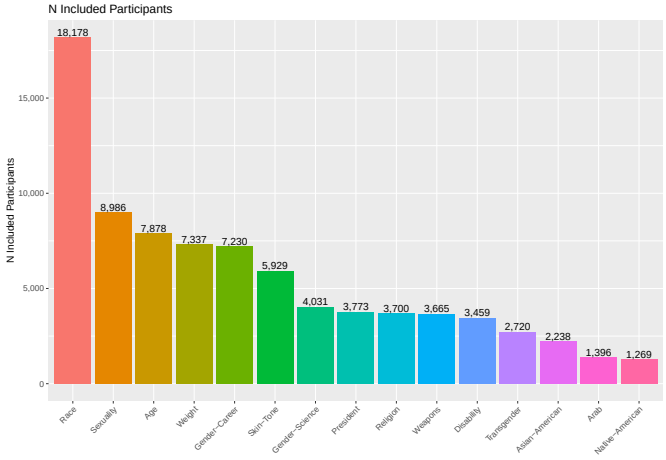
Data

- *Compressed*: demographics, d-score, explicit attitudes etc.
- *Raw*: trial-by-trial response data
- Koppeling via `session_id` als zijnde een 'proefpersoon'

In-/Exclusie


Included
Excluded
Missing
Compressed
0
200,000
400,000
600,000
Exclusion Criterion
Total
Age
Nationality
PCT 300
D−score
Completion
87.51 %
74.84 %
53.06 %
35.37 %
33.93 %
46.83 %
Percentage Excluded
N Participants
Raw
0
5,000,000
10,000,000
15,000,000
Exclusion Criterion
Total
% Errors
Latency
0.09 %
0 %
0.09 %
Percentage Excluded
N Trials

N Geïncludeerde Proefpersonen



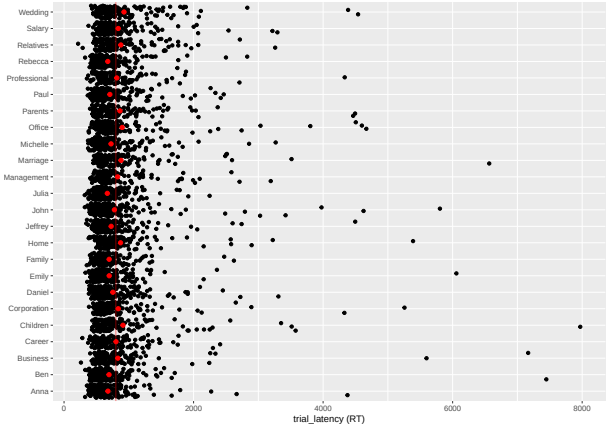
Analyses & Resultaten

Bootstraps

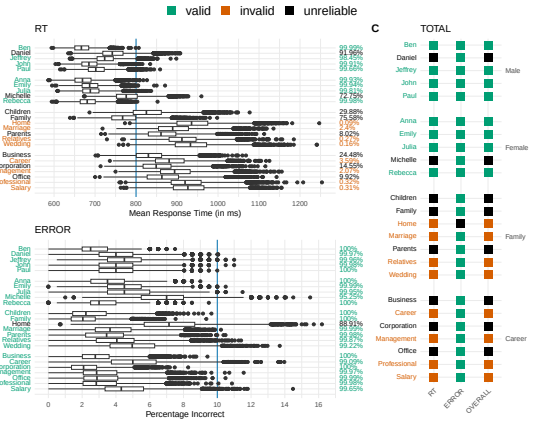
10,000 'experimenten' met 100 proefpersonen in elke sample.

- *10,000* samples = arbitrair groot getal
- *100* proefpersonen = een 'normale' sample size voor IAT experimenten (op basis van samplesizes in meta-analyses).

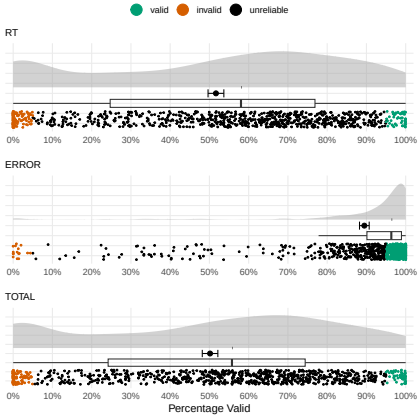
1 Sample van 100 Proefpersonen



10,000 Samples van 100 Proefpersonen

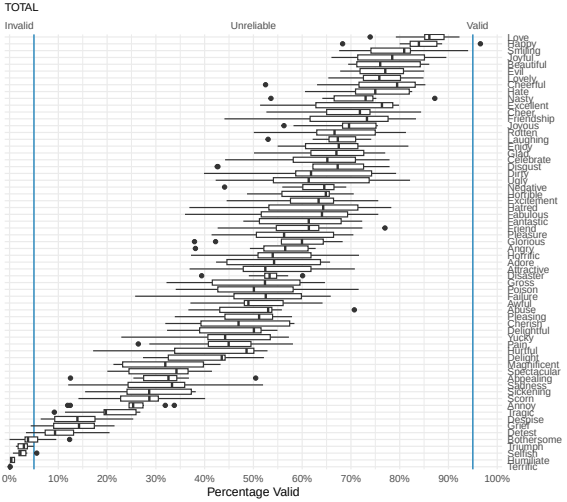


Analyse 1: Alle 395 Stimuli in 15 IATs



- 64 Exemplars worden in 9/10 IATs gebruikt.
- Positive/Negative/Bad/Good
- Bijvoorbeeld: “Yucky”, “Cheerful”, “Smiling”, “Grief”

Hoge spreiding = context dependent: de validiteit van een stimulus is afhankelijk van de IAT / populatie waar de stimulus in voorkwam.


TOTAL
Invalid
Unreliable
Valid
Love
Happy
Smiling
Joyful
Beautiful
Evil
Lovely
Cheerful
Hate
Nasty
Excellent
Cheer
Friendship
Joyous
Rotten
Laughing
Enjoy
Glad
Celebrate
Disgust
Dirty
Agony
Negative
Horrible
Excitement
Hatred
Fabulous
Fantastic
Friend
Pleasure
Glorious
Angry
Horrific
Adore
Attractive
Disaster
Gross
Poison
Failure
Awful
Abuse
Pleasing
Cherish
Delightful
Yucky
Pain
Hurtful
Delight
Magnificent
Spectacular
Appealing
Sickness
Sickening
Scorn
Annoy
Tragic
Despise
Guilt
Detest
Bothersome
Triumph
Selfish
Humiliate
Terrific
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Percentage Valid
ERROR
Invalid
Unreliable
Valid
Smiling

In welke mate is de validiteit verschillend voor stimulus types?

stimulus_type	N	Examples	MeM
Image	236	native5.jpg, us2.jpg, recent15.jpg, fatwoman6.jpg, hwallet.jpg...	Included
Adjective	69	Evil, Gay, Cheerful, Friendship, Rotten...	Included
Noun	57	Bible, Church, Girl, Children, Management...	Included
Name	29	Emily, Sharif, John, Takuya, Yousef...	Included
Multi-Word	4	Gay Men, Gay Women, Gay People, Straight People	Excluded

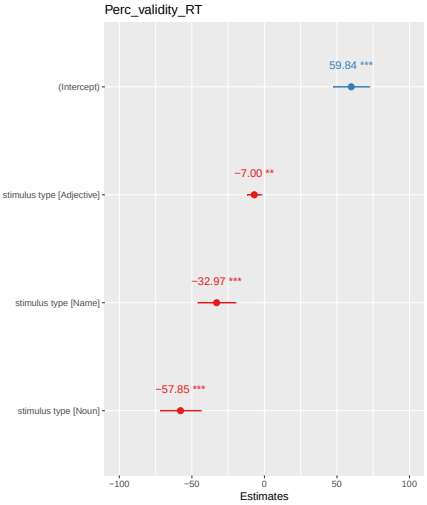
$$percantage\_valid \sim 1 + stimulus\_type + (1 + stimulus\_type \mid IAT)$$

- $(... \mid IAT)$: corrigeren voor de context van IATs
- $(1 + stimulus\_type$: corrigeren voor stimulus type binnen IATs

"Mixed-effects model with the `lme4` and `lmerTest` packages with a BOBYQA optimizer (Powell, 2009) and a maximum of 200,000 iterations (Miller, 2018).

Model convergeert niet vanwege issues met (near) singularity »> simpeler model:

$$percentage\_valid \sim 1 + stimulus\_type + (1 \mid IAT)$$


Perc_validity_RT
59.84 ***
(Intercept)
−7.00 **
stimulus type [Adjective]
−32.97 ***
stimulus type [Name]
−57.85 ***
stimulus type [Noun]
−100
−50
0
50
100
Estimates

Discussie

Algemeen: Samenvatting

- Van de 395 voorkomend in 15 IATs ($N_{total} = 935$ kunnen we slechts 201 keer betrouwbaar stimulus (in)validity concluderen (21.4973262%).
- Van de 935 validiteits checks zijn slechts 53 betrouwbaar ($>$ 95% vd 10,000 samples) valide (RT $<$ 800ms **EN** $<$ 10% ERROR).
- Van de 935 validiteits checks zijn 148 betrouwbaar ($>$ 95% vd 10,000 samples) invalide.
- Uit de algemene verdelingen blijkt telkens dat het `RT` criterium (RT $<$ 800 ms) veel vaker tot een beoordeling van invaliditeit zorgt dan het `ERROR` criterium ($<$ 10% errors).

Algemeen: Implicaties

- Het RT/ERROR criterium is te strikt/tolerant
  - Criteria herzien
  - Wat is een 'impliciete reactietijd'?
- De IATs hebben allen last van stimulus validiteit problemen
  - Hoe heeft 1 of meerdere invalide stimuli effect op de gevonden bias scores?
  - Kun je bias scores herberekenen op basis van post-hoc validatie?

Context Dependency: Samenvatting

- Het grootste deel van de 64 herbruikte exemplars verschilt in validiteit gegeven de IAT waar deze in voorkwam (grote spreiding).
- Sommige woorden (e.g., `Terrific`) zijn in alle IATs en voor beide criteria problematisch. Geen woorden zijn in alle IATs valide.
- Wederom verschil in effect `RT` vs. `ERROR` criterium.

Context Dependency: Implicaties

- Spreiding in validiteit over IATs heen kan meerdere oorzaken hebben:
  - Invloed van overige stimuli/categorieën (*cross-category associations*)
  - Invloed van de populatie die een IAT afmaakt (e.g., Psychology 101 students in Race-IAT vs. Sexuality-IAT).
- Ongeacht de oorzaak: de validiteit van stimuli verschilt per IAT/populatie - dus stimulus (her)validatie is genoodzaakt.

Stimulus Types: Samenvatting/Implicaties

- Under construction; er lijkt een verschil te zitten in de validiteit van verschillende stimulus types.
- Implicaties;
  - Validatie criteria aanpassen op stimuli types
  - Beperken tot 1 stimulus type per IAT? (nog te analyseren)
  - Een bepaalde stimulus_type in zijn geheel links laten liggen?
  - Culturele afhankelijkheid?

Vragen & Opmerkingen

Vragen & Opmerkingen

Heel benieuwd naar jullie reacties, gedachtes, input voor de discussie!