Investigating the relationship between TOEFL scores and international students' academic success: A meta-analysis

SLRF 2020, Tennessee

Ryo Maie & Wenyue Ma
Second Language Studies, Michigan State University

Background

1. A large body of research examining the relationship between international students' TOEFL scores and their academic achievement (e.g., Ginther & Yan, 2018; Light, Xu, & Mossop, 1987).
2. Grade point averages (GPA), although having received some criticism, has been widely used as a proxy indicator for academic success.
3. Studies tend to produce mixed fixed findings:
   a. Strong positive relationship (e.g., r = .654, Johnson & Tweedie, 2017)
   b. Small or even negligible (e.g., r = .07, Kwai, 2010)

Background

Possible explanations for the inconsistencies in the previous studies:

1. **The restricted range** in TOEFL scores available in the dataset (Cho & Bridgeman, 2012; Bridgeman, Cho, & DiPietro, 2016)
2. **Moderator variables**, including students' academic status, majors, different types of GPA, different versions of TOEFL, may complicate and add subtlety to the relationship.

Purpose of the study

1. The restricted range in TOEFL scores available in the dataset (Cho & Bridgeman, 2012; Bridgeman, Cho, & DiPietro, 2016)

   *We conducted **a quantitative synthesis study** to delineate the strength of the empirical relationship between TOEFL scores and academic success*

2. Moderator variables, including students' academic status, majors, different types of GPA, different versions of TOEFL, may complicate and add subtlety to the relationship.

   *We investigated **what kinds of factors (exploratorily)**, if any, **moderate the relationship** between the two variables.*

Research Questions

Questions

RQ1: What is the predictive relationship between TOEFL and academic performance operationalized by grade point average (GPA)?

RQ2: What are the moderating variables that mediate the relationship?

Domains

1. Graduate or undergraduate enrollment in U.S. or Canadian universities
2. All years of the degree process

Procedure

1. Defining research questions and domains
2. Literature search
3. Developing the coding book
4. Coding
5. Analysis

(Plonsky & Oswald, 2015)

Literature search

Inclusion/Exclusion criteria

- Fit into the research domain
- Report correlation or regression coefficients
- Written in English

Database

- Linguistics and Language Behavior Abstract
- PsycINFO
- Proquest database for dissertations and theses
- Web of Science
- Google Scholar

45 primary studies with **111** effect sizes

17502 independent participants

Developing the coding book

1. We brainstormed a list of potential study characteristics based on our knowledge of literature.
2. Each of us randomly coded 10 studies to validate.
3. We drafted the coding book and asked opinions from an expert in the domain.
4. We finalized the coding book.

The coding book was revised as we coded the primary studies.

- Dynamic and cyclical nature of the coding process

Study characteristics

Moderators

Publication status (un/published), *Academic status* (under/graduate),

Institution status (public, private), *TOEFL type* (iBT, PBT, CBT),

GPA type (cumulative, first year, first semester)

GPA mean, *TOEFL mean*

Coding

- Each of us independently coded all 45 studies and then compared and merged the data.
- Any discrepancies in coding were resolved through discussion.
- Intercoder reliability
  - Average agreement rate: **93.17%**
  - Average Cohen's Kappa: **0.8474**
  - All characteristics were of low inference

Analysis

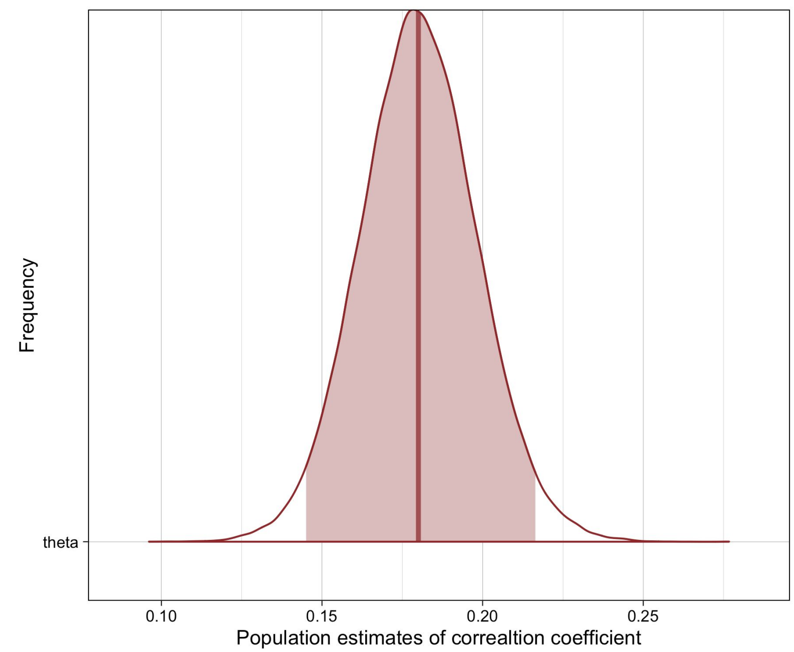
Synthesizing effect sizes for the overall effect

- Extracted effect sizes from **independent** participants
- **Bayesian multilevel models** (3 levels), which took into account nested data structure <u>at the study level</u> (i.e., multiple effects coming from the same study)
  - We used **Stan**, a probabilistic programming language, to estimate the posterior distribution of the population estimate through Markov chain Monte Carlo simulation.
  - We used Abunawas (2015) results as prior information

Moderator analyses

- Bayesian random effects models fit to **each subsample**

RQ1: Overall effect



The overall effect was **ρ = .178 [.143, .212]**

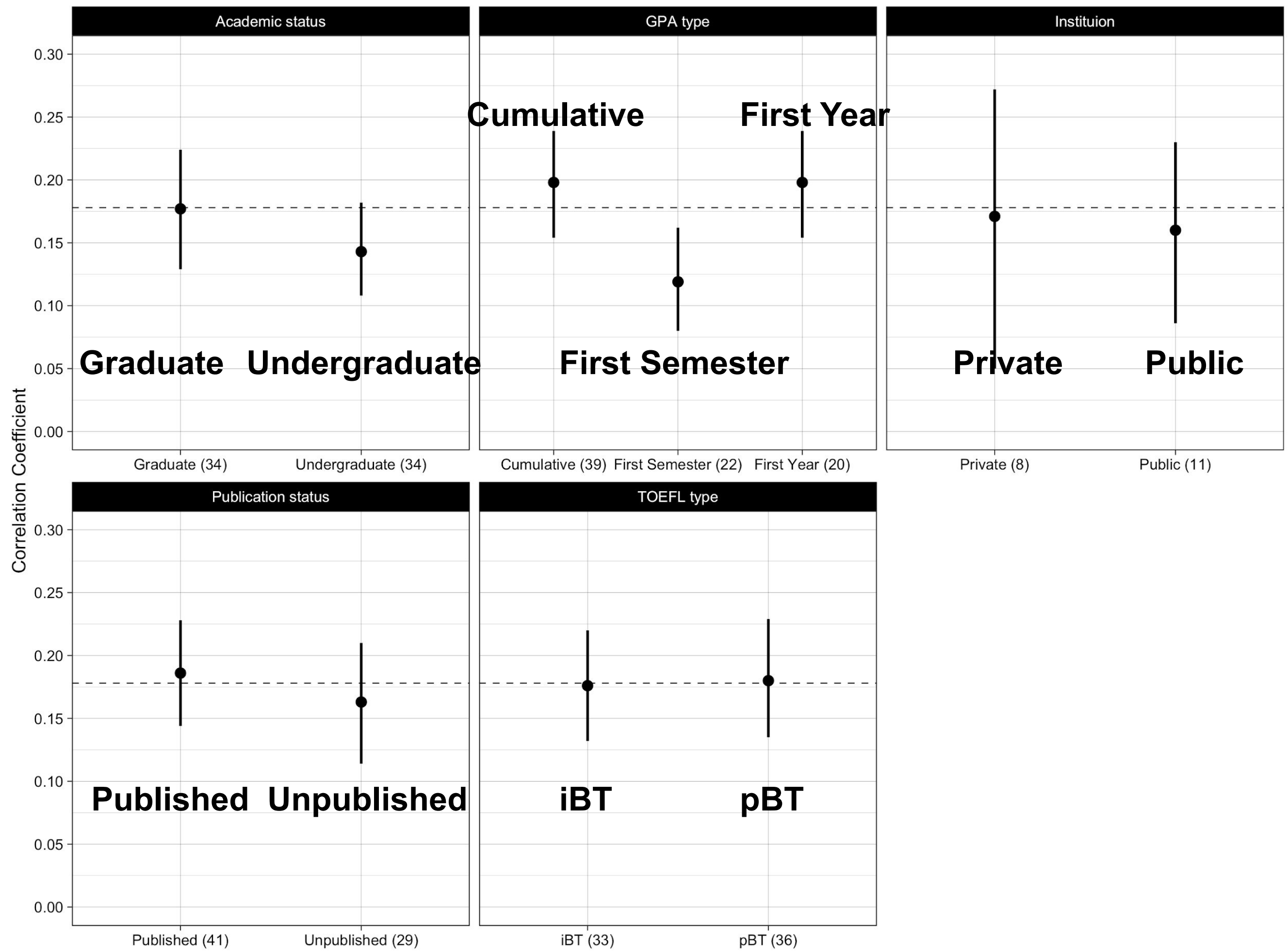
- 3.1% of shared variance

The study heterogeneity:

- Sampling error: v = .059 [.004, .115]
- Within-study: τ_2 = .058 [.003, .114]
- Between-study: τ_3 = .048 [.002, .110]

I^2_2 = .367 (36.7% of error variance by within)
I^2_3 = .251 (25.1% of error variance by between)

RQ2: Moderator analyses



Discussions

1. **The overall effect was ρ = .178 [.143, .212] (small!!!)**

Current study VS Abunawas (2015): ρ = .21 [.16, .26]

- Different samples
  - 45 studies with 111 effect sizes **VS** 40 studies (11 in an international setting) with 47 effect sizes
  - ESL (k = 45) context **VS** EFL (k = 11) + ESL (k = 29) context
- The importance of accounting for within-study dependencies

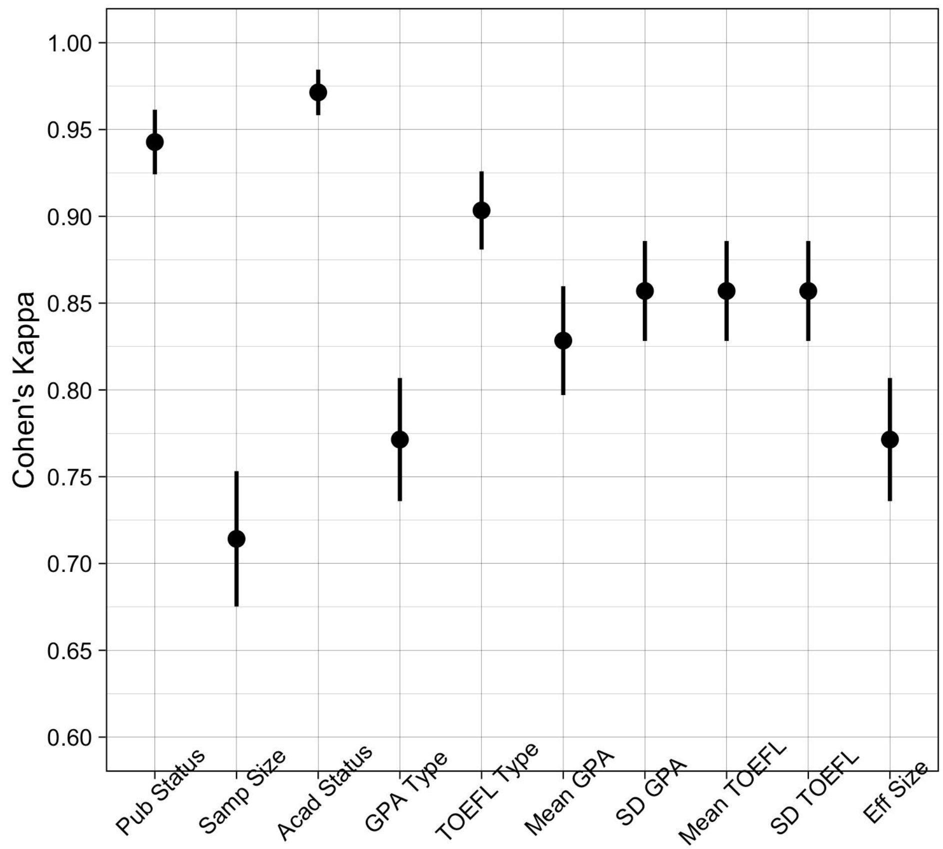
Discussions

2. Moderator analysis

- Student status: Graduate **slightly higher** than undergraduate
- GPA types: First semester GPAs yielded the **smallest** effect size
- iBT vs pBT: **similar-size** effects
  - The addition of **the speaking section**
  - Which components of language proficiency contributes to the correlation between academic achievements and TOEFL?

BUT, the overall effect was very small!

Cohen's Kappa



Bayesian multilevel models (overall)

Level 1: $y_{ij} = \lambda_{ij} + e_{ij}$

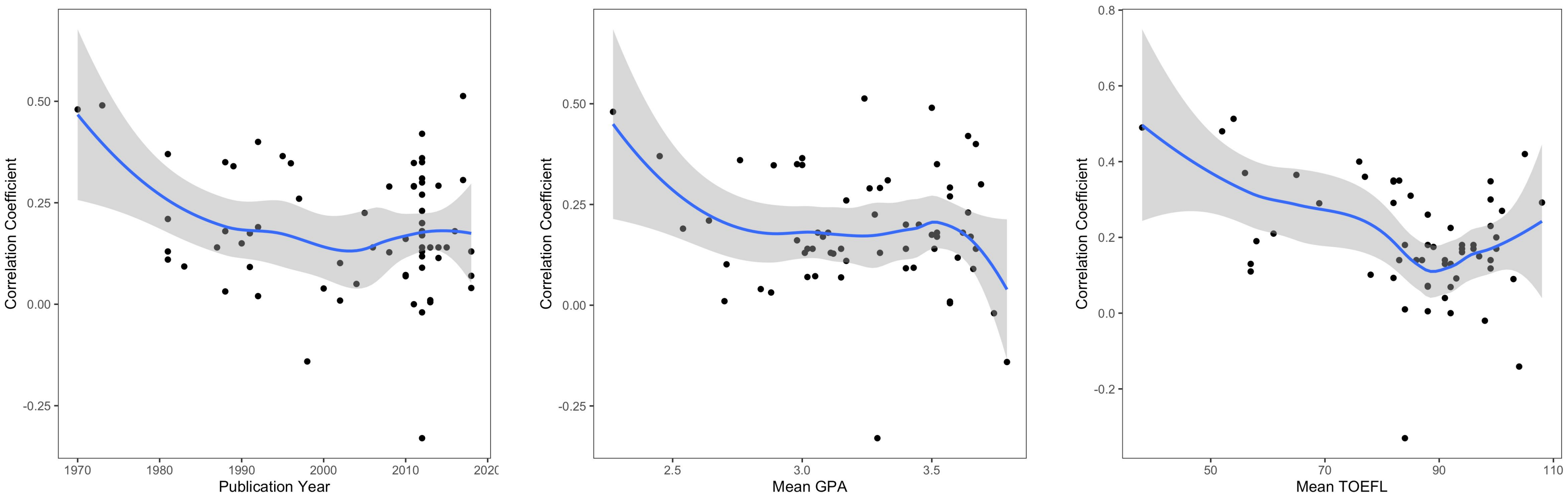
Level 2: $\lambda_{ij} = \kappa_j + u_{(2)ij}$

Level 3: $\kappa_j = \theta_0 + u_{(3)j}$

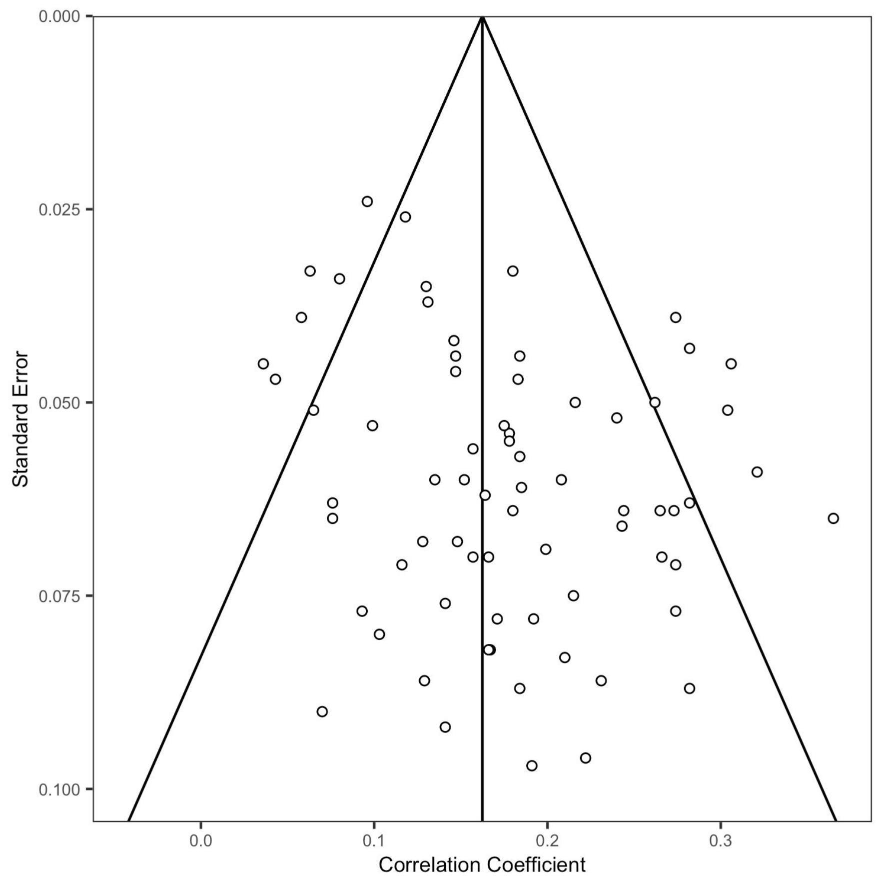
$$y_{ij} = \theta_0 + u_{(2)ij} + u_{(3)ij}$$
$$u_{(3)j} \sim \mathrm{N}(0, \tau_{(3)j})$$
$$u_{(2)ij} \sim \mathrm{N}(0, \tau_{(2)ij})$$
$$e_{ij} \sim \mathrm{N}(0, \sigma_{ij})$$

$$\theta_0 \sim \mathrm{N}(0.21, 0.1)$$
$$\tau_{(3)j} \sim \mathrm{Cauchy}(0, 1)$$
$$\tau_{(2)ij} \sim \mathrm{Cauchy}(0, 1)$$
$$\sigma_{ij} \sim \mathrm{Cauchy}(0, 1)$$

Exploratory analysis

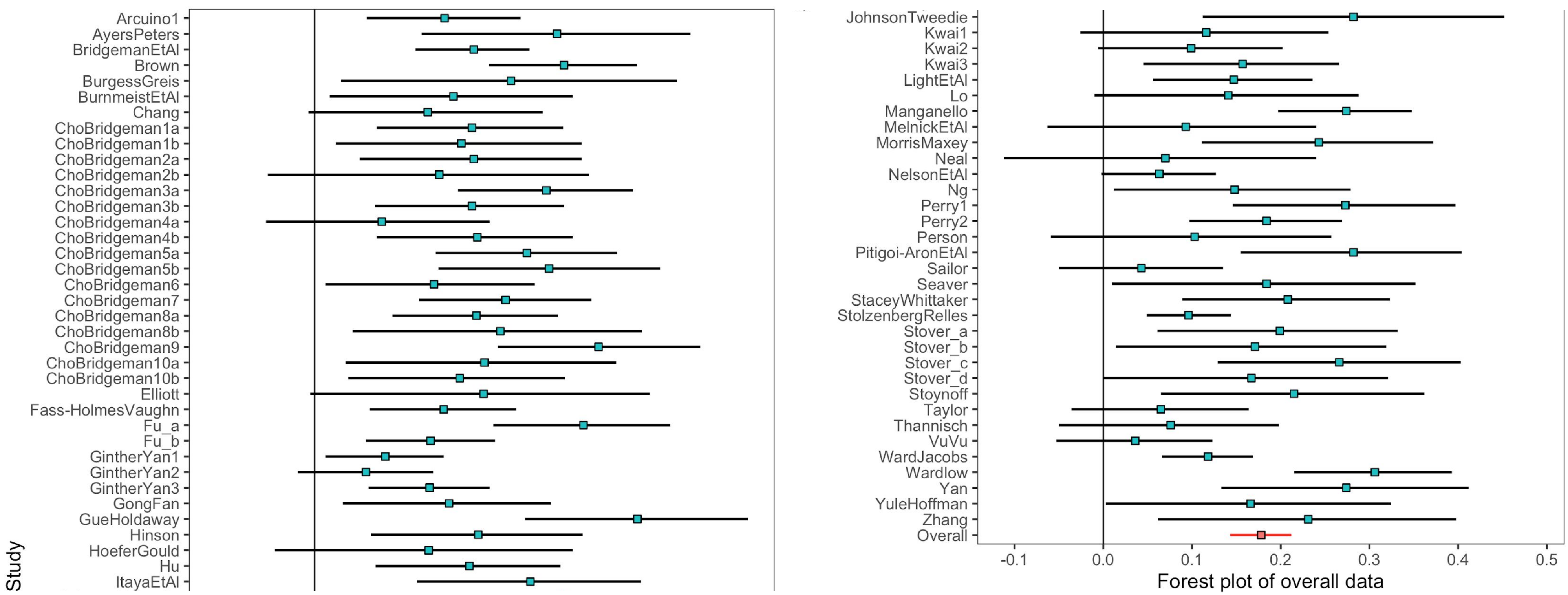


Funnel plot



0.000
0.025
0.050
0.075
0.100
Standard Error
0.0
0.1
0.2
0.3
Correlation Coefficient

Forest plot


Arcuino1
AyersPeters
BridgemanEtAl
Brown
BurgessGreis
BurnmeistEtAl
Chang
ChoBridgeman1a
ChoBridgeman1b
ChoBridgeman2a
ChoBridgeman2b
ChoBridgeman3a
ChoBridgeman3b
ChoBridgeman4a
ChoBridgeman4b
ChoBridgeman5a
ChoBridgeman5b
ChoBridgeman6
ChoBridgeman7
ChoBridgeman8a
ChoBridgeman8b
ChoBridgeman9
ChoBridgeman10a
ChoBridgeman10b
Elliott
Fass-HolmesVaughn
Fu_a
Fu_b
GintherYan1
GintherYan2
GintherYan3
GongFan
GueHoldaway
Hinson
HoeferGould
Hu
ItayaEtAl
Study
JohnsonTweedie
Kwai1
Kwai2
Kwai3
LightEtAl
Lo
Manganello
MelnickEtAl
MorrisMaxey
Neal
NelsonEtAl
Ng
Perry1
Perry2
Person
Pitigoi-AronEtAl
Sailor
Seaver
StaceyWhittaker
StolzenbergRelles
Stover_a
Stover_b
Stover_c
Stover_d
Stoynoff
Taylor
Thannisch
VuVu
WardJacobs
Wardlow
Yan
YuleHoffman
Zhang
Overall
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
Forest plot of overall data