Universität
Zürich UZH

# The Success of Linear Bootstrapping Models: Decision Domain-, Expertise-, and Criterion-Specific Meta-analysis

Esther Kaufmann (University of Zurich, Switzerland)

Werner W. Wittmann (University of Mannheim, Germany)

Society of Judgment and Decision Making, 2016 (Nov. 19)

Kaufmann, E., & Wittmann, W. W. (2016) The Success of Linear Bootstrapping Models: Decision Domain-, Expertise-, and Criterion-Specific Meta-Analysis. *PLoS ONE 11*(6): e0157914. doi:10.1371/journal.pone.0157914

Kaufmann, E., Reips, U.-D., & Wittmann, W. W. (2013). A critical meta-analysis of Lens Model Studies in human judgment and decision-making. *PLoS ONE 8*(12): e83528. doi:10.1371/journal.pone.0083528

# Introduction

Across a variety of settings, human judges are often replaced or ‘bootstrapped’ by decision-making models (in our examples, equations) in order to increase the accuracy of important - and often ambiguous – decisions

-> to save lives in medical science

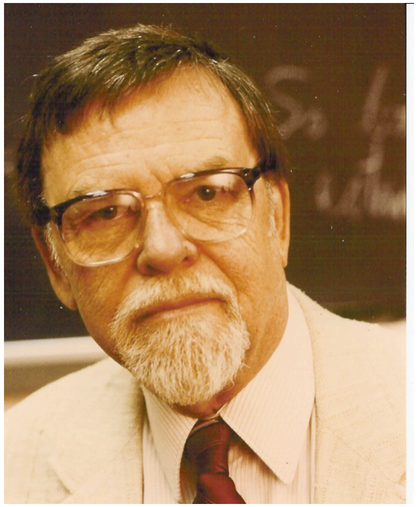
-> to improve students’ learning in education science

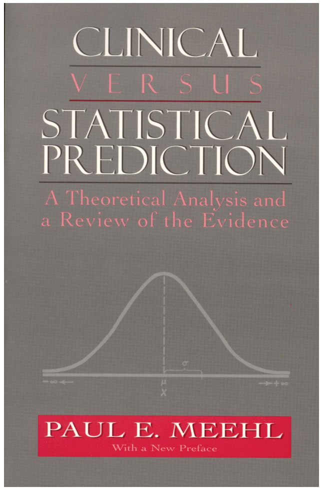
- Is it worthwhile to invest in developing such decision-making models, or is it just a waste of time?
- And how can we most precisely evaluate the success of bootstrapping models?

# Paul E. Meehl

Quantitative review of bootstrapping models (1954)

- Statistical vs. clinical predictions
- Frequency counting (box-score approach)



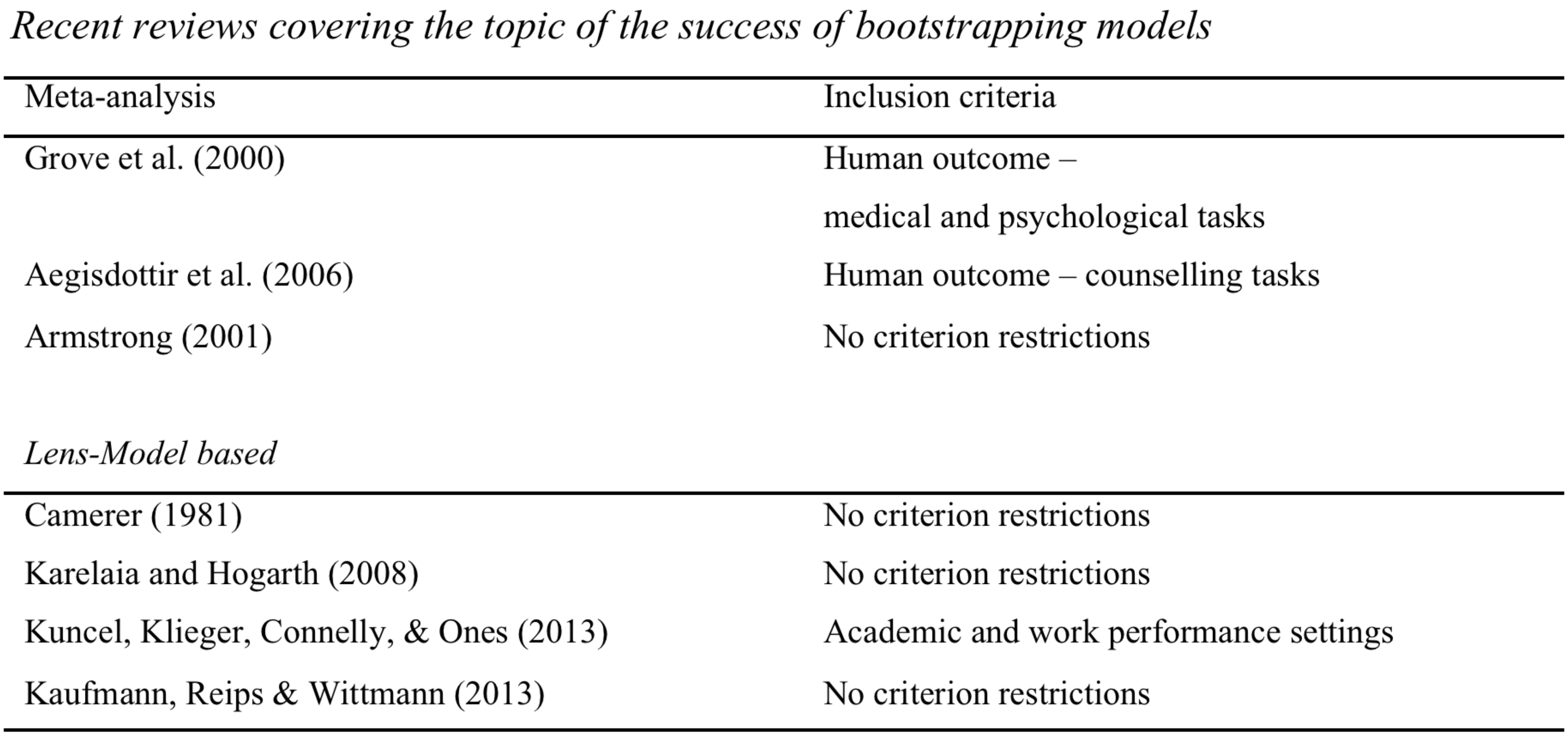


CLINICAL
VERSUS
STATISTICAL
PREDICTION

A Theoretical Analysis and
a Review of the Evidence

PAUL E. MEEHL
With a New Preface

# Recent reviews

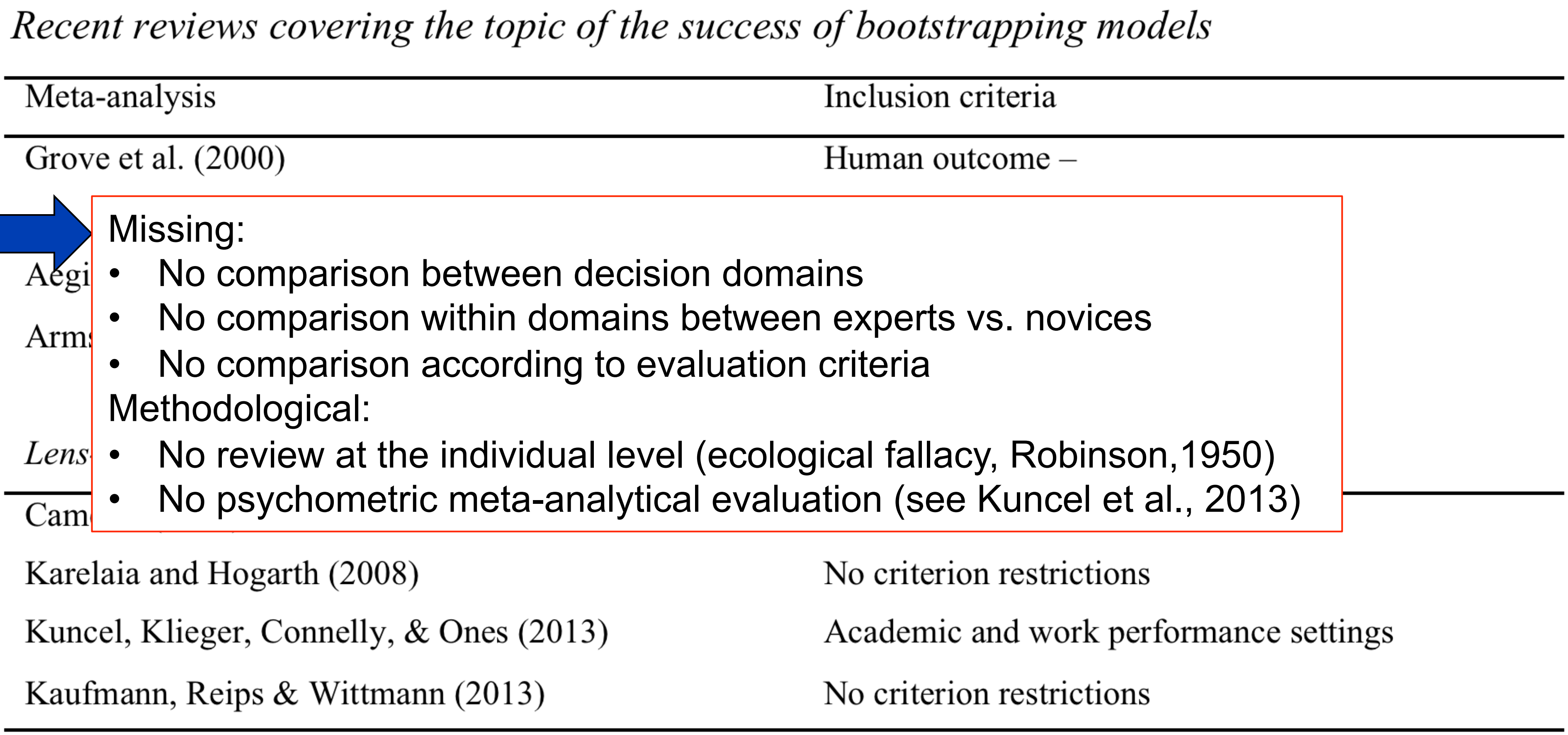


*Recent reviews covering the topic of the success of bootstrapping models*

| Meta-analysis | Inclusion criteria |
|---|---|
| Grove et al. (2000) | Human outcome –<br>medical and psychological tasks |
| Aegisdottir et al. (2006) | Human outcome – counselling tasks |
| Armstrong (2001) | No criterion restrictions |
| | |
| *Lens-Model based* | |
| Camerer (1981) | No criterion restrictions |
| Karelaia and Hogarth (2008) | No criterion restrictions |
| Kuncel, Klieger, Connelly, & Ones (2013) | Academic and work performance settings |
| Kaufmann, Reips & Wittmann (2013) | No criterion restrictions |

Universität Zürich^UZH

# Recent reviews



*Recent reviews covering the topic of the success of bootstrapping models*

| Meta-analysis | Inclusion criteria |
|---|---|
| Grove et al. (2000) | Human outcome – |
| Aegi | |
| Arms | |
| *Lens* | |
| Cam | |
| Karelaia and Hogarth (2008) | No criterion restrictions |
| Kuncel, Klieger, Connelly, & Ones (2013) | Academic and work performance settings |
| Kaufmann, Reips & Wittmann (2013) | No criterion restrictions |

Missing:
- No comparison between decision domains
- No comparison within domains between experts vs. novices
- No comparison according to evaluation criteria

Methodological:
- No review at the individual level (ecological fallacy, Robinson,1950)
- No psychometric meta-analytical evaluation (see Kuncel et al., 2013)

# Research questions

- Does the success of bootstrapping models vary across decision domains (e.g., medical versus business decisions)?
- Does the success of bootstrapping models vary within domains between expert and novice decision makers?
- Does the success of bootstrapping models vary according to the type of criterion for a „successful decision“ (objective, subjective, or based on a test score)?

# Success of bootstrapping models within the lens model approach

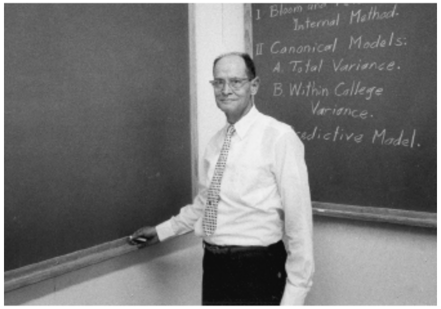
$$\Delta = GR_e - r_a$$

Judgment accuracy of human judge(s)

Model

Success of bootstrapping model

- Yes, if the value is positive
- No, if the value is negative



For more information on the Lens Model Equation, see Tucker (1964)

# Studies included in the meta-analysis (medical science)

Table 1

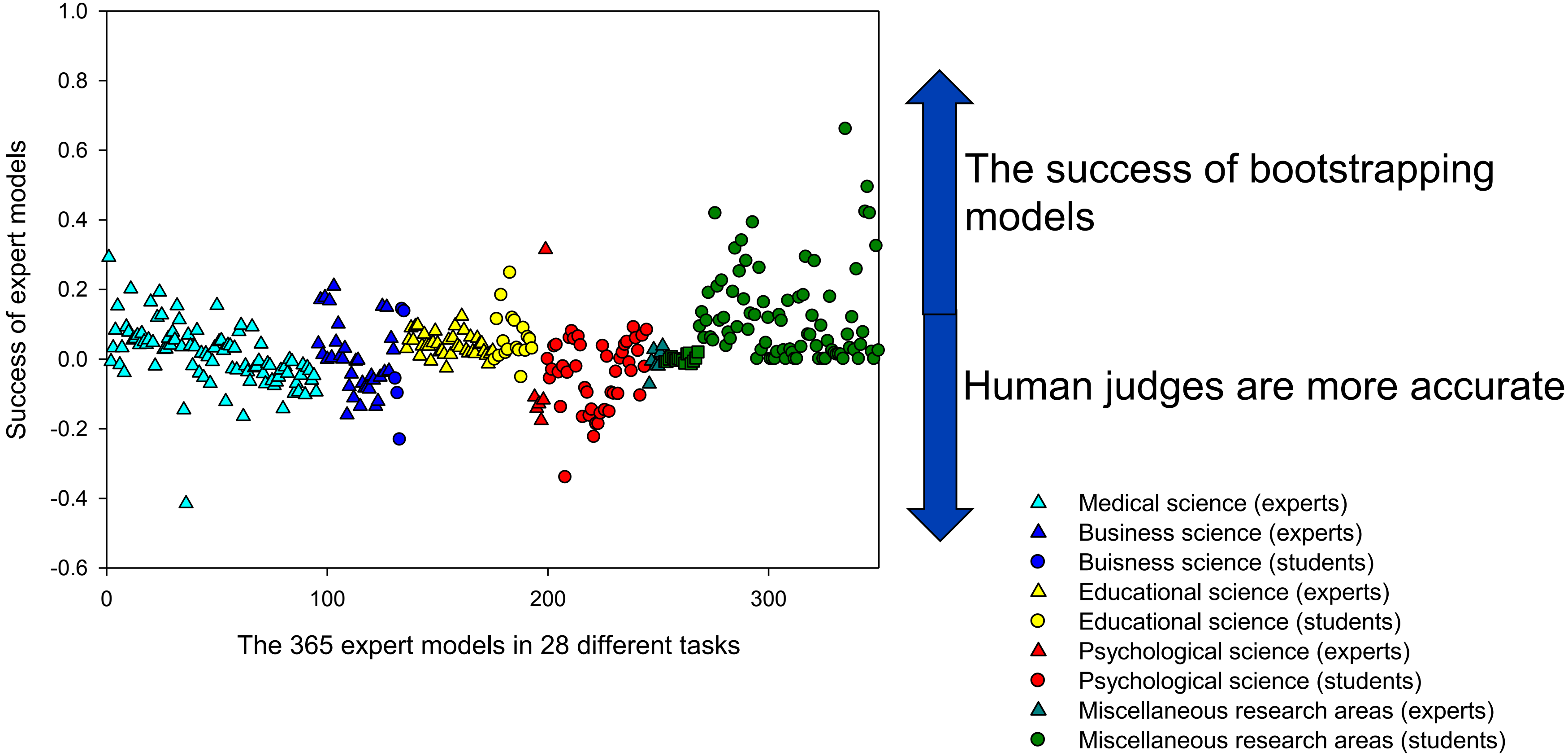
*Studies included in the meta-analyses by decision domain and decision-maker expertise*

| | Study | Judges | Number of judgments | Number of cues | Judgment task | Criterion | Results |
|---|---|---|---|---|---|---|---|
| *a)* | *Medical science, experts:* | | | | | | |
| 1) | Nystedt & Magnusson, 1975 | 4 clinical psychologists | 38 | 3 | Judge patients based on patient protocols :<br>*I:*intelligence<br>*II:* ability to establish contact<br>*III:* control of affect and impulses | Rating on three psychological tests (■) | *I:* $\Delta_1 = .11$<br>*II:* $\Delta_2 = .03$<br>*II:* $\Delta_3 =.12$<br>(*, +, s) |
| 2) | Levi, 1989 | 9 nuclear medicine physicians | 280<br>(60 replications) | 5 | Assess probability of significant coronary artery disease based on patient profiles | Coronary angiography | $\Delta_4 = .07$<br>(*, s) |
| 3) | LaDuca, Engel, & Chovan, 1988 | 13 physicians | 30 | 5 | Judge the degree of severity (congestive heart failure) based on patient profiles | A single physician's judgment (▲) | $\Delta_5 = .08$<br>(*, s) |
| 4) | Smith, Gilhooly, & Walker, 2003 | 40 general practitioners | 20 | 8 | Decision to prescribe an antidepressant based on patient profile | Guideline expert (▲) | $\Delta_6 = -.05$<br>(s) |
| 5a) | Einhorn, 1974<br>*Second study* | 3 pathologists | *III:* 193 | 9 | Evaluate the severity of Hodgkin's disease based on biopsy slides | Actual number of months of survival | *III:* $\Delta_7 = -.01$<br>(s) |
| 6a) | Grebstein, 1963 | 10 clinical experts (varying in amounts of clinical experience) | 30 profiles | 10 | Judge Wechsler-Bellevue IQ scores from Rorschach psychograms | IQ test scores (■) | $\Delta_8 = -.17$<br>$\Delta_9 = -.14$ |
| 5b) | Einhorn[1], 1974<br>*First study* | 29 clinicians | *I:* 77 MMPI profiles<br>*II:* 181 MMPI profiles | 11 | Judge the degree of neuroticism-psychoticism | Actual diagnosis (■) | *I:* $\Delta_{10} = .02$<br>*II:* $\Delta_{11} = -.05$<br>(*, +, s) |

# Database

- 35 studies (52 tasks)
- 1,110 bootstrapping models
- 532 experts versus 578 novices
- Five different decision domains (e.g., medical versus educational decisions)
- Individual-level data: 365 individual bootstrapping models across 28 tasks

# Individual level (to prevent any aggregation bias)

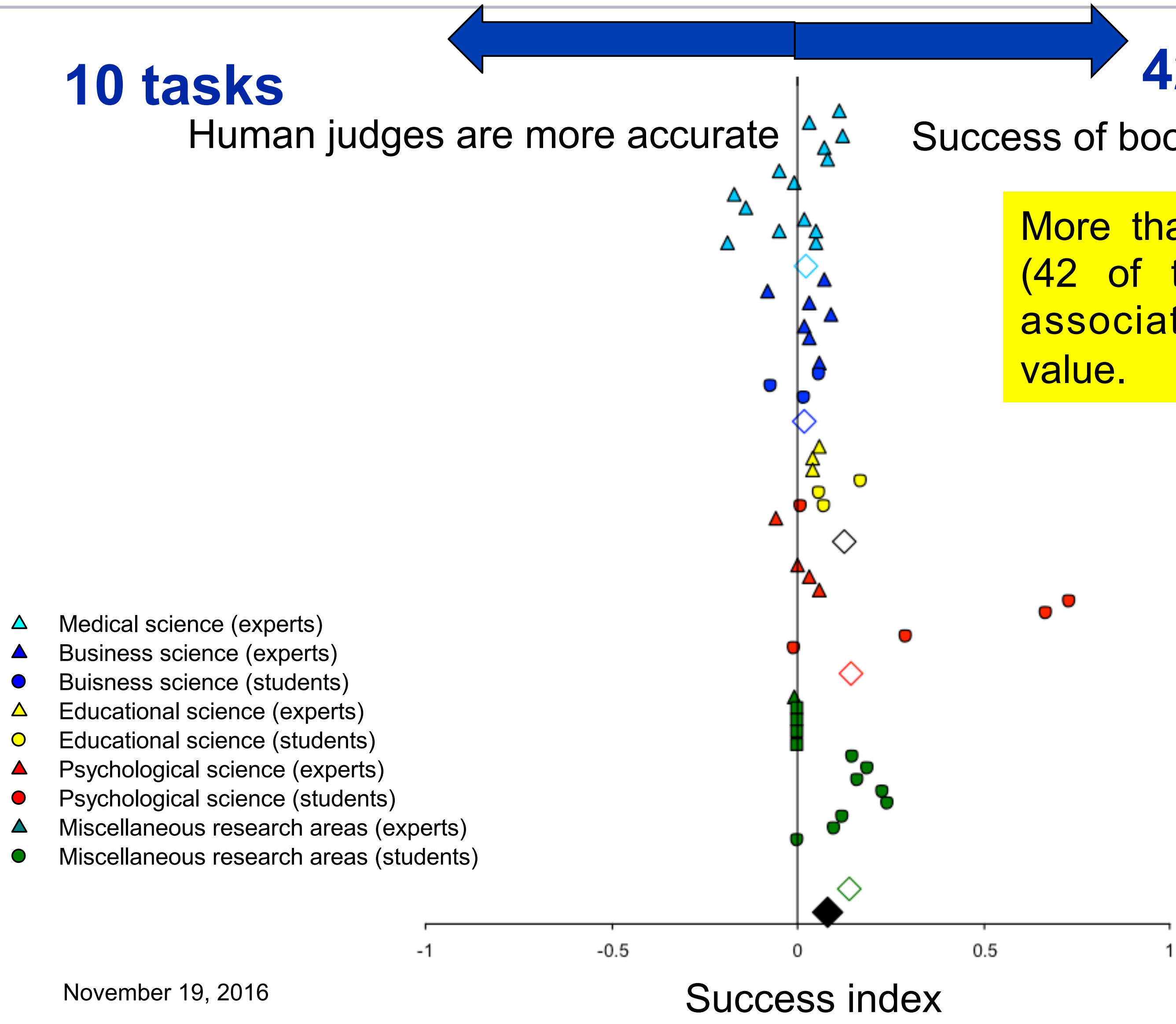


The success of bootstrapping models

Human judges are more accurate

Universität Zürich^UZH

# Forest plots of the sucess of bootstrapping models organized by decision domain and decision making expertise



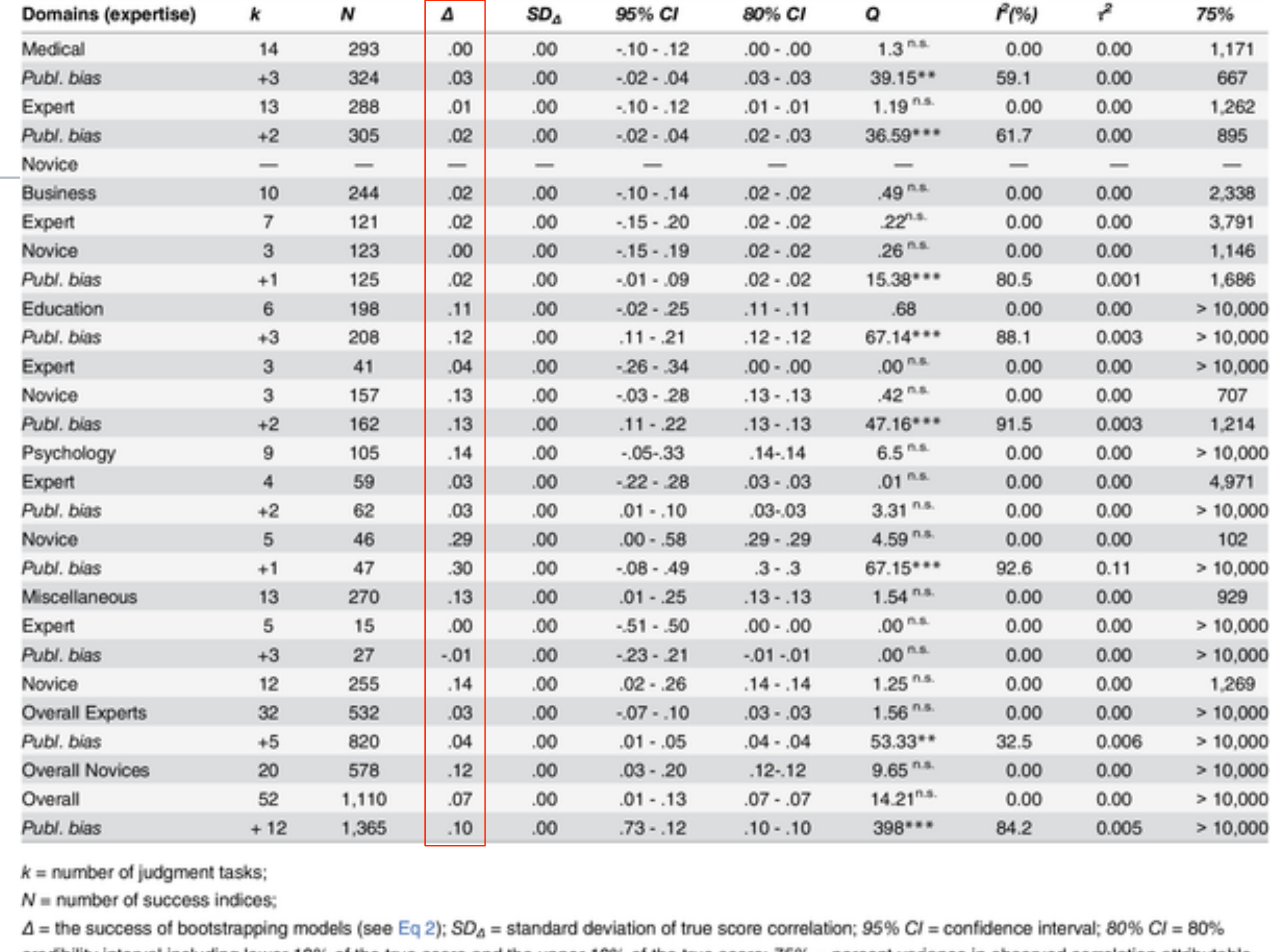
**10 tasks**

Human judges are more accurate

**42 tasks**

Success of bootstrapping models

More than 80% of the tasks (42 of the 52 tasks) were associate with a positive value.



| Domains (expertise) | *k* | *N* | *Δ* | $SD_{\Delta}$ | *95% CI* | *80% CI* | *Q* | $I^2$(%) | $\tau^2$ | 75% |
|---|---|---|---|---|---|---|---|---|---|---|
| Medical | 14 | 293 | .00 | .00 | -.10 - .12 | .00 - .00 | 1.3 $^{n.s.}$ | 0.00 | 0.00 | 1,171 |
| *Publ. bias* | +3 | 324 | .03 | .00 | -.02 - .04 | .03 - .03 | 39.15** | 59.1 | 0.00 | 667 |
| Expert | 13 | 288 | .01 | .00 | -.10 - .12 | .01 - .01 | 1.19 $^{n.s.}$ | 0.00 | 0.00 | 1,262 |
| *Publ. bias* | +2 | 305 | .02 | .00 | -.02 - .04 | .02 - .03 | 36.59*** | 61.7 | 0.00 | 895 |
| Novice | — | — | — | — | — | — | — | — | — | — |
| Business | 10 | 244 | .02 | .00 | -.10 - .14 | .02 - .02 | .49 $^{n.s.}$ | 0.00 | 0.00 | 2,338 |
| Expert | 7 | 121 | .02 | .00 | -.15 - .20 | .02 - .02 | .22$^{n.s.}$ | 0.00 | 0.00 | 3,791 |
| Novice | 3 | 123 | .00 | .00 | -.15 - .19 | .02 - .02 | .26 $^{n.s.}$ | 0.00 | 0.00 | 1,146 |
| *Publ. bias* | +1 | 125 | .02 | .00 | -.01 - .09 | .02 - .02 | 15.38*** | 80.5 | 0.001 | 1,686 |
| Education | 6 | 198 | .11 | .00 | -.02 - .25 | .11 - .11 | .68 | 0.00 | 0.00 | > 10,000 |
| *Publ. bias* | +3 | 208 | .12 | .00 | .11 - .21 | .12 - .12 | 67.14*** | 88.1 | 0.003 | > 10,000 |
| Expert | 3 | 41 | .04 | .00 | -.26 - .34 | .00 - .00 | .00 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| Novice | 3 | 157 | .13 | .00 | -.03 - .28 | .13 - .13 | .42 $^{n.s.}$ | 0.00 | 0.00 | 707 |
| *Publ. bias* | +2 | 162 | .13 | .00 | .11 - .22 | .13 - .13 | 47.16*** | 91.5 | 0.003 | 1,214 |
| Psychology | 9 | 105 | .14 | .00 | -.05-.33 | .14-.14 | 6.5 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| Expert | 4 | 59 | .03 | .00 | -.22 - .28 | .03 - .03 | .01 $^{n.s.}$ | 0.00 | 0.00 | 4,971 |
| *Publ. bias* | +2 | 62 | .03 | .00 | .01 - .10 | .03-.03 | 3.31 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| Novice | 5 | 46 | .29 | .00 | .00 - .58 | .29 - .29 | 4.59 $^{n.s.}$ | 0.00 | 0.00 | 102 |
| *Publ. bias* | +1 | 47 | .30 | .00 | -.08 - .49 | .3 - .3 | 67.15*** | 92.6 | 0.11 | > 10,000 |
| Miscellaneous | 13 | 270 | .13 | .00 | .01 - .25 | .13 - .13 | 1.54 $^{n.s.}$ | 0.00 | 0.00 | 929 |
| Expert | 5 | 15 | .00 | .00 | -.51 - .50 | .00 - .00 | .00 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| *Publ. bias* | +3 | 27 | -.01 | .00 | -.23 - .21 | -.01 -.01 | .00 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| Novice | 12 | 255 | .14 | .00 | .02 - .26 | .14 - .14 | 1.25 $^{n.s.}$ | 0.00 | 0.00 | 1,269 |
| Overall Experts | 32 | 532 | .03 | .00 | -.07 - .10 | .03 - .03 | 1.56 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| *Publ. bias* | +5 | 820 | .04 | .00 | .01 - .05 | .04 - .04 | 53.33** | 32.5 | 0.006 | > 10,000 |
| Overall Novices | 20 | 578 | .12 | .00 | .03 - .20 | .12-.12 | 9.65 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| Overall | 52 | 1,110 | .07 | .00 | .01 - .13 | .07 - .07 | 14.21$^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| *Publ. bias* | + 12 | 1,365 | .10 | .00 | .73 - .12 | .10 - .10 | 398*** | 84.2 | 0.005 | > 10,000 |

*k* = number of judgment tasks;

*N* = number of success indices;

*Δ* = the success of bootstrapping models (see Eq 2); $SD_{\Delta}$ = standard deviation of true score correlation; *95% CI* = confidence interval; *80% CI* = 80%

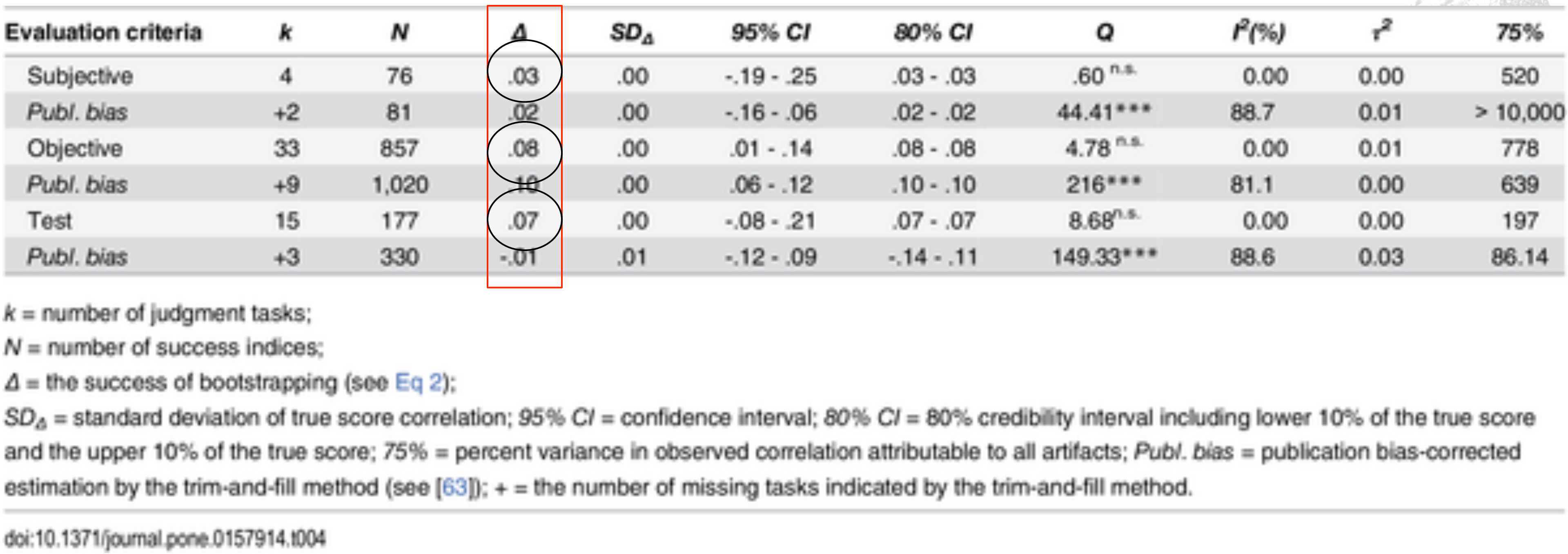
| Domains (expertise) | *k* | *N* | *Δ* | *SD*$_Δ$ | *95% CI* | *80% CI* | *Q* | *I*$^2$(%) | *τ*$^2$ | *75%* |
|---|---|---|---|---|---|---|---|---|---|---|
| Medical | 14 | 293 | .00 | .00 | -.10 - .12 | .00 - .00 | 1.3 $^{n.s.}$ | 0.00 | 0.00 | 1,171 |
| *Publ. bias* | +3 | 324 | .03 | .00 | -.02 - .04 | .03 - .03 | 39.15** | 59.1 | 0.00 | 667 |
| Expert | 13 | 288 | .01 | .00 | -.10 - .12 | .01 - .01 | 1.19 $^{n.s.}$ | 0.00 | 0.00 | 1,262 |
| *Publ. bias* | +2 | 305 | .02 | .00 | -.02 - .04 | .02 - .03 | 36.59*** | 61.7 | 0.00 | 895 |
| Novice | — | — | — | — | — | — | — | — | — | — |
| Business | 10 | 244 | .02 | .00 | -.10 - .14 | .02 - .02 | .49 $^{n.s.}$ | 0.00 | 0.00 | 2,338 |
| Expert | 7 | 121 | .02 | .00 | -.15 - .20 | .02 - .02 | .22$^{n.s.}$ | 0.00 | 0.00 | 3,791 |
| Novice | 3 | 123 | .00 | .00 | -.15 - .19 | .02 - .02 | .26 $^{n.s.}$ | 0.00 | 0.00 | 1,146 |
| *Publ. bias* | +1 | 125 | .02 | .00 | -.01 - .09 | .02 - .02 | 15.38*** | 80.5 | 0.001 | 1,686 |
| Education | 6 | 198 | .11 | .00 | -.02 - .25 | .11 - .11 | .68 | 0.00 | 0.00 | > 10,000 |
| *Publ. bias* | +3 | 208 | .12 | .00 | .11 - .21 | .12 - .12 | 67.14*** | 88.1 | 0.003 | > 10,000 |
| Expert | 3 | 41 | .04 | .00 | -.26 - .34 | .00 - .00 | .00 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| Novice | 3 | 157 | .13 | .00 | -.03 - .28 | .13 - .13 | .42 $^{n.s.}$ | 0.00 | 0.00 | 707 |
| *Publ. bias* | +2 | 162 | .13 | .00 | .11 - .22 | .13 - .13 | 47.16*** | 91.5 | 0.003 | 1,214 |
| Psychology | 9 | 105 | .14 | .00 | -.05-.33 | .14-.14 | 6.5 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| Expert | 4 | 59 | .03 | .00 | -.22 - .28 | .03 - .03 | .01 $^{n.s.}$ | 0.00 | 0.00 | 4,971 |
| *Publ. bias* | +2 | 62 | .03 | .00 | .01 - .10 | .03-.03 | 3.31 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| Novice | 5 | 46 | .29 | .00 | .00 - .58 | .29 - .29 | 4.59 $^{n.s.}$ | 0.00 | 0.00 | 102 |
| *Publ. bias* | +1 | 47 | .30 | .00 | -.08 - .49 | .3 - .3 | 67.15*** | 92.6 | 0.11 | > 10,000 |
| Miscellaneous | 13 | 270 | .13 | .00 | .01 - .25 | .13 - .13 | 1.54 $^{n.s.}$ | 0.00 | 0.00 | 929 |
| Expert | 5 | 15 | .00 | .00 | -.51 - .50 | .00 - .00 | .00 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| *Publ. bias* | +3 | 27 | -.01 | .00 | -.23 - .21 | -.01 -.01 | .00 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| Novice | 12 | 255 | .14 | .00 | .02 - .26 | .14 - .14 | 1.25 $^{n.s.}$ | 0.00 | 0.00 | 1,269 |
| Overall Experts | 32 | 532 | .03 | .00 | -.07 - .10 | .03 - .03 | 1.56 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| *Publ. bias* | +5 | 820 | .04 | .00 | .01 - .05 | .04 - .04 | 53.33** | 32.5 | 0.006 | > 10,000 |
| Overall Novices | 20 | 578 | .12 | .00 | .03 - .20 | .12-.12 | 9.65 $^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| Overall | 52 | 1,110 | .07 | .00 | .01 - .13 | .07 - .07 | 14.21$^{n.s.}$ | 0.00 | 0.00 | > 10,000 |
| *Publ. bias* | + 12 | 1,365 | .10 | .00 | .73 - .12 | .10 - .10 | 398*** | 84.2 | 0.005 | > 10,000 |

*k* = number of judgment tasks;

*N* = number of success indices;

*Δ* = the success of bootstrapping models (see Eq 2); *SD*$_Δ$ = standard deviation of true score correlation; *95% CI* = confidence interval; *80% CI* = 80% credibility interval including lower 10% of the true score and the upper 10% of the true score; *75%* = percent variance in observed correlation attributable

# Results of the bare-bones meta-analysis of the success bootstrapping organized by the type of evaluation criterion

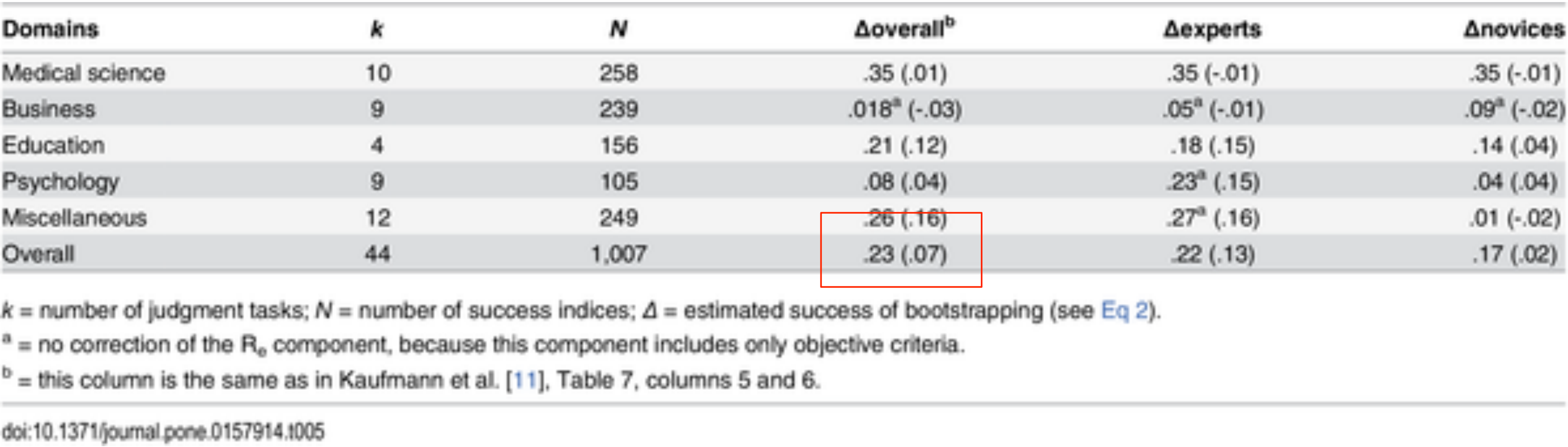


| Evaluation criteria | *k* | *N* | *Δ* | *SD*$_Δ$ | *95% CI* | *80% CI* | *Q* | *I*$^2$(%) | *τ*$^2$ | *75%* |
|---|---|---|---|---|---|---|---|---|---|---|
| Subjective | 4 | 76 | .03 | .00 | -.19 - .25 | .03 - .03 | .60 $^{n.s.}$ | 0.00 | 0.00 | 520 |
| *Publ. bias* | +2 | 81 | .02 | .00 | -.16 - .06 | .02 - .02 | 44.41*** | 88.7 | 0.01 | > 10,000 |
| Objective | 33 | 857 | .08 | .00 | .01 - .14 | .08 - .08 | 4.78 $^{n.s.}$ | 0.00 | 0.01 | 778 |
| *Publ. bias* | +9 | 1,020 | .10 | .00 | .06 - .12 | .10 - .10 | 216*** | 81.1 | 0.00 | 639 |
| Test | 15 | 177 | .07 | .00 | -.08 - .21 | .07 - .07 | 8.68$^{n.s.}$ | 0.00 | 0.00 | 197 |
| *Publ. bias* | +3 | 330 | -.01 | .01 | -.12 - .09 | -.14 - .11 | 149.33*** | 88.6 | 0.03 | 86.14 |

*k* = number of judgment tasks;
*N* = number of success indices;
*Δ* = the success of bootstrapping (see Eq 2);
*SD*$_Δ$ = standard deviation of true score correlation; *95% CI* = confidence interval; *80% CI* = 80% credibility interval including lower 10% of the true score and the upper 10% of the true score; *75%* = percent variance in observed correlation attributable to all artifacts; *Publ. bias* = publication bias-corrected estimation by the trim-and-fill method (see [63]); + = the number of missing tasks indicated by the trim-and-fill method.

doi:10.1371/journal.pone.0157914.t004

# The success of bootstrapping according to bare-bones (in brackets) and psychometrically-corrected lens model indices



| Domains | *k* | *N* | Δoverall[b] | Δexperts | Δnovices |
|---|---|---|---|---|---|
| Medical science | 10 | 258 | .35 (.01) | .35 (-.01) | .35 (-.01) |
| Business | 9 | 239 | .018[a] (-.03) | .05[a] (-.01) | .09[a] (-.02) |
| Education | 4 | 156 | .21 (.12) | .18 (.15) | .14 (.04) |
| Psychology | 9 | 105 | .08 (.04) | .23[a] (.15) | .04 (.04) |
| Miscellaneous | 12 | 249 | .26 (.16) | .27[a] (.16) | .01 (-.02) |
| Overall | 44 | 1,007 | .23 (.07) | .22 (.13) | .17 (.02) |

*k* = number of judgment tasks; *N* = number of success indices; *Δ* = estimated success of bootstrapping (see Eq 2).

[a] = no correction of the $R_e$ component, because this component includes only objective criteria.

[b] = this column is the same as in Kaufmann et al. [11], Table 7, columns 5 and 6.

doi:10.1371/journal.pone.0157914.t005

# Conclusions

- Models are more accurate than both novice and expert human judges.
- The success of bootstrapping models is underestimated (without a psychometric meta-analytic evaluation).
- But, we only evaluated linear models, although non-linear models are more user-friendly (Katsikopoulos, Machery, Pachur, & Wallin, 2008)

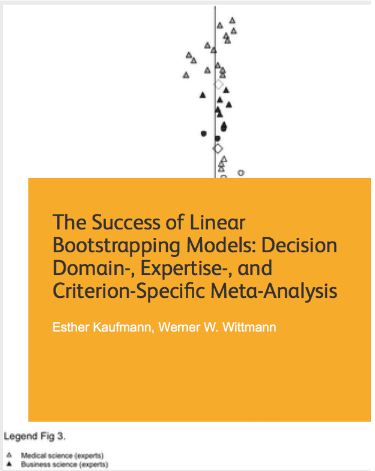
# Thank you

Also on behalf of Professor Wittmann

esther.kaufmann@gmx.ch

wittmann@xi.psychologie.uni-mannheim.de

ResearchGate RG



The Success of Linear Bootstrapping Models: Decision Domain-, Expertise-, and Criterion-Specific Meta-Analysis

Esther Kaufmann, Werner W. Wittmann

Legend Fig 3.

△ Medical science (experts)
▲ Business science (experts)

Brunswik Society Newsletter 2016

Kaufmann, E., & Wittmann, W. W. (2016) The Success of Linear Bootstrapping Models: Decision Domain-, Expertise-, and Criterion-Specific Meta-Analysis. *PLoS ONE 11*(6): e0157914. doi:10.1371/journal.pone.0157914

Kaufmann, E., Reips, U.-D., & Wittmann, W. W. (2013). A critical meta-analysis of Lens Model Studies in human judgment and decision-making. *PLoS ONE 8*(12): e83528. doi:10.1371/journal.pone.0083528