Rethinking External Validity and Official Systematic Reviews

Jacob Alex Klerman, Abt Associates

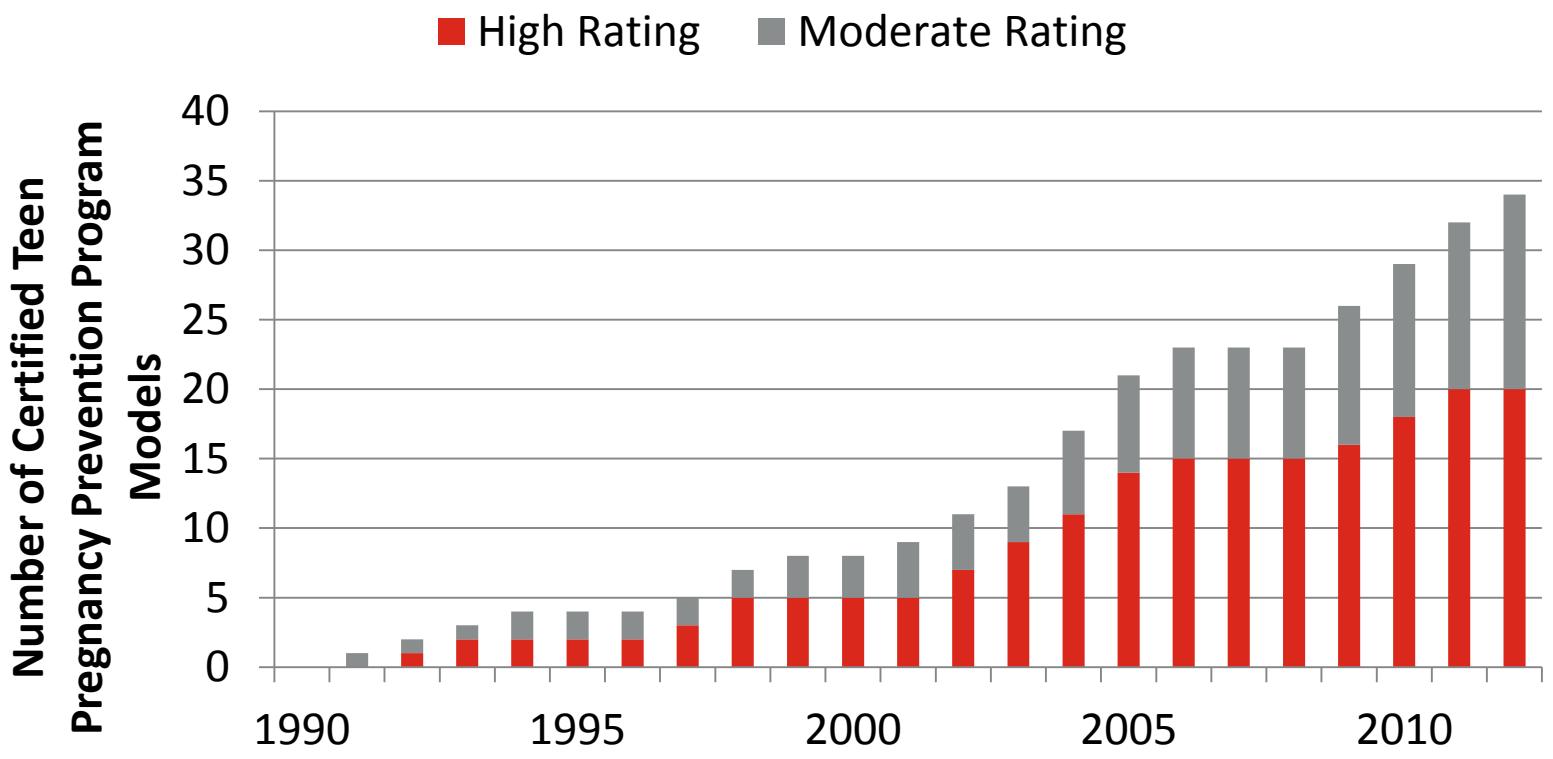
Research and Evaluation Conference on Self-Sufficiency (RECS)

June 2016, Washington DC

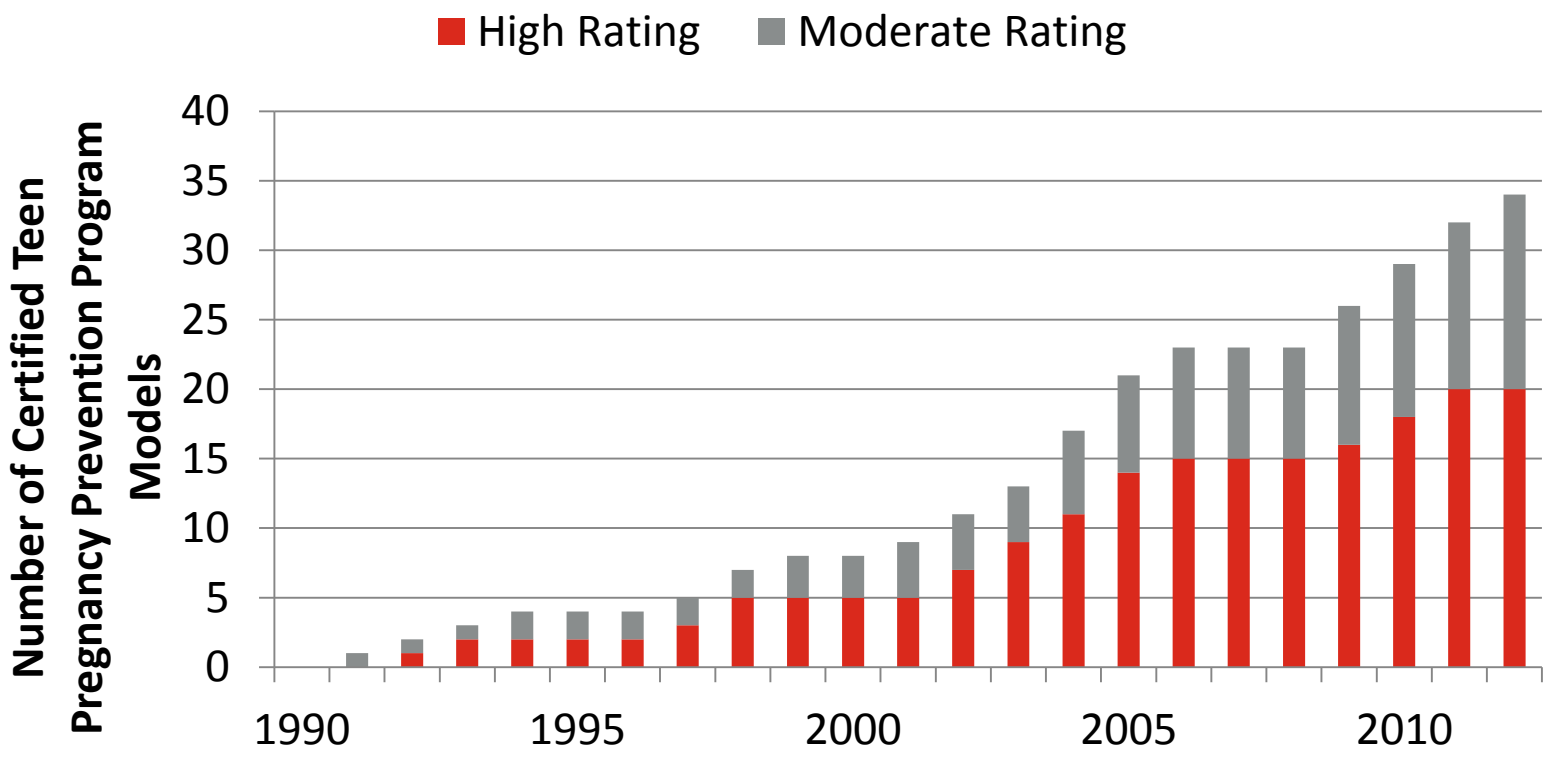
Context

- Ad hoc meeting on “Principles of External Validity for Systematic Evidence Reviews Meeting”
  - Sponsored by the Teen Pregnancy Prevention Evidence Review
  - Washington DC, May 2014
- Special Issue of *Evaluation Review* on “External Validity in Official Reviews”
  - Guest Editor: T’Pring Westbrook
  - To appear later in 2016
- This presentation is the core of my “Editor’s Essay” in the Special Issue
  - Presenting my reaction to (better, opposition to) one of the themes raised by the authors of the articles in the Special Issue
  - Consistent with that purpose, talk is purely conceptual
  - But with implications for official review practice!

“Everyone should have these Problems”



“Everyone should have these Problems”



How should sites choose a program model?

Considerations

- Internal Validity
- Any Impact
- Magnitude of Impact
- Precision
- External Validity

Considerations

- Internal Validity
- Any Impact
- Magnitude of Impact
- Precision
- External Validity

These considerations are often conflicting. How should they be balanced?

Response: Official Systematic Reviews

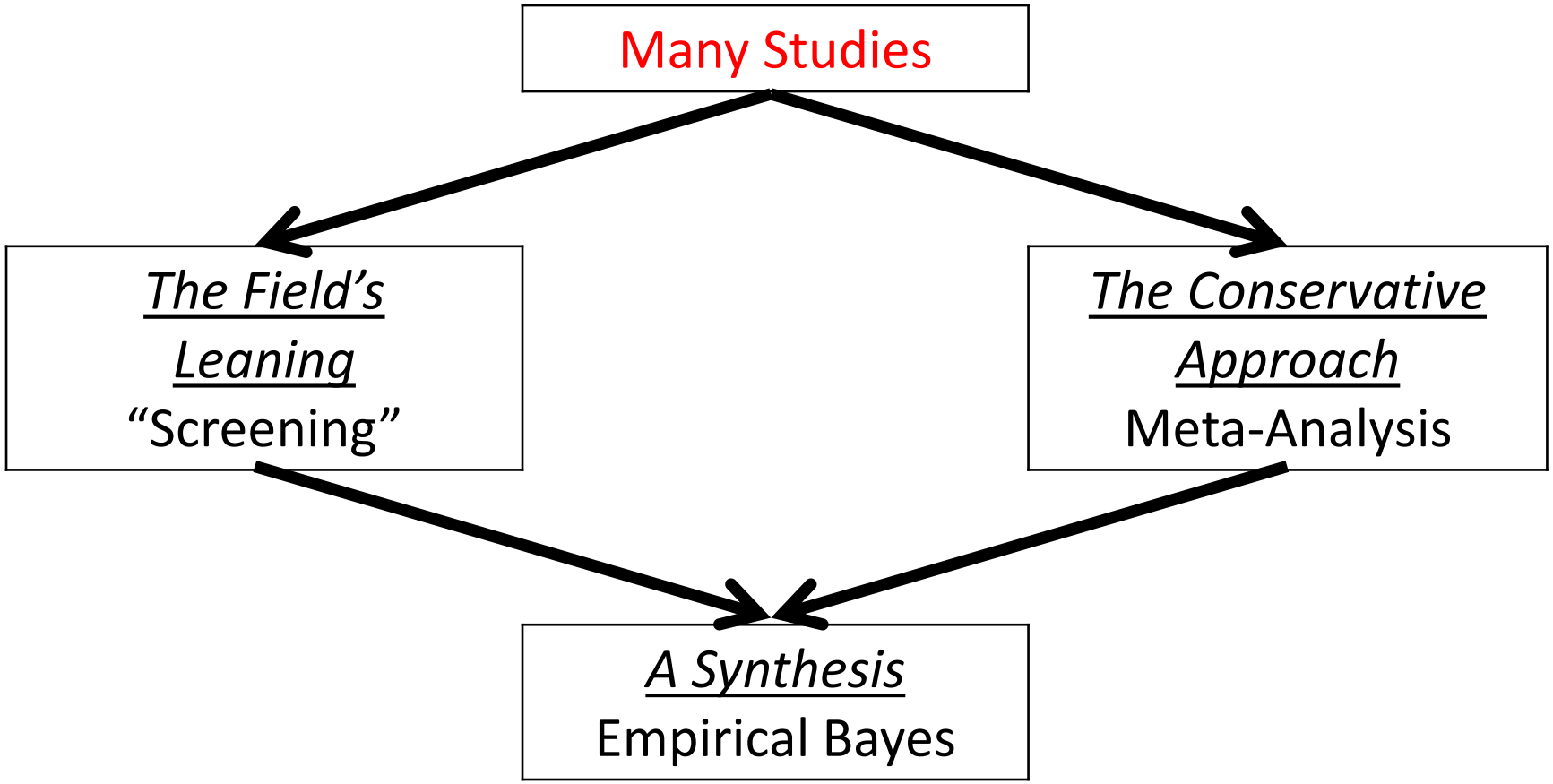
- Descriptive
  - WWC/What Works Clearing House (Education)
  - CLEAR/Clearinghouse for Labor Evaluation and Research
- Prescriptive (i.e., review results affect funding)
  - TPP/Teen Pregnancy Prevention Evidence Review
  - HomVEE/Home Visiting Evidence of Effectiveness

Response: Official Systematic Reviews

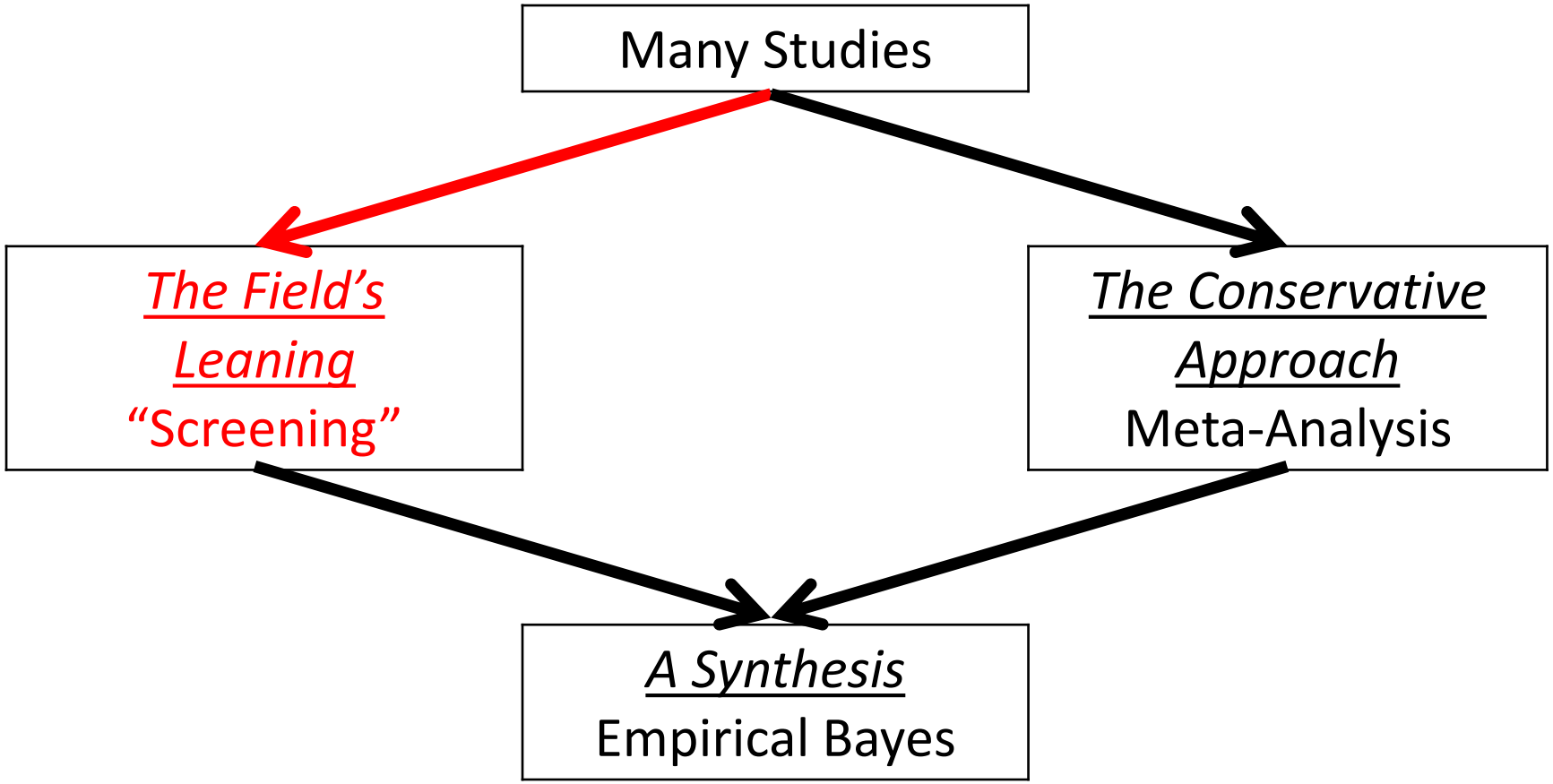
- Descriptive
  - WWC/What Works Clearing House (Education)
  - CLEAR/Clearinghouse for Labor Evaluation and Research
- Prescriptive (i.e., review results affect funding)
  - TPP/Teen Pregnancy Prevention Evidence Review
  - HomVEE/Home Visiting Evidence of Effectiveness

1. Review Literature
   - Assemble papers
   - Assess quality: meets/does not meet standards
2. Disseminate Results
   - Reviews of individual programs
   - Reviews of program areas
   - <sometimes> Web-based choice tool

Plan for the Talk



Plan for the Talk



Web Based Choice Tool

U.S. Department of Health and Human Services

Search

Teen Pregnancy Prevention Evidence Review

HOME | FIND A PROGRAM | PUBLICATIONS | ABOUT THE REVIEW | REVIEWED STUDIES | FAQS | CONTACT US

Home > Programs

Programs

Find a program based on...

Program Type

- ☐ Abstinence-based
- ☐ Clinic-based
- ☐ Sexuality education
- ☐ Special populations
- ☐ Youth development

To filter programs, use the filter criteria in the left-hand panel. To sort results, click on the column heading.

Below you can see all 37 programs that have met the review criteria for evidence of effectiveness.

Program Name	High-Quality Randomized Trial	Moderate-Quality Randomized Trial	Moderate-Quality Quasi-Experiment	Number of Reviewed Studies
¡Cuídate!	✔			1
Aban Aya Youth Project		✔		1

http://tppevidencereview.aspe.hhs.gov/EvidencePrograms.aspx

Comments from TPP Meeting

- *Ignoring external validity is a frequent criticism of systematic reviews*
- *Treatment effectiveness for whom?*
- *How can systematic reviews assist decision-makers determine which program models are the best fit for their needs?*
- *Lack of widely used or accepted standards to assess external validity; i.e., whether causal relationship generalizes beyond the study*

One Response: Screening

- "Screening" is one possible approach
  - i.e., (at least when possible) select among studies demonstrated effective in the (demographic) group of interest
  - Better: Select the "best" (i.e., largest impact) among studies demonstrated effective in the (demographic) group of interest
- Several of the papers in the *Evaluation Review* Special Issue seem to be advocating "screening" (or something similar)

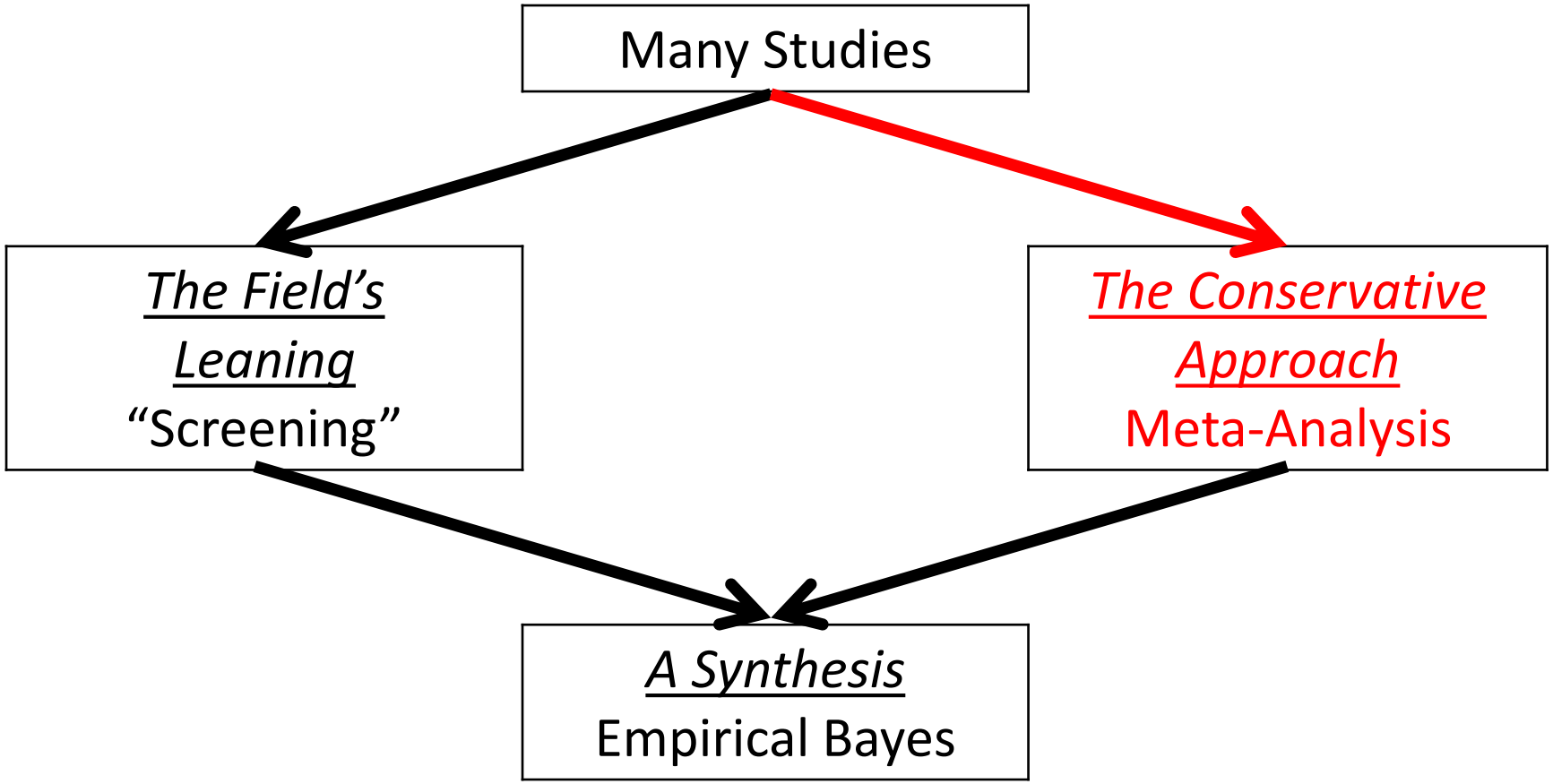
One Response: Screening

- "Screening" is one possible approach
  - i.e., (at least when possible) select among studies demonstrated effective in the (demographic) group of interest
  - Better: Select the "best" (i.e., largest impact) among studies demonstrated effective in the (demographic) group of interest
- Several of the papers in the *Evaluation Review* Special Issue seem to be advocating "screening" (or something similar)

I want to argue that "screening" is a bad idea

To see why consider Meta-Analysis …

Plan for the Talk



(Regression) Model for Impacts

Reported Impact: for a given study, *s*, of a given model *m*, on a given (demographic) group, *g*

Estimated (pooled) impact for model *m*

Differential impact for group *g* for model *m*

Regression residual

$$I_{m,g,s} = \alpha_m + d_g \gamma_{m,g} + \varepsilon_{m,g,s}$$

- Single estimate of magnitude
  - By optimally weighting across available studies
  - Pooling across all (demographic) groups

(Regression) Model for Impacts

Reported Impact: for a given study, *s*, of a given model *m*, on a given (demographic) group, *g*

Estimated (pooled) impact for model *m*

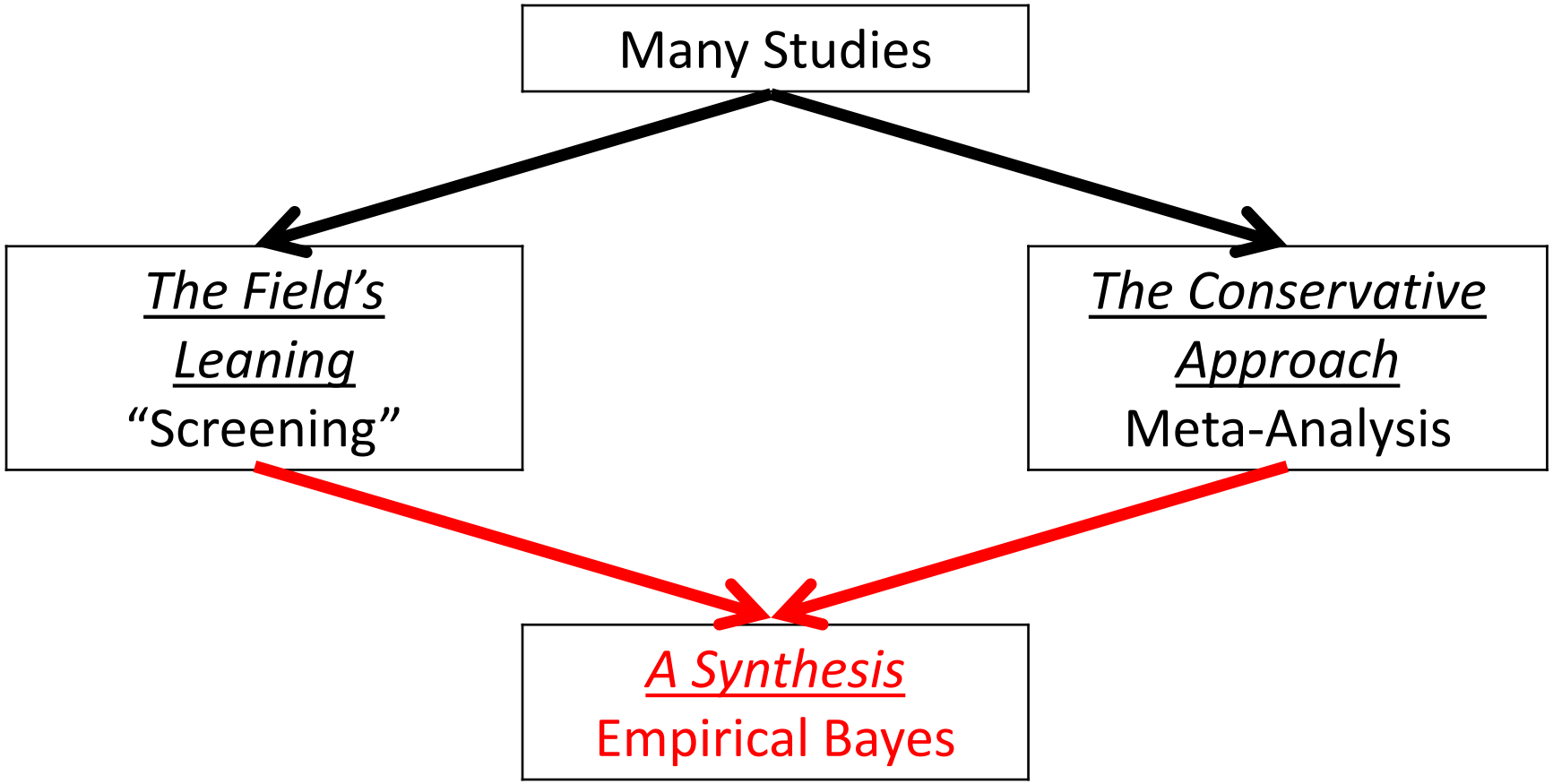
Differential impact for group *g* for model *m*

Regression residual

$$I_{m,g,s} = \alpha_m + d_g \gamma_{m,g} + \varepsilon_{m,g,s}$$

- Single estimate of magnitude
  - By optimally weighting across available studies
  - Pooling across all (demographic) groups
- Implicit presumption of homogeneous impacts
  - Unless there is clear evidence to the contrary

Plan for the Talk



(Regression) Model for Impacts

Reported Impact: for a given study, *s*, of a given model *m*, on a given (demographic) group, *g*

Estimated (pooled) impact for model *m*

Differential impact for group *g* for model *m*

Regression residual

$$I_{m,g,s} = \alpha_m + d_g \gamma_{m,g} + \varepsilon_{m,g,s}$$

- Single estimate of magnitude
  - By optimally weighting across available studies
  - Pooling across all (demographic) groups
- Implicit presumption of homogeneous impacts
  - Unless there is clear evidence to the contrary

"Clear evidence" (i.e. statistical significance) is too high a standard

Empirical Bayes

- Goal: Help sites to identify the “best program” given available evidence
- i.e., a focus on magnitude of the impact, adjusting for
  - Precision of estimate
  - Evidence of heterogeneity of impact by (demographic) group

Empirical Bayes

- Goal: Help sites to identify the "best program" given available evidence
- i.e., a focus on magnitude of the impact, adjusting for
  - Precision of estimate
  - Evidence of heterogeneity of impact by (demographic) group

- Empirical Bayes formalizes this intuition
- Generates a predicted impact for each program model *x* demographic group
  - While—appropriately—"shrinking" noisy estimates (e.g., interactions) towards the overall mean
  - The less precise, the closer to the overall mean

Empirical Bayes

- Goal: Help sites to identify the "best program" given available evidence
- i.e., a focus on magnitude of the impact, adjusting for
  - Precision of estimate
  - Evidence of heterogeneity of impact by (demographic) group

- Empirical Bayes formalizes this intuition
- Generates a predicted impact for each program model *x* demographic group
  - While—appropriately—"shrinking" noisy estimates (e.g., interactions) towards the overall mean
  - The less precise, the closer to the overall mean

See paper for (much) more (formal) detail

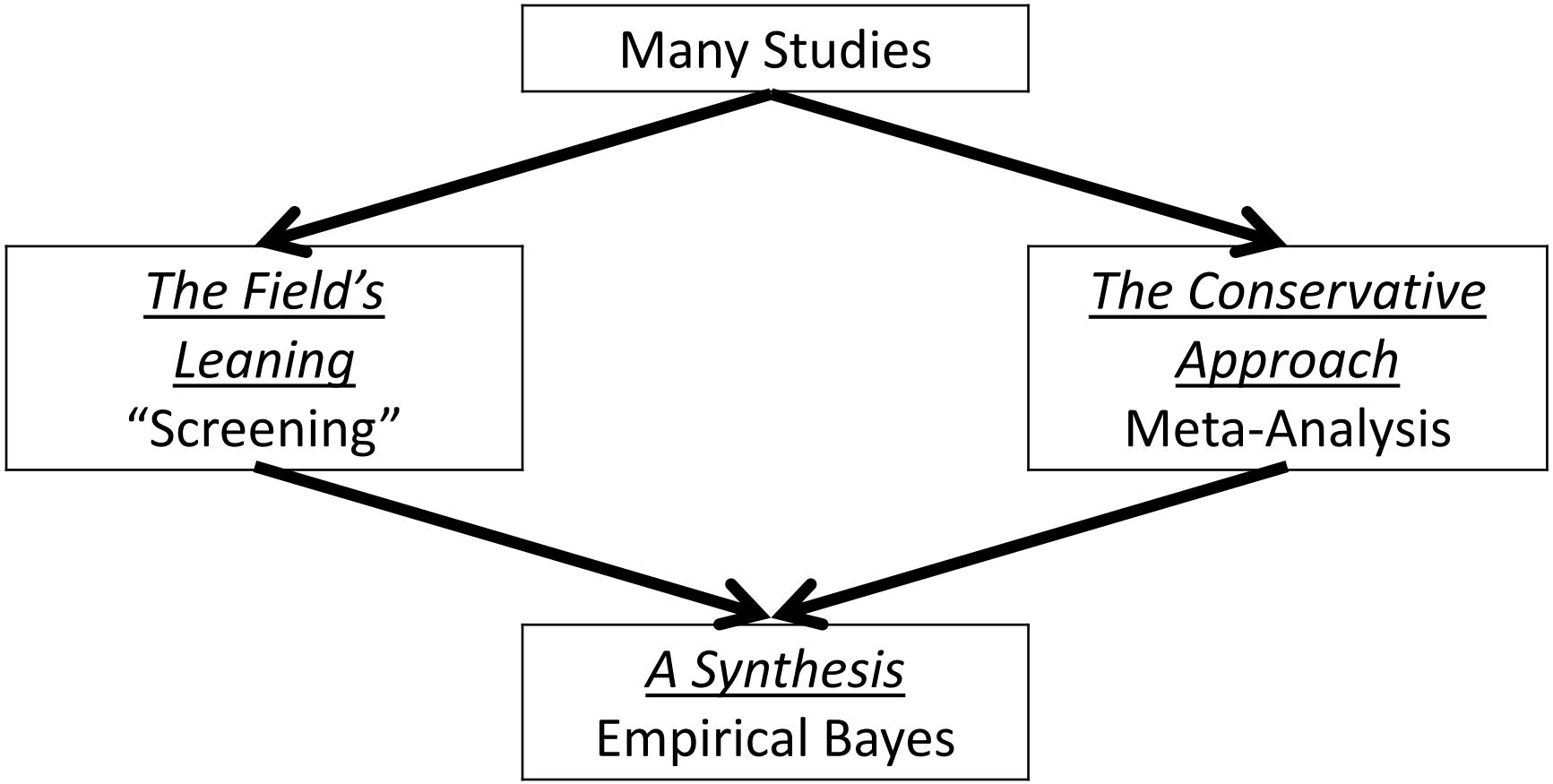
Empirical Bayes

$$I_{m.g,s} = \alpha_m + \beta_g + d_g \gamma_{m,g} + \varepsilon_{m,g,s}$$

- Treat α and γ as random
- Estimate variance terms across programs
- Treat posterior (shrunk) means as best predictiorm: for this program, for this demographic group

See paper for (much) more formal discuss

Plan for the Talk



Discussion

- Given that our goal is to identify the “Best Program”
  - Empirical Bayes seems like the “ideal” approach
  - But, it requires a major “study” (for TPP: underway at Abt)
- Study seems worth doing
  - Will help us to understand to what extent my conjectures here are correct/useful
- Short of that, remember that …
  - Individual estimates are extremely noisy
  - Evidence for heterogeneity by demographic group appears—at least to me—to be weak
  - So, “best program” will often not have been tested on—or will not have clear evidence for effectiveness in—the demographic group of interest

Discussion (cont'd)

- The decision rule implicitly advocated by some of the papers in the Special Issue
  - Screen on demonstrated impact in the demographic group of interest
  - Choose the maximum estimate among those with demonstrated impact

  … is unlikely to choose the "best program" (given the available evidence)

- Instead, there is likely to be considerable information in the estimates for "other" demographic groups
  - Probably more than in the estimates for "this" demographic group
  - Don't ignore that evidence!

Rethinking External Validity and Official Systematic Reviews

Jacob Alex Klerman, Abt Associates

Research and Evaluation Conference on Self-Sufficiency (RECS)

June 2016, Washington DC

So what is the Implicit Weighting?

- Internal Validity
- Any Impact
- Magnitude of Impact
- Precision
- External Validity

1. Internal Validity
2. Any Impact
3. <sort of> Internal Validity
4. <sort of> Precision
5. External Validity
6. Magnitude of Impact

So what is the Implicit Weighting?

- Internal Validity → 1. Internal Validity
- Any Impact → 2-4. Any Impact
- Magnitude of Impact → 2-4. Magnitude of Impact
- Precision → 2-4. <sort of> Precision
- External Validity → 5. External Validity

(Regression) Model for Impacts

Reported Impact: for a given study, *s*, on a given demographic group, *g*

Estimated (pooled) impact

Differential impact for demographic group *g*

Pure (unmodelled) inter-study variation

Pure sampling variability

$$I_{g,s} = \alpha + d_g \gamma_g + \eta_{g,s} + \varepsilon_{g,s}$$

- WLS/Weighted Least Squares to account for
  - Sampling variability of each estimated impact
  - Other inter-study variation in impact

(Regression) Model for Impacts

$$I_{g,s} = \alpha + d_g \gamma_g + \eta_{g,s} + \varepsilon_{g,s}$$

- Single estimate of magnitude
  - By optimally weighting across available studies
  - Pooling across all demographic subgroups
- Implicit presumption of homogeneous impacts
  - Unless there is clear evidence to the contrary

(Regression) Model for Impacts

$$I_{g,s} = \alpha + d_g \gamma_g + \eta_{g,s} + \varepsilon_{g,s}$$

- Single estimate of magnitude
  - By optimally weighting across available studies
  - Pooling across all demographic subgroups
- Implicit presumption of homogeneous impacts
  - Unless there is clear evidence to the contrary

Magnitude is primary concern;
external validity is only a (weak) secondary concern

(Regression) Model for Impacts

$$I_{g,s} = \alpha + d_g \gamma_g + \eta_{g,s} + \varepsilon_{g,s}$$

- Single estimate of magnitude
  - By optimally weighting across available studies
  - Pooling across all demographic subgroups
- Implicit presumption of homogeneous impacts
  - Unless there is clear evidence to the contrary

Magnitude is primary concern;
external validity is only a (weak) secondary concern

Can We Detect Heterogeneity?

- Heterogeneity of impacts by (demographic) group is an “interaction”
  - Not: How does impact vary with program model?
  - Not: How does impact vary with (demographic) group? <does not matter for differential impact>
  - But: How does impact vary, for a given program, by (demographic) group?
- We know interactions are hard to detect
  - Most observed variation across subgroups will be “noise”
- So, we want to discount (“shrink” towards the mean impact for the model)—but not ignore—such evidence
  - Empirical Bayes provides a formal/optimal way to do so
  - See paper for (much) more formal detail

Web Based Choice Tool

U.S. Department of Health and Human Services

Search

Teen Pregnancy Prevention Evidence Review

HOME | FIND A PROGRAM | PUBLICATIONS | ABOUT THE REVIEW | REVIEWED STUDIES | FAQS | CONTACT US

Home > Programs

Programs

Find a program based on...

Program Type

- ☐ Abstinence-based
- ☐ Clinic-based
- ☐ Sexuality education
- ☐ Special populations
- ☐ Youth development

To filter programs, use the filter criteria in the left-hand panel. To sort results, click on the column heading.

Below you can see all 37 programs that have met the review criteria for evidence of effectiveness.

Program Name	High-Quality Randomized Trial	Moderate-Quality Randomized Trial	Moderate Quality Quasi-Experiment	Number of Reviewed Studies
¡Cuídate!	✔			
Aban Aya Youth Project		✔		1

http://tppevidencereview.aspe.hhs.gov/EvidencePrograms.aspx

Web Based Choice Tool

U.S. Department of Health and Human Services

Search

Teen Pregnancy Prevention Evidence Review

HOME | FIND A PROGRAM | PUBLICATIONS | ABOUT THE REVIEW | REVIEWED STUDIES | FAQS | CONTACT US

Home > Programs

Programs

Find a program based on...

Program Type

- ☐ Abstinence-based
- ☐ Clinic-based
- ☐ Sexuality education
- ☐ Special populations
- ☐ Youth development

To filter programs, use the filter criteria in the left-hand panel. To sort results, click on the column heading.

Below you can see all 37 programs that have met the review criteria for evidence of effectiveness.

Program Name	High-Quality Randomized Trial	Moderate-Quality Randomized Trial	Moderate-Quality Quasi-Experiment	Number of Reviewed Studies
¡Cuídate!	✔			1
Aban Aya Youth Project		✔		1

http://tppevidencereview.aspe.hhs.gov/EvidencePrograms.aspx

Selection Guided via Check Boxes

Find a program based on …

- Program Type
- Program Length
- Target Population
- Research Shows Impact On

In Particular, “Target Population”

Find a program based on …

- Program Type
- Program Length
- ***Target Population***
- Research Shows Impact On