Mathematica
Progress Together

Response to the Request for Comment

Protocol for the Pathways to Work Evidence Clearinghouse: Methods and Standards, Version 2.0

Federal Register: Notice 90 FR 21310
June 18, 2025

Submitted to:

Administration for Children and Families
U.S. Department of Health and Human Services
330 C Street, SW
Washington, DC 20201
Attention: Lauren Supplee, Deputy Assistant Secretary for Planning, Research, and Evaluation

Submitted by:

Mathematica
1100 First Street, NE, 12th Floor
Washington, DC 20002-4221
Phone: (202) 484-9220
Fax: (609) 228-4958

This page has been left blank for double-sided copying.

Contents

This page has been left blank for double-sided copying.

I. Introduction

The Pathways to Work Evidence Clearinghouse (Pathways) represents the Administration for Children and Families' (ACF's) commitment to supporting employment as a policy priority and to supplying employment and training program leaders and other decision makers with evidence-based information on allocating program resources efficiently and improving the employment-related outcomes of people with low incomes. Through the proposed updates to the Pathways review protocol, ACF has a unique opportunity to advance its mission and serve as an innovative leader in the federal evidence clearinghouse ecosystem, particularly by establishing standards for cost analyses.

Mathematica's response to this request for comment recognizes and applauds ACF's efforts to update the review protocol while highlighting opportunities for additional changes that could further improve the transparency of the Pathways evidence review process. This work could enhance program effectiveness and advance ACF's goal of providing decision makers with useful, accurate, and clear information to inform practical decision making and achieve the best possible outcomes. To identify these opportunities, we have drawn on our experience operating multiple evidence review clearinghouses for the federal government and other partners, including the What Works Clearinghouse for the U.S. Department of Education; the Home Visiting Evidence of Effectiveness for ACF's Office of Planning, Research, and Evaluation; the Clearinghouse for Labor Evaluation and Research for the U.S. Department of Labor; the Teen Pregnancy Prevention Evidence Review for the Office of Population Affairs; and Pathways and its predecessor the Employment Strategies for Low-Income Adults Evidence Review. To complement this perspective, we also draw on our experience conducting rigorous evaluations of employment programs, including the Next Generation of Enhanced Employment Strategies (NextGen) and Employment Coaching Projects. Mathematica has worked with federal project officers and experts in the field to develop and update review standards, protocols, and other tools to elevate the best evidence to support decision makers. We appreciate the chance to apply our experience through ACF's request for comments on Pathways to Work Evidence Clearinghouse: Methods and Standards, Version 2.0.

II. Responses

A. Identifying eligible studies (Chapter 2, Sections 2.1–2.3, pages 7–11)

The proposed updates to Chapter 2 of the draft protocol are reasonable and appropriate for identifying eligible studies. However, to **keep pace with the rapidly evolving technologies** used to perform literature searches and screening, including artificial intelligence (AI), and to increase the **transparency** of the procedures used to identify eligible studies, we recommend ACF consider additional changes to this chapter:

- **Describe search and screening procedures in a separate, updateable document rather than embedding them in the static review protocol.** Maintaining the search and screening procedures in a separate, updatable document would give ACF the flexibility to make timely

revisions as new tools and technologies become available. For example, the current draft protocol specifies the use of trained (human) screeners to apply a set of eight eligible criteria to identified studies. If ACF's agencywide technology policy changes to encourage or require the use of large language models or other AI tools for routine activities such as study screening, maintaining a separate, updatable document would enable ACF to adapt to this change without having to revise and seek public comment on the entire review protocol.

- **Clarify the objectives of the search process.** ACF could increase the transparency of the review protocol by adding information to Chapter 2 on the purpose of the search process. For example, given that Pathways already includes published reviews of more than 300 studies, is the objective of additional search and screening activities to ensure that the latest research is reviewed, to build a comprehensive inventory of unreviewed studies, or another purpose? Elaborating the existing text on research prioritization (Section 2.3, page 11) would help clarify the objectives, supporting users' confidence that Pathways content is up to date and researchers' understanding about how and when their research will be reviewed.
- **Explain differences in re-reviews.** Misunderstandings can arise when two clearinghouses come to different conclusions about a study. These differences are often due to legitimate distinctions in mission or methods. However, without sufficient explanation, users could mistakenly interpret the differences as resulting from errors or oversights. In the section of the review protocol on coordination with other federal evidence reviews (Section 2.1.5, page 9), we recommend ACF develop and explain a transparent approach for how Pathways will document any differences that might result from re-reviews.
- **Preserve documentation of prior search procedures.** We support ACF's efforts to keep the search procedures up to date and aligned with current best practices and technologies. However, in line with Preferred Reporting Items for Systematic reviews and Meta-Analyses standards, we also recommend preserving documentation of prior search procedures, so users can understand how those procedures have evolved over time and were applied in each review effort.

B. Baseline adjustments and equivalence (Chapter 3, Section 3.2.5, Steps 2 and 3, pages 21–22)

We support the proposed change to prioritize baseline effect sizes over statistical significance in large studies, which will improve the rigor of the study quality ratings by more accurately differentiating between low- and moderate-quality evidence. Baseline effect sizes are informative about potential bias in a study, while statistical significance can be misleading and arbitrary (Ho et al., 2007; Imai et al., 2008). To further improve the **rigor** of the study quality ratings, promote **consistency** in the application of ratings across studies, and **create positive incentives** for researchers in the field, we recommend ACF consider three additional changes to the baseline equivalence standards:

- **Eliminate statistical significance as a criterion so Pathways uses baseline effect sizes to assess baseline equivalence in all studies.** The draft protocol continues to use statistical significance for assessing baseline equivalence in smaller studies, despite acknowledging the flaws of this approach in the context of larger studies. This approach enables underpowered studies with a high risk of bias to meet Pathways standards because they cannot detect large group differences. It also penalizes researchers for designing adequately powered studies by holding them to a higher standard. Although using effect sizes to assess baseline equivalence in all studies might require additional information requests to authors and increase the number of studies failing to meet the revised standards, we believe the benefits outweigh costs and will provide positive incentives for researchers to provide adequate reporting on baseline equivalence going forward.
- **Do not require statistical adjustments when baseline differences are small.** Assessing baseline equivalence using effect sizes for all studies would also provide opportunity to simplify, improve, and support efficient application of another part of the review standards—namely, requirements for statistical adjustments. When studies have small baseline effect sizes (such as those smaller than 0.05 standard deviations), statistical adjustments have little impact. Therefore, it is not necessary to require that researchers control for covariates when baseline differences are small.
- **Clarify how Pathways will calculate baseline effect sizes, especially for categorical variables.** A large baseline effect size in a low-incidence group (such as when most of the group are in one condition with a sample composed of less than 3 percent of that group) may not indicate meaningful bias. Furthermore, for categorical variables, if Pathways assesses equivalence for all reported categories, researchers have incentive to combine low incidence categories into a single, larger, "other" category, to avoid reporting imbalances for one or more individual groups. To avoid this incentive and ensure studies are not rated low unnecessarily, the protocol could specify that Pathways will assess baseline equivalence only for sufficiently large categories of a pre-determined size.

C. Determining programs' evidence of effectiveness (Chapter 4, Section 4.2, pages 28–31)

The way Pathways defines program effectiveness plays a crucial role in shaping public understanding of which programs are evidence based. Although not reflected in the current updates to the review protocol, we recommend strengthening the program effectiveness definitions to provide **positive incentives for researchers** and improve their **rigor and usefulness**. These changes will lead to effectiveness ratings that better reflect the **strength of evidence** and **support better decisions**. We have three recommendations to improve the strength of evidence:

- **Replace vote counting with a meta-analysis approach, such as a weighted or unweighted average of effect sizes.** Vote counting is inadequate for assessing the strength of evidence for a program. Counts of studies with statistically significant findings may not identify evidence-based

programs reliably because of known limitations with the approach (McKenzie & Brennan, 2024). For example, Hedges and Vevea (1998) show that the vote counting approach cannot reliably distinguish positive from null effects except under narrow circumstances.

We recognize that a downside of the meta-analysis approach is the need to extract more data from studies or request missing data from authors, which can add time and cost to reviews. If too few authors report data on the magnitude and precision of findings needed for meta-analysis, Pathways could consider alternatives to meta-analysis suggested by McKenzie and Brennan (2024), which are also more reliable alternatives to vote counting.

- **Either do not use study counts in the effectiveness ratings or update the approach Pathways uses to define studies.** The definition of a study (Section 3.1, pages 12–13) influences how Pathways reports on findings from individual studies and how it assigns effectiveness ratings. Although for the first purpose, the Pathways definition of a study may support clear and understandable reporting at the study level, using the definition of a study to count studies in assigning effectiveness ratings provides poor incentives for researchers. Because Pathways defers to study authors to decide whether multiple site-level findings represent one or multiple studies, the study definition provides poor incentives for researchers when used to measure the strength of evidence for programs. This approach creates incentives for researchers to report findings in a way most favorable to the program, but inaccurately, such as achieving a rating of Well-Supported by selectively reporting positive results from different sites. Researchers might search for combinations of sites that provide positive results, even when the overall average across sites is not statistically significant.

  Adopting a meta-analysis approach and avoiding criteria related to the number of studies would remove the incentives, because Pathways would focus on the average impact across sites and studies. Furthermore, in avoiding criteria related to the number of studies, Pathways can define studies in a way that best supports simple and clear reporting of findings at the study level without concern about poor incentives. If it is not feasible to adopt a meta-analysis approach, ACF could instead refine the definition of a study to limit authors' influence over evidence classification. For example, ACF could define studies using criteria that are not easily manipulable by researchers, such as whether two analyses share a group formation procedure, such as the same matching or random assignment process. Under such a definition, Pathways might always measure a study's impact finding across multiple sites, whether authors report pooled or site-by-site impacts, or in one or more manuscripts. A limitation of this approach is that carefully balancing the goals of both uses of the study definition to support clear reporting and avoid poor incentives could force tradeoffs that satisfy one goal at the expense of the other. As another alternative, ACF could replace study counts in the definition of effectiveness ratings with criteria about the breadth of evidence, such as a minimum combined number of sites, breadth of sample characteristics, and implementation approaches. This approach recognizes that the purpose of requiring evidence from multiple studies to achieve a Well-Supported rating is to ensure a reasonably broad evidence base. That purpose can be achieved without reference to a study definition.

- **Update the plain-language definitions of the evidence classifications.** The plain-language definitions provided in Section 4.2 of the protocol risk overstating evidence strength in a way that could lead decision makers to be overconfident in achieving positive results. Specifically, the draft protocol states that a program in the Well-Supported category "is likely to improve outcomes in a domain if the program was replicated." Although the text also cautions that "no two implementations are identical" and that users "should not view this rating as a guarantee of success," the initial claim about the likely effects of replication cannot be supported by statistical significance alone (Deke & Finucane, 2019). Considered claims about the likelihood of replication depend on careful examination of the dispersion of program effects across sites and studies.

  To address this concern, we recommend updating the protocol to match the definitions ACF provides elsewhere in the Pathways online glossary, which are appropriately more descriptive than predictive. For example, for the Well-Supported evidence category, the glossary definition describes the strength of the supporting evidence without the added claim about the likelihood of replication. Alternatively, if ACF decides to adopt a meta-analysis approach to synthesize evidence—especially a Bayesian approach that addresses limitations of statistical significance, as Mathematica has implemented for Pathways and the What Works Clearinghouse—it could provide plain-language and specific statements about the likelihood of replicating positive effects that are consistent with the analytic approach.

D. Assessing cost study information (Chapter 5 and Appendixes E and F)

The addition of cost study standards is a major advance in the updated protocol and will greatly enhance Pathway's contributions to the field. Cost information is an important determinant when an agency is considering what program or set of services to implement to advance employment-related outcomes for their clients. To make efficient use of public resources, total cost or measures of cost efficiency can be the deciding factor in determining what program or set of programs to implement. To achieve ACF's goal of providing decision makers with reliable, practical information on program cost, the Pathways standards for cost analysis, cost-effectiveness analysis, and cost-benefit analysis must be clear and rigorous to meaningfully and consistently differentiate between estimates of different calibers and to motivate the research field to deliver high-quality cost studies alongside their impact evaluations. We have five recommendations to improve the cost standards and the accuracy, usability, and accessibility of cost information that Pathways shares with decision-makers:

- **Clarify the threshold standards for cost study.** The draft Version 2.0 protocol establishes a set of threshold standards that assess important aspects of the cost study methodology, including those necessary to understand whether the cost estimates are likely to be comprehensive. Only studies meeting the threshold standards are assessed against the quality rating standards, which focus on the transparency and completeness of the cost study report. Given their importance in determining which studies are assessed for quality ratings, we offer four suggestions to make the threshold standards as clear and transparent as possible:

- **Specify how a cost analysis study "demonstrates" the use of a "valid resource-based approach" (Section 5.1, page 33).** If a study must explicitly state that the costs were estimated using a resource cost model or the ingredients method, then very few cost studies will be eligible for review, and Pathways will provide decision makers with cost information for very few programs. Pathways could allow more studies to meet the threshold standard by spelling out what it means to use a "valid resource-based approach" apart from referencing a specific method.
- **Clarify how Pathways reviewers will assess whether a study has accounted for the full set of resources.** Although a resource-based approach provides the best mechanism to capture all program costs, it does not prevent cost omissions or misestimation completely (Cost Analysis Standards Project, 2021). Consider how Pathways reviewers will assess potential omissions. Is it sufficient that the resources or resource categories listed seem to map to the full set of program activities, to the best of the Pathways reviewers' judgement? If so, what processes will Pathways use to maximize, measure, and address issues of inter-assessor agreement? Alternatively, will Pathways defer to the study author's judgement about whether the list of resources is complete?
- **Allow flexibility in the level of detail necessary in reporting cost line items.** Consider what level of detail is necessary in the list of resources, quantities, and prices. We recommend allowing flexibility for cost studies that may have been discouraged from including greater detail in their reports, including due to space constraints. For example, Pathways might accept studies that report the total amount spent on staff. If Pathways requires that studies list the individual positions and salaries, then very few cost studies are likely to pass this criterion.
- **Specify the level of detail necessary in the "description" of the comparison condition and a per-participant cost of this condition for cost-effectiveness and cost-benefit analyses.** We recommend the review protocol specify whether the Pathways reviewers will assess the completeness of the comparison condition cost estimate and how any concerns about its completeness will factor into the cost-effectiveness or cost-benefit study's assessment against the threshold and ratings standards. Consider whether it is sufficient if a study indicates that people in the comparison condition may have participated in other training programs available in the community. Is it sufficient if the study describes one of these programs in greater detail, but not others? How will the Pathways reviewer assess if the description of alternative activities is complete? The draft protocol does not indicate whether costs of the comparison condition must also be developed using a resource-based approach, which could lead to assignment of a moderate or with distinction rating to a cost-effectiveness or cost-benefit study that relies on incomplete costs for the comparison condition.

- **Adjust the quality rating standards to also capture the degree of confidence in the completeness of the cost estimates.** Although the threshold standards are meant to advance only studies Pathways has judged to have estimated program and comparison costs accurately, variation will inevitably remain in the completeness of the cost estimates that pass the threshold standards. As a result, we recommend the quality rating standards include criteria to capture the degree of confidence in the completeness of the cost estimates.

  For example, assuming Pathways establishes a more flexible set of criteria to confirm the use of a resource-based approach in the threshold criteria, it could apply a quality rating standard that captures the level of confidence that the cost study methods yielded accurate and complete costs. This could include a requirement that studies earn 3 of 3 possible points on this criterion to be eligible for a With Distinction rating and at least 2 of 3 possible points to be eligible for a Moderate rating. A best practice for cost analyses is to publish information about sources of uncertainty (Cost Analysis Standards Project, 2021); studies that include this information and that convey the impact of the uncertainty as minor could pass the threshold standard with 3 of 3 possible points. Studies that convey larger levels of uncertainty could pass the threshold with 2 of 3 possible points. Studies that are not transparent about the potential sources of uncertainty but appear to have identified and estimated costs for all ingredients could pass the threshold with only 1 of 3 possible points, meaning they would be eligible for only a Low Quality rating. Pathways may want to revise these criteria depending on the line-item specificity required, how it confirms the use of a resource-based approach, and its plans to report low-rated cost findings or not. Pathways users already associate the Low, Moderate, and High rating scheme used for impact analyses as conveying the level of confidence in the accuracy of the estimates. Factoring confidence in the cost estimates into the quality rating will mean that the cost analysis quality ratings will be more similar to the impact analysis ratings, thereby making them more intuitive to Pathways users.

- **Account for common factors that affect the magnitude and interpretation of the cost estimates.** Not all cost estimates are alike. Factors such as reference point, perspective, and type of analysis can influence cost estimates in important ways that users must understand to avoid misinterpretation. We offer five suggestions to help ensure that users interpret and use the cost estimates correctly:
  - **Only post cost-analysis findings that meet standards with a moderate quality rating or with distinction.** This mirrors the approach to publishing impact evaluation findings and will help users understand the rating earned by a given finding and what that means for the credibility of the cost estimate.
  - **Clearly differentiate between cost estimates, cost-benefit estimates, and cost-effectiveness estimates on the website.** These three types of cost estimates are not interchangeable, and the studies will vary in which they report. Pathways website users must clearly understand the data available to them, so they do not mistakenly compare the cost estimate from one program to the cost-effectiveness estimate from another.

  - **Note the perspective used in calculating the estimates.** Users need to understand who might incur the presented costs so that they can accurately assess the feasibility of their organization, community, or clients to absorb those costs. Knowing the perspective used to calculate costs will also prevent inappropriate comparisons of a participant-level estimate from one program to the societal-level estimate from another.
  - **Be transparent about the outcome domains used in the calculation of cost-effectiveness findings.** For cost-effectiveness findings, studies present costs in reference to impacts on specific outcome measures, such as short-term earnings. Not all studies use the same outcomes for calculating cost effectiveness. Specifying the outcome domains used in the analysis can affect whether and how users should make comparisons across programs.
  - **Inflate all costs to represent a common year and standardize them to represent national costs.** Costs for some resources may vary widely at the local level or over time, making it difficult for decision makers to assess whether differences in cost between programs are due to differences in program ingredients or differences in context. Reporting all costs as national costs for the same reference year will support increased accuracy for users when comparing across programs. If this is not possible, we recommend reporting at least the year and region for which the costs were calculated and presenting them as part of a range rather than a precise estimate, to discourage users from assigning meaning to small differences in cost that may primarily or entirely reflect differences in context.

- **Query authors to ensure accurate rates.** The draft protocol does not indicate that Pathways intends to query authors for clarifying information about the cost analyses. Given the importance of missing information for the threshold standards in particular, we recommend that Pathways query authors when needed to ensure the threshold standards are applied accurately, and that Pathways state the author query plans explicitly in the protocol.
- **Publish findings from cost analyses on the program pages where decision makers can easily retrieve them.** Understanding where and how cost study findings will be incorporated into the Pathways to Work Evidence Clearinghouse website, and therefore how website users might interpret and use this information, is central to designing and considering the suitability of cost analysis standards. Net costs are currently available on some Pathways website study pages, but not on the program pages or program search tools that are of more interest to practitioners. Publishing cost information where decision makers can more easily retrieve and use it not only supports more effective decision making, it also requires that Pathways present the information clearly to support accurate interpretation and use, and that Pathways establish and follow standards that differentiate meaningfully between cost information that is a reasonably accurate representation of the resources necessary to implement the program and cost information that is not accurate.

References

Cost Analysis Standards Project. (2021). *Standards for the economic evaluation of educational and social programs*. American Institutes for Research.

Deke, J., & Finucane, M. (2019). *Moving beyond statistical significance: The BASIE framework for interpreting findings from impact evaluations.* Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Education. https://acf.gov/opre/report/moving-beyond-statistical-significance-basie-bayesian-interpretation-estimates.

Hedges, L., & Vevea, J. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3,* 486–504.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15,* 199–236.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, *171*(2), 481–502.

McKenzie, J. E., & Brennan, S. E. (2024). Synthesizing and presenting findings using other methods. In J. Higgins et al. (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions, version 6.5*. Cochrane. https://training.cochrane.org/handbook.

Mathematica Inc.

Our employee-owners work nationwide and around the world.

Find us at **mathematica.org** and **edi-global.com**.