Assessing Change: Integrating Kirkpatrick's and Phillips' Models into The Five Levels of Evaluation

Steven H. Cady, PhD
Bowling Green State University

Briana Medley, MOD
Bowling Green State University

C. Theodor Stiegler, MOD
NEXUS4change

Dr. Steve Cady serves as a graduate faculty member in the Master of Organization Development program at Bowling Green State University. His focus is on motivation, along with collaborative change and innovation. He is author of The Change Handbook 3rd Edition and Stepping Stones to Success, along with a DVD titled Life Inspired: Six Ways to a Passionate Soul. He received his PhD in Organizational Behavior from FSU, along with an MBA and BSBA in Finance from UCF.

Briana Medley is a Corporate Trainer at BoatU.S. and graduate of Bowling Green State University's Master of Organization Development program. Her professional and scholarly aim is to design impactful learning and development experiences that empower people to reach their maximum potential.

C. Theodor Stiegler holds a masters degree in Organization Development from Bowling Green State University and has worked in advocacy, arts management, trauma informed care, and music. He is currently serving as the executive director for NEXUS4change.

Abstract

The use and scope of evidence-based solutions in organization development and change is on the rise; this trend is bringing renewed focus on ensuring interventions are well designed, followed by an objective evaluation. One of the most unattended functions in organization development and change is the use of good evaluation techniques to examine the impact of planned interventions. Often, the reason for the lack of evaluation is the ambiguous nature of the purpose and outcomes; for it is through an objective appraisal of an intervention that you can reconcile intuitive judgments (i.e. impact evaluation; Khandker, Koolwal, & Samad, 2010). In this article, we examine the history and role of evaluation across disciplines; we then describe an integrated model describing how the five levels of evaluation drawn from Kirkpatrick (1998) and Phillips (1996) can be utilized to determine the efficacy of change interventions. Within this

discussion, we address the paradox of competing demands in conducting evaluations and provide a model for choosing an effective assessment strategy.

Keywords: organization development; performance improvement; consulting intervention; evaluation; resource constraints

---

Contact Information:

Steven H. Cady
Phone: 419-372-9388
Fax: 419-372-6057
Bowling Green State University
3022 Business Administration
Bowling Green, OH 43403

Email: scady@bgsu.edu

Briana Medley
Bowling Green State University
3022 Business Administration
Bowling Green, OH 43403

Email: brianasimmons90@gmail.com

C. Theodor Stiegler
NEXUS4change
10555 Five-Point Rd.
Perrysburg, OH 43551

Email: theo@nexus4change.com

The use and scope of evidence-based solutions in organization development and change is on the rise and this trend is bringing renewed focus on ensuring interventions are well designed then followed by an objective evaluation. Building on the concept of evidence-based management practice (Rynes & Bartunek, 2017), evidence-based organization development and change is defined as the discerning use of knowledge and expertise to guide in the design of interventions. This approach requires using evidence informed by research, case reports, behavioral science principles, and informed opinions. When evidence-based approaches are judiciously used, the organizations and communities undergoing the change benefit from the robustness of the intervention.

What is a robust intervention? A robust intervention is one that has the intended impact in spite of violations to the basic assumptions considered in its design. For example, a change initiative that fails to meet some of the procedural requirements may still turn out to be a huge success (Hatry, Newcomer, & Wholey, 2015). Determining the robustness of an intervention requires a well-constructed evaluation. In order to determine the best evaluation strategy, professionals are faced with a dilemma—whether to prove the impact of the intervention or to improve the intervention for later use (Cady & Milz, 2015; Cady et al., 2010). Newcomer, Hatry, & Wholey (2010) offer that this is the crux of good evaluation, suggesting that it is important to reconcile what to evaluate and how many resources to allocate.

One of the most unattended functions in organization development and change is the use of good evaluation techniques to examine the impact of planned interventions. Often, the reason for the lack of evaluation is the ambiguous nature of the purpose and outcomes—for it is through an objective appraisal of an intervention that you can reconcile intuitive judgments (i.e. impact evaluation; Khandker, Koolwal, & Samad, 2010). Furthermore, a thorough analysis will enable adaptation, improvement, and learning from the evaluation (Balasubramanian et al., 2015) ensuring the intervention will be more robust in the future. The more robust the intervention, the more efficacious it is. The more efficacious the intervention, the more confident one can be in it achieving the desired results.

In this article, we examine the history and role of evaluation across disciplines; we then describe how the five levels of evaluation, drawn from Kirkpatrick (1998) and Phillips (1996), can be utilized to determine the efficacy of organization development and change interventions. Within this discussion, we address the paradox of competing demands in conducting evaluations and provide a

model for choosing an effective assessment strategy.

Approaches to Evaluation: An Interdisciplinary Perspective

An examination of evaluations across disciplines shows that assessment criteria and methods are not always universal or straightforward. In most cases, the criteria are ambiguous and the outcomes are compounded with other factors. Nonetheless, the value of an evaluation cannot be underestimated. Evaluation experts across multiple disciplines have developed a multitude of measurement criteria and methodologies in the fields of organization development, education, social programs, finance and accounting, and training and development.

Education

The development of educational evaluation started as early as the 1840s, when Horace Mann proposed the assessment of institutional effectiveness and teachers' competency as part of educational assessment criteria in addition to the traditional evaluation of student achievements. In the past, the purported use of evaluations was to test a set of hypotheses based on the behavioral objectives set in advance by educators (Alkin & Christie, 2004). Today, educational institutions require the use of evaluations as a way to raise funds from stakeholders or to attract students by showcasing high student achievements (Ortiz & Rubio, 2009).

Organization Development

In the field of organization development the role of evaluation has been emphasized as a core process of "action research" developed by Kurt Lewin. With a relatively short history of evaluations in this field, there are many more issues yet to be resolved (Cady & Caster, 2000). For example, the criteria of evaluations are often unclear, with the ambiguity of the objectives of interventions (e.g. to improve organizational culture and mood). However, even when the objectives are clear, it is often to resolve an urgent problem at hand in an organization, thus preventing a controlled evaluation to measure the net impact (Terpstra, 1981a). In addition, the evaluation outcomes are rarely generalizable because of the limited sample of age, occupation, and status of the participants. All of these elements vary within the organization implementing the intervention (Terpstra, 1981b).

Finance and Accounting

The advancement of impact evaluation has created conditions for evaluation practices in finance and accounting to flourish. Individual firms are utilizing accounting and financial data primarily to evaluate the overall performance or simulate different options. Although these evaluations primarily consist of quantitative measures, those for public finance have expanded to include

individual behavioral responses to financial policies (Pomeranz, 2017). Evaluations in these fields allow for more controlled and rigorous designs and methods at the individual rather than just the organizational level. This is because the impact of the policies on individuals' attitudes and behaviors is indirect.

Training and Development

The evaluation of training and development interventions is important and is a key part of the human resource management techniques. These interventions are a way for human resource management to deliver their benefits to the trainees, trainers, and organizations that support the interventions (Aguinis & Kraiger, 2009). Training differs from education in that it is conducted within an organizational setting to teach practical skills. It also differs from organization development, as it is conducted in a more structured setting with a limited focus on individuals' on-the-job performance. Thus, evaluations in training and development may be more straightforward than in other areas, as their objectives are less ambiguous. However, due to the wide scope within training and development, it may be more difficult to measure the effects from such broad perspectives.

Five Levels of Evaluation: An Integrated Perspective

Despite the differences in the approaches to evaluations across disciplines, it has been consistently suggested that evaluation must capture the actual impacts on individuals, groups, and organizations (Aguinis & Kraiger, 2009)(Khandker, Koolwal, & Samad, 2010) (Pomeranz, 2017). The information gained from an evaluation can help those implementing the intervention determine where improvements can be made (Scriven, 1967; Balasubramanian et al., 2015). Kirkpatrick (1998) and Phillips (1996), from the field of training and development, offer a hierarchical model of data-driven evaluation strategies that incorporates multiple layers of outcomes at both individual and organizational levels. When combined, the five levels of evaluation, Kirkpatrick's Four Levels of Evaluation (1998), and Phillips' ROI (1996) address the following questions.

Level 1: Reaction – how satisfied are the participants with the intervention?

Level 2: Learning – what did the participants get to know through the intervention?

Level 3: Behavior – what are the participants doing as a result of the intervention?

Level 4: Results – what outcomes have been achieved by the intervention?

Level 5: Return – what is the Return on Investment (ROI) of the intervention?

Although originally designed for evaluations in the field of training, the levels of evaluation have been

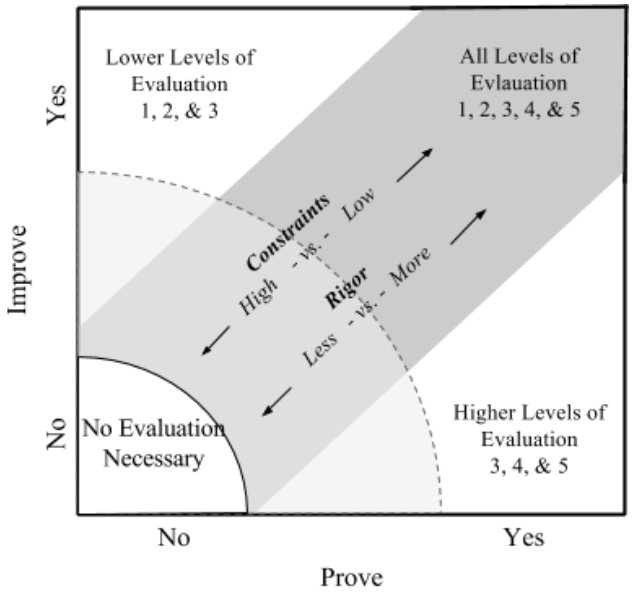
applied to a variety of organizational interventions (Phillips, Phillips, & Zuniga, 2013; Russ-Eft et al., 2008). Because of its comprehensiveness and applicability, we will use this framework as a lens to evaluate organizational change interventions in organizations.

Choosing an Evaluation Strategy

It is often thought that one must design and implement the most rigorous and comprehensive evaluation at all times through conducting all five levels of evaluation. Such an approach is not ideal nor realistic (Cady, Auger, & Foxon, 2010). Obtaining every level of data may not be necessary, much less plausible if one level of evaluation meets the purpose or requirement (see Figure 1). This is particularly true when it comes to the implementability of the designed evaluation: The more comprehensive and rigorous an evaluation, the higher the associated costs (Cady & Kim, 2017).

Change initiatives have budgets that are allocated to design and delivery of an intervention. The question is, how much of the budget does the organization want to spend on evaluation? The problem lies in the fact that a rigorous evaluation often necessitates a greater amount of resource allocation which can potentially decrease the resources that are available for implementation. To resolve this dilemma we suggest that consultants first identify the purpose of an evaluation since this determines the pertinence and utility of that evaluation. Secondly, identify the constraints, and finally adjust the level of rigor that can be achieved depending on the resource availability.

Figure 1. Choosing an Evaluation Strategy



Step 1. Determine Purpose of the Evaluation

The purpose, or the why, of an evaluation can be to prove the effect of an intervention, to improve the existing intervention, or a combination of the two. We frame this purpose definition as yes or no questions in order to demonstrate the decision making process: [A] Do you need to prove the intervention works?—and—[B] Do you need to improve the intervention? While the model allows for some degree of agreement, we suggest starting with the more finite response of "yes" or "no."

The answer to these two questions shows what levels of evaluation are suggested for consideration.

For example, higher levels of evaluations (*levels 3 to 5*) are desirable for assessing the impact on the organization as a whole. On the other hand, lower levels of evaluations (*levels 1 to 3*) provide the more detailed information necessary for improving an initiative. However, if an intervention is a one-time event with a small group of participants, spending additional resources on evaluating the results may not be necessary. In this case, it may be better to conduct a Level 1 evaluation with an open ended question. This enables reactions to be assessed and allows for a basic level of feedback that can be leveraged for future applications.

Step 2. Identify Constraints on the Evaluation

When no absolute criterion exists for assessing resource constraints, one should consider the amount of time and monetary resources that can be used for the evaluation relative to the total amount allotted to the intervention. If the resources do not allow for upholding the highest rigor then an adjustment process may be necessary to consider the appropriate level of diligence.

The tension between proving the intervention works and improving it for future use is referred to as the paradox of competing demands (Cady & Kim, 2017). On one hand, an evaluation must provide useful information to prove the impact of an intervention, or inform improvements of the intervention. On the other hand, the information obtained from an evaluation might lose its utility if it was generated at the expense of the resources—time and/or money—required to implement the improved intervention.

Step 3. Adjust Rigor to Achieve Purpose within Constraints

Once the purpose and constraints of an evaluation have determined the level of rigor, the design of an evaluation can be adjusted depending on the availability of the resources. The higher the need to prove, or improve, an intervention the more desirable a rigorous evaluation becomes. However, because of the increasing nature of evaluation costs the level of rigor may be compromised by the constraint of resources.

Figure 1 illustrates how assessment strategies can be determined depending on resource constraints. The horizontal axis represents the need for proving, while the vertical axis the need for improving. If the need for proving is greater than the need for improving, you can choose from higher levels of evaluations (*levels 3-5*); if the need for improving is greater, lower levels (*levels 1-3*) can be used. If the need for proving and improving is equivalent, a mixed design is preferable. On the other hand the dotted concave curve shows the level of rigor that can be compromised by the resource constraint. The greater the constraints on time or money, the lower the level of rigor that is available

to evaluation methodologies. Due to resource constraints one may choose to conduct all the levels of evaluation with less rigor, which will require fewer resources, yet may still allow the intended purpose to be achieved.

Evaluation Practices and Examples

Table 1 illustrates examples of the five-levels of evaluation with less or more rigor. Although the optimal choice of an evaluation tool can vary with each situation, the integration of suitable technologies may enable an equivalent option at a lower cost (for a comprehensive review, see Materia, et al. 2016).

Higher levels of evaluation have been underutilized for decades (Foxon, 1989; Rosset, 2007) due to the misconception that they must be more costly. For instance, Kennedy, Chyung, Winiecki, and Brinkerhoff (2012) report that Level 3 evaluations are used by 26% of training professionals as a major tool, while Level 4 evaluation by 13% due to limited resource availability, managerial support, and expertise in Level 4 tools. However, higher levels of evaluations do not necessarily require a higher amount of resource allocation if accompanied by the appropriate level of rigor. Caution should be taken when using a less rigorous evaluation approach to infer the effectiveness of an intervention, since the sacrifice of rigor tends to trigger a positive-outcome bias (Terpstra, 1981a).

Concluding Remarks

Without appropriate evaluations, many organizational leaders and executives, as well as their shareholders, will continue to doubt the validity of targeted organizational development and change interventions. How professionals in the field address the paradox of competing demands is important to the future of the field of organization development. Our recommendation is to take a contingency approach. There is no one size fits all, when it comes to evaluating change interventions. The model and steps proposed in this article offers a clear roadmap for choosing an evaluation strategy by determining purpose, constraints, and appropriate rigor that will give credibility to the interventions and leverage its impact for the future. We offer that intentional evaluation strategies allow for advancements in evidence based change practices, even when lower levels of rigor are applied. When appropriately designed, evaluations will be more apt to promote confidence, support, and even praise by those who demand clear evidence of the impact of organizational change interventions.

Table 1

Examples of the Five Levels of Evaluation with Low and High Rigor

Evaluation	Less Rigorous	More Rigorous
Level 1. Reaction	***Quantitative*** - Send a text message prompting individuals to rate their overall satisfaction with an intervention in a short survey; the survey can be created with online services such as Google Forms or LimeSurvey and distributed within a few days after the implementation. ***Qualitative*** - Have a group dialogue during or after an intervention, asking for brief feedback or suggestions for improving the intervention; each intervention leader can take a note from a group conversation and discuss the key points during a separate meeting. Alternatively, emails or text messages can be sent with open-ended questions for feedback/ suggestions.	***Quantitative*** - Distribute a paper-and-pencil or an email survey (using web-based services such as Qualtrics, SurveyMonkey, or PollDaddy) with a series of specific questions about their experience during and/or after an intervention (e.g. effectiveness, relevance, etc.). Conduct a real-time polling by posing a poll question during an intervention (e.g. Poll Everywhere). ***Qualitative*** - Conduct a structured interview or a focus-group interview with selected groups of people (e.g. task force team, the target group of an intervention, etc.) to seek feedback from the whole community; each interview must be recorded and transcribed for thorough analysis. Use a validated software for qualitative data analysis (e.g. Atlas.ti).
Level 2. Learning	***Quantitative*** - Prepare a simple quiz about the intended changes from an intervention (e.g. procedural changes from technology update, policy change, or organizational redesign, etc.); embed the quiz in a text message or an email, as in the Level 1 evaluation. ***Qualitative*** - Have managers ask employees what they know about or learned from an intervention during an informal meeting, take a note, and report it.	***Quantitative*** - Develop a structured test to assess the individual understanding of the planned changes. Ask a series of hypothetical questions to capture the applicability of the planned intervention. ***Qualitative*** - Develop a simulation task to observe the extent to which individuals apply the knowledge/change from an intervention to the task. Observe and track any patterns of behaviors (e.g. a pattern of mistakes) to supplement and improve the existing intervention.
Level 3. Behavior	***Quantitative*** - Develop a simple survey asking individuals the degree to which any behavioral changes have taken place in the workplace - for example, how much a work process has been simplified. Distribute the survey through text messages or emails. ***Qualitative*** - Have key personnel observe whether the intended changes have been applied by individuals, ask people about any influences the intervention has had on their behaviors, and report in a written paragraph.	***Quantitative*** - Identify key on-the-job behaviors that can be changed by the intervention. Observe individual behaviors pre- and post-implementation and rate on a predetermined scale; employ evaluators blind to the purpose of the assessment and average the score across evaluators to minimize any bias. Alternatively, have individuals self-report their behaviors before and after an intervention. Administer the assessments multiple times with regular intervals to compare short-term and long-term effects. ***Qualitative*** - Have a selected group of people write online diaries on a regular basis (e.g. daily, each week or every other week, depending on the intervention time frame). Consider having them write during a scheduled meeting occasionally to ask real-time questions during the diary entry.
Level 4. Results	***Quantitative*** - Using the online survey services outlined above, develop a survey asking individuals about any noticeable performance-related outcomes that can be attributed to the intervention. Distribute the survey a reasonable period of time after the implementation through text messages/emails. Ask survey participants to identify themselves if they wish to give some feedback or tell stories about their experiences. ***Qualitative*** - Interview managers about the overall experience of an intervention and its impact on individual- and organizational-level outcomes, if any. Alternatively, select interviewees based on the survey data for an integrative assessment.	***Quantitative*** - Analyze individual and organizational performance to compare the results before and after implementation; be sure to assess the outcomes at the whole organizational level as well, including those who are not directly affected by the intervention. Account for external factors that may influence outcomes to construct an approximate causal effect of the intervention. Administer the assessments multiple times with time intervals to compare short-term and long-term effects. [See Poister (2015) for the measurement] ***Qualitative*** - Observe and record any incremental changes that are yet to be reflected in numerical outcomes to supplement the above quantitative method. Interview those with extreme changes in their performance after an intervention.
Level 5. ROI	***Quantitative-*** Conduct an abbreviated cost benefit analysis mean for the company. Estimate the cost that were removed/ incurred as a result of the intervention. Also estimate the revenue that was generated and the savings for the company. Calculate an estimated ROI. ***Qualitative-*** Utilize the estimated calculations to guide focus group discussions and interpret the *why* behind the numbers. If the return does not turn out as expected discuss the intervention shortcomings as well as successes and gather themes from the discussions. Generate key talking point themes for learnings and best practices moving forward.	***Quantitative-*** Convert all of the program benefits to an exact dollar amount then utilize the ROI formula. After calculating the ROI run a full cost benefit analysis to show the financial impact of the intervention to the organization. ***Qualitative-*** Hold a meeting with key stakeholders. Through dialogue, estimate the dollar amounts that relate to each of the program benefits by utilizing the ROI formula. Clarify key assumptions and apply them to the calculation. Have each stakeholder conduct their calculation independently, report on the degree of agreement among stakeholder as interrater reliability. The formula will give the estimated return on investment for the intervention.

References

Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual review of psychology,* 60, 451-474.

Alkin, M. C., & Christie, C. A. (2004). An evaluation theory tree. In M. C., Alkin. (Ed.) *Evaluation roots: Tracing theorists' views and influences*, (pp. 12-65). Thousand Oaks, CA: Sage.

Balasubramanian, B. A., Cohen, D. J., Davis, M. M., Gunn, R., Dickinson, L. M., Miller, W. L. & Stange, K. C. (2015). Learning Evaluation: blending quality improvement and implementation research methods to study healthcare innovations. *Implementation Science, 10*(31), 1-11.

Cady, S., Auger, J. & Foxon, M. (2010). Situation evaluation. In W., Rothwell, J. M., Stavros, & R. L., Sullivan (Eds.). *Practicing organization development: A guide for leading change* (3rd ed.) (pp. 363-376). San Francisco, CA: Jossey-Bass.

Cady, S.H., & Caster, M. (2000). A DIET for action research: A problem and appreciative focused approach. *Organization Development Journal, 18*(4), 79 – 93.

Cady, S., & Kim, J. (2017). What we can learn from evaluating OD interventions: The paradox of competing demands. *OD Practitioner, 49*(1), 50-55.

Cady, S., & Milz, S. A. (2015). Evaluating Organizational Transformation. In W., Rothwell, J. M., Stavros, & R. L., Sullivan (Eds.). *Practicing Organization Development: Leading Transformation and Change* (4th ed.) (pp. 195-210). San Francisco, CA: Jossey-Bass.

Foxon, M. (1989). A process approach to the transfer of training. *Australian Journal of Educational Technology, 5*(2), 89-104.

Hatry, H. P., Newcomer, K. E., & Wholey, J. S. (2015). Evaluation challenges, issues, and trends. In K. E. Newcomer, H. P., Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (pp. 668-680). Hoboken, NJ: John Wiley & Sons.

Kennedy, P. E., Chyung, S. Y., Winiecki, D. J., & Brinkerhoff, R. O. (2014). Training professionals' usage and understanding of Kirkpatrick's Level 3 and Level 4 evaluations. *International Journal of Training and Development, 18*(1), 1-21.

Khandker, S. R., Koolwal, G. B., & Samad, H. A. (2010). *Handbook on impact evaluation: quantitative methods and practices.* World Bank Publications.

Kirkpatrick, D. (1998). *Evaluating training*

programs: The four levels. San Francisco, CA: Berrett-Koehler Publishers.

Materia, F. T., Miller, E. A., Runion, M. C., Chesnut, R. P., Irvin, J. B., Richardson, C. B., & Perkins, D. F. (2016). Let's get technical: Enhancing program evaluation through the use and integration of internet and mobile technologies. *Evaluation and program planning,* 56, 31-42.

Newcomer, K. E., Hatry, H. P., & Wholey, J. S. (2010). Planning and designing useful evaluations. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (pp.). Hoboken, NJ: John Wiley & Sons.

Ortiz, M., & Rubio, C. (Eds.). (2009). *Educational Evaluation: 21st Century Issues and Challenges.* Nova Science.

Phillips, J. (1996). *Measuring ROI: The fifth level of evaluation.* Alexandria, VA: American Society for Training and Development.

Phillips, P. P., Phillips, J. J., & Zuniga, L. (2013). *Measuring the Success of Organization Development.* Alexandria, VA: ASTD Press.

Pomeranz, D. (2017). Impact evaluation methods in public economics: A brief introduction to randomized evaluations and comparison with other methods. *Public Finance Review, 45*(1), 10-43.

Reio, T.G., Rocco, T.S., Smith, D.H., & Chang, E. (2017). A critique of Kirkpatrick's evaluation model. *New Horizons in Adult Education and Human Resource Development,* 29, 35-53.

Rossett, A. (2007). Leveling the levels. *Training and Development*, Feb, 49-53.

Russ-Eft, D., Bober, M., de la Taja, I., Foxon, M., & Koszalka, T. (2008). *Evaluator competencies: Standards for the practice of evaluation in organizations.* San Francisco, CA: Jossey-Bass Publishers.

Rynes, S.L. & Bartunek, J.M. (2017). Evidence-based management: Foundations, development, controversies and future. *Annual Review of Organizational Psychology and Organizational Behavior, 4*(1), 235-261.

Scriven, M. S. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of Curriculum Evaluation* (Vol. 1, pp. 39-83). Chicago, IL: Rand McNally.

Terpstra, D. E. (1981a). Relationship between methodological rigor and reported outcomes in organization development evaluation research. *Journal of Applied Psychology, 66*(5), 541.

Terpstra, D. E. (1981b). The organization

development evaluation process: Some problems and proposals. *Human Resource Management, 20*(1), 24-29.

Trank, Christine. (2014). "Reading" evidence-based management: The possibilities of interpretation. *Academy of Management Learning & Education.* 13. 381-395. 10.5465/amle.2013.0244.