Evaluation and Program Planning 58 (2016) 88–97

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Steps towards incorporating heterogeneities into program theory: A case study of a data-driven approach

Sanjeev Sridharan[a,*], Bobby Jones[b], Barry Caudill[c], April Nakaima[d]

[a] The Evaluation Centre for Complex Health Interventions, St. Michael's Hospital and University of Toronto
[b] University of Pittsburgh School of Medicine
[c] Independent Consultant
[d] The Evaluation Centre for Complex Health Interventions, St. Michael's Hospital

ARTICLE INFO

Article history:
Received 7 January 2016
Received in revised form 5 May 2016
Accepted 7 May 2016
Available online 11 May 2016

ABSTRACT

This paper describes a framework that can help refine program theory through data explorations and stakeholder dialogue. The framework incorporates the following steps: a recognition that program implementation might need to be multi-phased for a number of interventions, the need to take stock of program theory, the application of pattern recognition methods to help identify heterogeneous program mechanisms, and stakeholder dialogue to refine the program. As part of the data exploration, a method known as developmental trajectories is implemented to learn about heterogeneous trajectories of outcomes in longitudinal evaluations. This method identifies trajectory clusters and also can estimate different treatment impacts for the various groups. The framework is highlighted with data collected in an evaluation of an alcohol risk-reduction program delivered in a college fraternity setting. The framework discussed in the paper is informed by a realist focus on "what works for whom under what contexts." The utility of the framework in contributing to a dialogue on heterogeneous mechanism and subsequent implementation is described. The connection of the ideas in paper to a 'learning through principled discovery' approach is also described.

1. Introduction

Programs can work through active reasoning of key stakeholders (Pawson, 2006). The idea of how programs work has found expression in the concept of mechanisms. (Pawson & Tilley, 1997) define a mechanism as, " . . . not a variable but an account of the makeup, behavior, and interrelationships of those processes that are responsible for the outcome. A mechanism is thus a theory—a theory that spells out the potential of human resources and reasoning. It is through the notion of program mechanisms that we take the step from 'asking whether a program works to understanding what it is about a program that makes it work' (pp. 408–409). It is important to understand program mechanisms in order to develop a theory of change (Pawson & Tilley, 1997); (Pawson, 2006). These ideas are important because the hypothesized mechanisms by which the program can work are often unclear. In our experience, even detailed applications of theory-driven evaluations often implicitly assume a standardized, homogeneous mechanism: the implicit assumption is that program recipients experience the interventions similarly, have similar needs, and react to the intervention in a similar way. There is very little discussion about the heterogeneous mechanisms by which interventions can work. In our experience, very rarely are theories of change explicit about the " . . . contingencies on which program effectiveness depends . . . " (Mark, Henry, & Julnes, 2000)). This is surprising because one of the purposes of an evaluation is to understand how and why a program might work differently for different individuals Pawson, 2013; (Pawson & Tilley, 1997).

In our experience, evaluating programs in multiple sectors, heterogeneity is ubiquitous; it is more the norm rather than the exception. For example, (Schlattman, 2009) discusses why examining heterogeneity is important in medicine:

> Patients are not alike! This simple truth is often ignored in the analysis of medical data, since most of the time results are presented for the "average" patient. As a result, potential variability between patients is ignored when presenting, e.g., the results of a multiple linear regression model; In medicine there are more and more attempts to individualize therapy; thus, from the author's point of view biostatisticians should

* Corresponding author.
E-mail address: sanjeevsridharan@msn.com (S. Sridharan).

http://dx.doi.org/10.1016/j.evalprogplan.2016.05.002
0149-7189/© 2016 Published by Elsevier Ltd.

support these efforts. Therefore, one of the tasks of the statistician is to identify heterogeneity of patients and, if possible, to explain part of it with known explanatory covariates (p. V)

An interest in examining heterogeneous mechanisms are growing in fields as widely varied as medicine, landscape ecology, child development, and criminology (Nagin & Tremblay, 2005; Pickett & Cadenasso, 1995; Schlattman, 2009). In this light, it is especially intriguing that most evaluations as well as program theories of interventions assume a homogeneous response without explaining the steps programs have taken to address the differential needs for different individuals.

A focus on tailoring programs for heterogeneous sub-populations has implications for evaluation design. (Davidoff, 2009) writes: “Rigorous experimental methods suppress differences among study participants (noise) to detect true intervention effects (signals). But suppressing participants’ heterogeneity obscures an essential dimension of biological and clinical knowledge. Medicine is therefore ambivalent about the influence of heterogeneity on outcomes and struggles to find ways to take it properly into account in both clinical practice and research.” Davidoff’s comments, while made in a medical setting, are also true about other areas like evaluations of education, health care, crime and substance use treatment. Heterogeneous impacts are often assessed in evaluations by doing a sub-group analysis, although in our experience these analyses are frequently not informed by theory (Davidoff, 2009).

1.1. Focus of paper

This paper explores the role of data-driven approaches in understanding heterogeneities associated with programs. Following (Mark et al., 2000), this paper argues that in the absence of clear theoretical knowledge of program’s mechanisms, data needs to be one of multiple sources used to develop the program theory. Using a case study from an alcohol risk reduction program implemented in multiple fraternities in the United States (Caudill et al., 2007; Crosse, Ginexi, & Caudill, 2006), we discuss the role of methods in uncovering heterogeneous patterns in the trajectories of outcomes; this in turn can help refine the program theory through dialogues with program stakeholders.

A statistical method called developmental trajectories (Nagin, 1999); Nagin, 2005; Nagin & Jones, 2001 is described to help identify ‘interesting’ heterogeneous patterns. We argue that the results from the statistical analysis can serve to facilitate a dialogue between program planners, implementers, evaluation methodologists and program theorists in refining the program over time. An important implication of the arguments above is that program theory is often ‘incomplete’ (Sridharan & Nakaima, 2012) especially when faced with the complexities of interventions (Pawson, Greenhalgh, Harvey, & Walshe, 2004) that have the flexibility to evolve and modify over time. Our interest in this paper is to develop a framework to explore if data-driven approaches can help us learn more about the heterogeneous mechanisms by which programs can work.

The focus of this paper is about learning about heterogeneous mechanisms in longitudinal settings. The application of developmental trajectories is only intended to be illustrative. For example, one of the reviewers noted that Qualitative Comparative Analysis as another technique that it can help with complex causation and heterogeneous effects. Our interest is to provoke dialogue on the varieties of methods that can promote clarity on heterogeneous mechanisms; while we have chosen developmental trajectories in this paper, we certainly don’t think this is the only method to understand heterogeneities

We focus on two related claims that we call identifying heterogeneous patterns and the knowledge translational problem (see Fig. 1).

(i) Identifying heterogeneous patterns: The primary argument in this paper is that there is a role for methods to help improve understanding of heterogeneity; the statistical results can help identify heterogeneities in the sample, point towards factors associated with the heterogeneities and also help understand heterogeneous impacts. In this paper, these ideas are demonstrated through an application of the method of developmental trajectories Nagin, 2005: *as noted earlier, this application is only intended to be illustrative; the approach outlined in this paper can be applied using a range of different methodologies* (Mark et al., 2000); (Mark, 2006).
(ii) The knowledge translation problem: There is also a need to pay attention to issues of knowledge translation of heterogeneous mechanisms. Attention to heterogeneous mechanisms for a target population at the outset can help program planners answer: “*whatcan* work for whom?” This question needs to be addressed not at the end of a program but as the program is being implemented, so that the programs better deliver diverse services to meet heterogeneities in individual needs. Therefore, we argue that for some interventions program implementation perhaps needs to be thought of as a multi-phased process in which the results of statistical analysis and stakeholder dialogue inform program development and implementation over time.

The arguments presented in this paper can be illuminated by a consideration of Fig. 2. The top panel of Fig. 2 (scenario 1) describes the standard approach by which evaluations proceed. The ‘standard’ approach assumes that the program is stable and that the program theory is well understood. The objective of evaluations is to estimate ‘average’ program impacts. An alternative view (as seen in the bottom panel of Fig. 2, scenario 2) is that the initial program theory is incomplete; evaluation design and methods can be used to develop estimates of initial program impacts; however an additional focus can also be to learn about areas in which there are uncertainties about program theory including areas of lack of clarity about heterogeneous mechanisms. One methodological focus would be to implement methods that help with initial learnings about the heterogeneity of program mechanisms. Over time such learning can lead to more refined, emergent program theory that incorporates knowledge of program mechanisms.

1.2. Why would this paper be useful?

The focus on heterogeneity is especially relevant in situations where the evaluation has a role to help with the formative/development of the intervention itself. One of the roles that evaluations can play is to build knowledge on how interventions

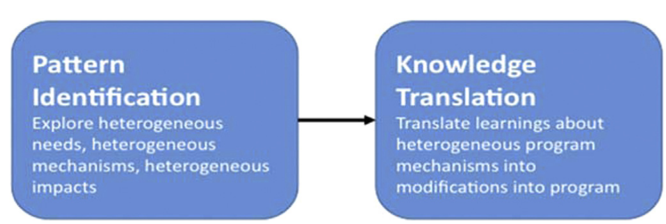

Fig. 1. An illustration of the pattern identification and knowledge translation process.

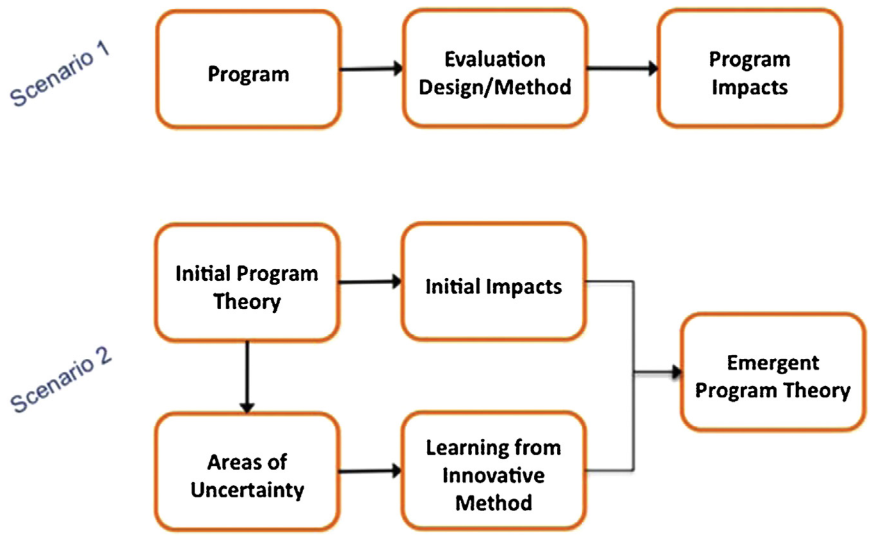

Fig. 2. An illustration of two methods of program evaluation.

can work for very different groups of individuals. Such information can help with 'translating' the intervention to make it more effective across the heterogeneous group of program recipients.

An evaluation can also serve in building dialogue with program implementers on developing knowledge of both the program's context as well as the individual's context that can map both the heterogeneous needs of individuals and the heterogeneous responses that individuals might have to programs. There are different types of knowledge on heterogeneities that an evaluation can generate. For example, in our case study, we demonstrate that there are heterogeneities in longitudinal patterns of alcohol use – this is captured with the idea of 'latent classes' Nagin, 2005; different sets of factors also predict membership into the different 'classes.' Further, a program might 'work' differentially for different groups of individuals; in other words, programs can have heterogeneous impacts. Our claim is that statistical methods can help identify patterns that might point towards possible mechanisms. Moving from statistical patterns to mechanisms is a process of "sense-making" (Julnes & Mark, 1998; Mark et al., 2000) that requires dialogue, understanding of theory and the evidence within the field.

The approach described here will apply for interventions in which the knowledge of heterogeneous mechanisms is limited (in our experience, this would include a large number of interventions); It will also be useful in settings where there is an expectation that an evaluation itself will help with the formative aspects of the intervention. One of the implications of our work is that formative evaluation can be enhanced with a clearer focus on heterogeneous mechanisms.

2. A brief look at the evaluation literature on heterogeneity

The evaluation literature has addressed issues of heterogeneity in at least two ways—the first is to focus on *heterogeneous program impacts* (Djebbari & Smith, 2008; Elbers, Gunning, & de Hoop, 2008). Much of evaluation's focus on heterogeneity in the traditional sense comes from the economic literature where the focus has been on seeing heterogeneity as an estimation problem. The question addressed in this literature is how does one estimate distributional impacts of programs across a range of individuals based on specific characteristics of the individuals or communities? There has been a more limited focus on understanding the knowledge translational implications of such heterogeneity. In other words, how can such programs be tailored to address the differential needs of individuals?

Another attempt to address heterogeneity related to evaluations comes from realist evaluation which argues that evaluators need to pay attention to heterogeneous configurations of context-mechanisms-outcomes in order to properly understand how programs work for different people (Mark et al., 2000; Pawson, 2006; Pawson & Tilley, 1997). Different individuals under different contexts might need different program mechanisms in order for a program to work. *Program design and development need to be informed by knowledge of the heterogeneous mechanisms by which programs can work.*

There is also a rich tradition of evaluation on social justice that we also believe intersects with a focus on heterogeneities. For example, consider Ernie House's (2014, p. 10) focus on social justice as part of a definition of evaluating with validity: "Put simply, my broadening of the concept of validity was based on the idea that if an evaluation is untrue, or incoherent, or unjust, it is invalid. In other words, an evaluation must be true, coherent, and just. All three criteria are necessary...Truth is the attainment of arguments soundly made, beauty is the attainment of coherence well wrought, and justice is the attainment of politics fairly done." A broader view of evaluating with validity that incorporates both justice and beauty also implies a move away from an "average" impacts view of interventions. By incorporating a justice perspective, one argument is that is moves away from an average perspective and pays attention to the 'equity' aspects of evaluation. A justice view ideally should pay attention to *all* individuals in the continuum, not just the average.

In our work we have been strongly influenced by a realist perspective. The implication from a realist perspective is that under heterogeneous contexts, different packages of mechanisms are likely to lead to different outcomes. Yet the challenge we find across multiple evaluations is that the program theory rarely incorporates knowledge of heterogeneous mechanisms. In our experience, program implementers often struggle with the idea of mechanisms, leave aside heterogeneous mechanisms.

Our warrant for viewing programs as evolving dynamic systems comes from multiple sources including the literature on complex interventions (Sterman, 2006; Morell, 2010; Sridharan & Nakaima, 2011). Additionally another source is from the literature on realist evaluation (Pawson, 2013; (Pawson & Tilley, 1997). Realist evaluation is especially useful for evaluations of complex interventions. Four key precepts of interventions from the realist evaluation literature (Pawson et al., 2004) that are especially relevant to this paper are: (i) “The intervention consists of a chain of steps or processes”; (ii) “These chains of steps or processes are often not linear, and involve negotiation and feedback at each stage”; (iii) “Interventions are prone to modification as they are implemented. To attempt to ‘freeze’ the intervention and keep it constant would miss the point, that this process of adaptation and local embedding is an inherent and necessary characteristic”; (iv) “Interventions are open systems and change through learning as stakeholders come to understand them.”

Adaptation, tinkering, changes in responses to learning and contexts are key to program implementation within a realist evaluation worldview. This paper reflects on heterogeneous patterns as part of such adaptation and tinkering.

2.1. Learning through principled discovery

The ideas discussed in this paper find resonance in the ideas of what (Mark et al., 2000) call principled discovery. (Mark, Henry, & Julnes, 1998) posed the question, “How do we ask the data, rather than practitioners or social science theory, to provide the program theory to further guide us?” (Mark et al., 2000) described principled discovery as a method that, “ . . . can allow discovery via induction within the complexities of an open system but that are principled in that the discoveries are subsequently disciplined by data and are not simply post hoc explanations that exploit chance variations in a particular sample.”

Mel Mark has been at the forefront of encouraging a principled discovery approach in evaluation. As an example, consider his course taught at the CDC Summer Evaluation Institute (Mark, 2006) that discussed a variety of methods that can be helpful in taking such a ‘principled discovery’ approach.[1] This paper builds on the methods that describe principled discovery approaches with a particular focus on longitudinal methods while bringing a focus on the role of both methods and stakeholder dialogue in developing refined program theory over time.

While we share a similar interest in understanding such contingencies, our interest is on how such knowledge can help improve the same intervention over time. We stress that even though the focus in this paper is on a specific type of statistical technique called developmental trajectories that can be useful for understanding heterogeneity, our claims in this paper are not confined to developmental trajectories. As noted earlier, our broader claim is there is a need to embed such data-driven processes in learning about program theories over time. We also argue that there is a need to develop dialogues around how learning from such data can help inform what to do about such heterogeneities.

2.2. Organization of paper

The rest of the paper is organized as follows. The next section describes the case study including the program, the developmental trajectories methodology and the results from the statistical modeling. Next we describe our proposal for how such pattern recognition techniques can be integrated with dialogue with program stakeholders to address issues of knowledge translation. Finally, we outline the implications of this approach.

[1] See Mel (Mark, 2006). What Works When: Unraveling How Context Affects Program Effectiveness; CDC Summer Evaluation Institute http://www.eval.org/summerinstitute/06SIHandouts/SI06.Mark.BO12.Final.pdf

3. Case study

In our original evaluation for this intervention was an experimental trial with a focus on summative average level effects. Our original evaluation question was “Does the intervention work on the average?” As we highlight in this paper our attention turned to questions that focused on the heterogeneity of the population of program recipients: What can be done to assure that an intervention serves the differential needs of individuals?

3.1. Developmental trajectories

Biological, attitudinal, behavioral and social processes and behavior tend to change and evolve with time. Psychologists have termed this evolution the “developmental trajectory” Nagin, 2005 of the process or behavior measured over time. In this example, we demonstrate how an application of the developmental trajectories methodology can be part of a methodological process (that includes both statistical methods and stakeholder dialogues) in *exploring* if different mechanisms are operating for drinking behaviors for varied groups of individuals in an evaluation of a program delivered in a college fraternity setting. While the developmental trajectories methodology can be used for confirmatory purposes (Haviland & Nagin, 2005); (Haviland, Jones, & Nagin, 2011); (Haviland, Nagin, & Rosenbaum, 2007), the approach adopted in this paper is exploratory.

The developmental trajectories methodology (Jones, Nagin, & Roeder, 2001); (Nagin, 1999); Nagin, 2005) is used to model the trajectories of drinking. A formal description of developmental trajectories is available in Nagin (2001) and (Jones et al., 2001). Developmental trajectories (Jones and Nagin, 2007); (Jones et al., 2001); (Roeder, Lynch, & Nagin, 1999) identify groups of individuals that follow similar trajectories of a variable (in our illustrative example the variable is a measure of drunkenness) measured repeatedly over time.

3.2. Analytical questions and some modeling details

Key analytical questions that are addressed by this method include determining the number of different classes of trajectories that exist in the data, identifying factors that differentiate the different classes, estimating treatment effects and assessing if the treatment effects vary across different groups (Nagin, 2001; (Jones & Nagin, 2007). The capability to estimate the trajectory of group specific treatment effects also allows researchers to examine the association of the treatment effect estimates with characteristics of trajectory group members. This provides a basis for an enhanced understanding of how treatment effects vary across variables that distinguish trajectory course. We discuss some of the technical details in the following endnote.[2]

The choice of the final set of models is based both on the Bayesian Information Criterion (D’Unger, Land., McCall, & Nagin, 1998); (Jones et al., 2001); (Kass & Raftery, 1995); (Kass & Wasserman, 1995) and on substantive domain knowledge (Nagin, 1999); p.148). Maximum likelihood methods are used for the

[2] A polynomial relationship is used to model the link between time and outcome measure as well as to model a variety of developmental trajectories. The PROC TRAJ software (Jones et al., 2001) allows estimations of up to a fourth order polynomial permitting a variety of different shapes/trajectories to be modeled. The developmental trajectories model implemented in PROC TRAJ can model various distributions of outcomes including Poisson (for count data), logistic (for dichotomous data) and censored normal (for psychometric scales). As the dependent variable in our case study is a count (number of drinking 8+ drinks in the last 28 days) a Poisson distribution was utilized in this paper.

Table 1
Measures in the developmental trajectory model.

Sensation seeking Zuckerman (1979) is a widely used personality scale that measures the extent to which a person displays a need for new and varied experiences through uninhibited behavior including dangerous activities, a non-conventional lifestyle, and a rejection of monotony. The total scale score reflects the percent of positively endorsed items, with higher scores meaning higher levels of sensation seeking. Sensation seeking scores have been highly positively correlated with a number of risky behaviors, including high risk drinking (Hittner & Swickert, 2006).
A modified subscale of Beck et al. (1995) Social Context of Drinking scale was used to measure the extent to which peer acceptance was a primary motive for drinking. Scores on this 5-item measure can range from 1 to 5, with higher scores indicating greater frequency of drinking to obtain peer acceptance and approval.
The desire to obtain peer approval can be an important motivator for drinking, particularly in college contexts and even more so in contexts where group norms are extremely powerful such as the fraternity environment. In a context where heavy drinking is both normative and socially acceptable, those who drink to obtain peer approval may be expected to drink more than those who do not.
Other covariates included in the model were age of the individual at the time of the baseline interview and the year in school measured by the year in college a person is at the time of the baseline study along with each person's GPA at baseline. Self reported grade point averages (GPAs) at baseline ranged from 4 (A) to 1 (D). Where freshmen students had not yet received a semester grade, they were asked to report their grade point average in high school.

estimation of the model parameters. Missing measurements are assumed to be missing at random. With the missing at random assumption, the maximum-likelihood parameter estimation approach used in PROC TRAJ provides unbiased parameter and error estimates.[3]

3.3. Program background

As this case study is intended to highlight the utility of the developmental trajectories method in understanding heterogeneous mechanisms, only brief descriptions of the intervention are provided. Participants were study members of a national evaluation of a social skills (risk reduction) training program (Caudill et al., 2007) involving 98 chapters of a national college fraternity within "mainland" United States.

The intervention consisted of a training session that utilized a standard curriculum emphasizing active involvement of participants throughout the process.

The intervention included the following three components: (1) *Information:* Participants learned basic information about alcohol and its effects on the body, including information about blood alcohol concentration and alcohol absorption rates. Participants also learned about the effects of alcohol on drinkers, the development of behavioral tolerances to alcohol, and ways to intervene with others who drink excessively (e.g., offering food). Video presentations taught participants how to recognize behavioral cues of intoxication and illustrated some potential ways to intervene in high-risk situations;

(i) *Interpersonal Intervention Skills-Training:* This component aimed to teach participants skills for intervening with high-risk alcohol-related behaviors. Students viewed video segments that illustrated varying levels of alcohol risk behaviors as well as illustrated attempts to intervene with an individual who was drinking excessively. In addition, this component included review and discussion of the vignettes, in addition to role playing suggested intervention strategies.

(ii) *Behavioral Rehearsal:* In this component, participants described personally relevant examples of situations where intervention to prevent a friend's alcohol-related risk might be warranted. The group then discussed and role played possible solutions and ways of dealing with the scenario behaviorally. Participants practiced challenging each other about their drinking motives and suggesting alternative methods of achieving a similar objective.

3.4. Research design in the original evaluation

A randomized design was implemented in the original study to examine program impacts. The original evaluation was not informed by a theory-driven evaluation approach. As part of the randomized design, research assessments were conducted at baseline, and again at 6-, 12- and 18-months post-baseline. Research assessments for the original national study (Caudill et al., 2007) were conducted using personalized visits to every chapter, and audio-enhanced computer-assisted self-administered interviews. Assessments were always conducted at least 30 days after summer break (in the fall) and spring break (in the spring) to avoid any potential seasonalities (e.g., dips or spikes in drinking) in student drinking practices that may occur when students were not yet at school (for our fall assessments) or were on spring break (for our spring assessments).

3.5. Measures

While a number of outcome measures were collected as part of the original design, the focus of the developmental trajectories in this paper is on drunkenness over a twelve-month span. Drunkenness was measured by taking the total number of days in the last twenty-eight days on which 8 or more drinks were consumed in one day. This paper's focus on drunkenness is driven by an interest in exploring the program's ability to modify heavy drinking. Given the illustrative focus of the analysis, the analysis was restricted to the first three measurement waves and only included frequent binge drinkers at baseline.[4] Attendance in treatment was modeled as a dichotomous measure.

While a number of measures relating to the individual context of drinking behaviors were collected as part of the national evaluation, the focus in this paper is on defining trajectory classes as a function of five measures: sensation seeking, drinking for peer approval, school year at baseline, age, and baseline grade point average (see Table 1).

4. Results

The results are divided into the following sections: heterogeneous classes of individuals in the data (heterogeneity within the populations), factors that predict such heterogeneity, and heterogeneities in program impacts.

4.1. How many different classes of drinking exist in the sample?

A number of models were examined. Based on the Bayesian Information Criteria (Jones et al., 2001); Nagin, 2001), a solution

[3] The procedure along with online documentation including examples is available from the authors' website free of charge at http://www.andrew.cmu.edu/user/bjones/index.htm.

[4] Frequent binge drinkers were defined as individuals who reported consuming 5 or more drinks on at least 3 occasions in the 2 weeks preceding the baseline interview.

Table 2
Descriptions of the five-cluster solution: average scores of key variables.

	All Subjects	Lowest Level	Moderate Level	High Level Stable	High Level Declining	Highest Level
# at Baseline (n)	1382	204	392	554	102	130
# at 6 Months (n)	1030	155	296	378	91	103
# at 12 Months (n)	804	106	225	293	91	82
# Treated (n (%))	689 (49.8%)	97 (47.5%)	207 (52.8%)	273 (49.3%)	42 (41.2%)	70 (53.8%)
Average Sensation Seeking Score at Baseline (SD)	0.59 (0.13)	0.53 (0.13)	0.56 (0.13)	0.62 (0.12)	0.60 (0.12)	0.66 (0.12)
Average Baseline GPA (SD)	3.20 (0.69)	3.41 (0.58)	3.24 (0.66)	3.07 (0.72)	3.49 (0.58)	3.08 (0.74)
Average Drinking for Peer Approval Score at Baseline (SD)	0.48 (0.96)	0.39 (0.85)	0.40 (0.85)	0.41 (0.85)	0.52 (0.92)	1.07 (1.54)
Age in years (SD)	21.3 (1.5)	21.7 (1.5)	21.2 (1.5)	21.0 (1.6)	21.6 (1.7)	21.6 (1.5)
Years in school (SD)	3.4 (1.1)	3.7 (1.0)	3.6 (1.1)	3.2 (1.1)	3.6 (1.0)	3.6 (1.1)

consisting of five trajectory classes was chosen for those individuals originally classified as frequent binge drinkers. The trajectory classes provide an indication of the heterogeneities in the longitudinal patterns of drunkenness. The first class is those with a *low level of drunkenness* (14.8% of sample); these individuals mean number of drunken days was at the lowest level (relative to other groups) in the first two time periods and the second lowest at the second follow-up.[5] The second class is the *moderate level of drunkenness* group (28.4% of the sample). These are individuals who had a mean number of 8+ drinks for approximately 5 days out of the last 28 days at baseline. There was a slight increase in mean levels of drinking between the baseline and the second follow-up for this class. The third class is defined as a *high level stable drunkenness* group (40.1% of sample). These are individuals whose mean drunken days were close to 9 days out of the last 28 days for all three waves of data collection. Class 4 is the *high level declining drunkenness* group (7.4% of the sample). At baseline, this group drinking was at high level with an average of approximately 9.8 days drunk days out of the last 28 days. This average declined to 1.4 days of the last 28 days by the second follow-up and such a pattern might be suggestive of a pattern on 'regression to the mean' (Shadish, Cook, & Campbell, 2002)). The last class, class 5, is the *highest level drunkenness* class (9.4% of the sample). This group had close to 15 drunken days out of the last 28 days for each of the three data collection waves. Table 2 describes the mean values for the key measures for each of the five trajectory classes. Fig. 3 describes the 'average' trajectory for drinking 8+ drinks across the three periods for both the groups that attended treatment and did not attend treatment. Fig. 4 describes the five trajectory-class solutions and provides empirical support for heterogeneous trajectories in heavy drinking.

4.2. What factors separate the different classes?

We then explored what factors predict membership into the five classes. We explored this both descriptively and by running a multinomial logistic regression model (see Table 3).[6] Key predictors included in the multinomial logistic regression model were sensation seeking, drinking for peer approval and baseline GPA. Class 1 (*low level of drunkenness*) was the reference category in the regression model. None of the included measures were statistically distinguishable between trajectory class 1 and class 2 (*moderate level of drunkenness*; see Table 3). Class 3 (*high level stable drunkenness)* had significantly higher levels of sensation seeking and lower grades compared to class 1. Class 4 (*high level declining drunkenness)* had higher levels of sensation seeking at baseline compared to class 1 but lower levels of sensation seeking than members of class 3 (see Tables 2 and 3). Class 5 (*highest level drunkenness*) had significantly higher levels of sensation seeking and drinking for peer approval as well as lower grades than class 1. Based on the descriptive analysis, it can be seen that class 5 had the highest average of drinking for peer approval and high levels of sensation seeking compared to all other classes (see Table 3). This can be seen more clearly by examining Figs. 5 and 6, which describe the differences in mean levels of sensation seeking and drinking for peer approval respectively across the five classes.

4.3. What are the treatment effects? Does attendance in the treatment make a difference?

Attendance was modeled as a time varying covariate.[7] Two parameters (per group) were modeled to represent the effect at 6 and 12 months. This allows us to see any group rebound or continued reduction in drinking for the intervention subjects. The average number of drunken days was compared between those who attended the treatment and those who did not within each trajectory class for each time point. A *point noted by one of the reviewers was that the technical details required to follow the results would detract from the overall claims of the paper. We do not discuss the results in this section*. We note however that differential impacts of attending the program were obtained across the different classes.

5. Discussion

5.1. Learnings about heterogeneous patterns

The use of the developmental trajectories methodology yielded multiple findings relating to the heterogeneities with the program recipients as well as potential heterogeneous impacts:

(i) There are heterogeneous groups with very different trajectories of drinking behaviors within a single category of drinkers (frequent binge drinkers);
(ii) It is possible to identify factors that predict membership into different heterogeneous groups that may be useful for program planning;
(iii) The intervention has had differential effects on different 'classes of individuals'; this finding may have been hidden if the analysis had assumed the population was homogeneous.

[5] Note that within the definition of frequent binge drinkers – individuals who reported consuming 5 or more drinks on at least 3 occasions in the 2 weeks preceding the baseline interview – it is possible to have individuals whose drinking behaviors was close to 0 days per month less than 8+ drinks per day.

[6] The multinomial logistic regression model is standard output from *Proc Traj*.

[7] There was a modeling choice to be made as to whether the group assignment would be associated with treatment assignment (risk modeling) or if trajectory of drinking behavior would be modified by treatment assignment within trajectory group (time-varying covariate modeling). Both can't be used because the model wouldn't be identified. Both modeling strategies were investigated and the time-varying covariate model proved to provide a better fit to the data.

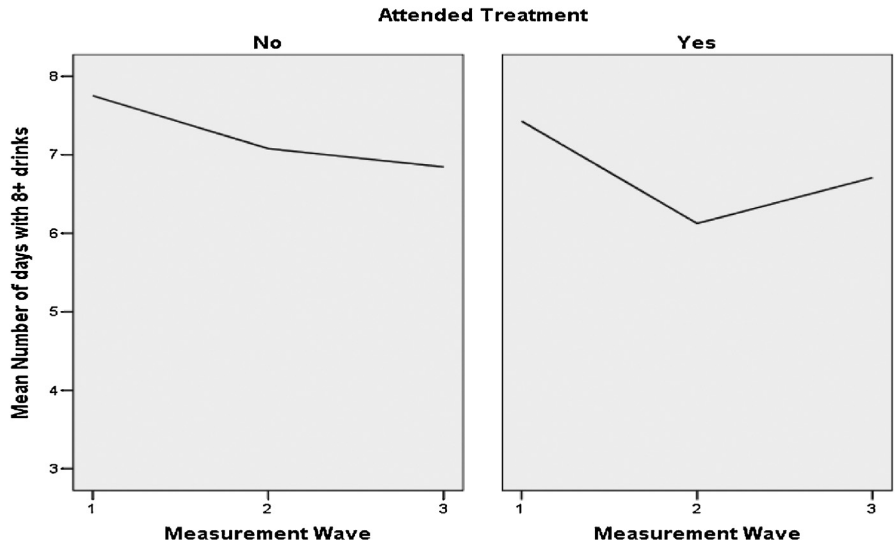

Fig. 3. Average drinking across the three periods for groups that attended treatment and did not attend treatment.

Table 3
Multinomial logistic model predicting membership in the five classes.

	Moderate level		High level stable		High level declining		Highest level	
	Log-odds	P-value	Log-odds	P-value	Log-odds	P-value	Log-odds	P-value
Average Sensation Seeking Score	1.24	0.198	**4.50**	**0.000**	**3.61**	**0.003**	**8.01**	**0.000**
Baseline GPA	−0.31	0.124	**−0.57**	**0.001**	0.19	0.483	**−0.67**	**0.002**
Year in School	0.20	0.295	−0.15	0.306	−0.03	0.874	−0.13	0.486
Age	−0.24	0.067	−0.12	0.227	−0.06	0.675	0.02	0.866
Average Drinking for Peer Approval Score	0.04	0.796	0.08	0.515	0.08	0.618	**0.43**	**0.001**

Bolded values indicate statistical significant at α=0.05.

The above analysis highlights evidence consistent with existence of heterogeneous mixtures within the population.

5.2. Stakeholder dialogues

However, in our view, the above statistical results are by themselves not enough to establish the existence of heterogeneous mechanisms. There is a need to establish how these heterogeneous patterns are useful and actionable so that results are presented in a way that is engaging to stakeholders as well as facilitates action by knowledge users. The above findings should serve to raise a number of focused questions that help stimulate further program development among key stakeholders as well as help to refine program theory over time.

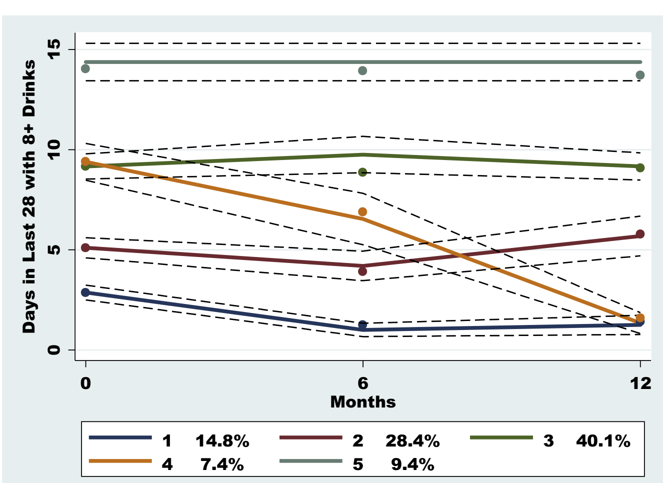

Fig. 4. Estimated heavy drinking trajectories (solid lines), observed group means at each time point (dot symbols), and estimated group percentages. Dashed lines are 95% pointwise confidence intervals on the estimated trajectories.

Statistical results are useful though also incomplete by themselves to suggest changes in programming over time; thus additional steps should be undertaken to improve validity including triangulation of quantitative data with qualitative longitudinal research (Calvey, 2004; Saldana, 2003) and seeking alternative explanations of the observed patterns. Additionally, stakeholder dialogue can also throw light on the local context of implementations (such local contextual measures are often not captured well by 'global' statistical models) (Anselin, Sridharan, & Gholston, 2007).

In our experience, the actual practice of evaluation tends to be siloed: program developers, methodologists and program implementers rarely discuss how learnings about heterogeneities can be utilized to build the program theory as well as to improve the program over time.[8] We highlight the utility of the above findings

[8] A notable exception where this dialogue does take place is in quality improvement (QI) programs that are delivered within hospital or community settings (Provonost, 2011; (Gawande, 2007). The QI programs are designed so that the monitoring and evaluation teams provide feedback to program implementers in almost real time thereby providing information to improve programs in a timely manner.

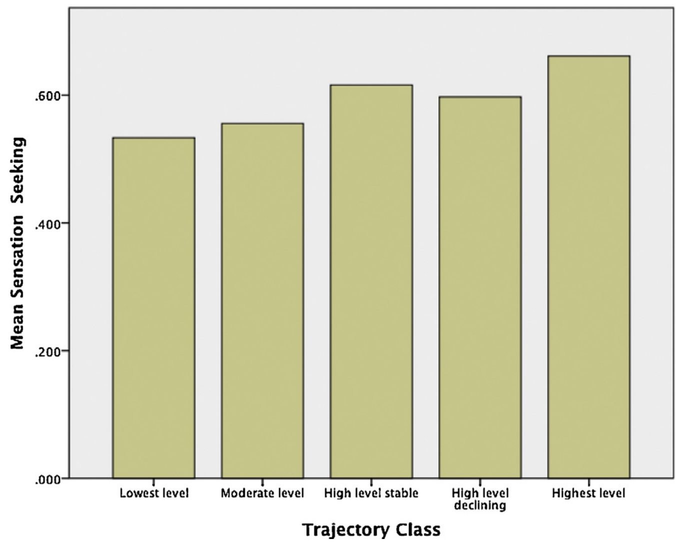

Fig. 5. Mean Sensation Seeking Score for each class of the frequent binge drinkers.

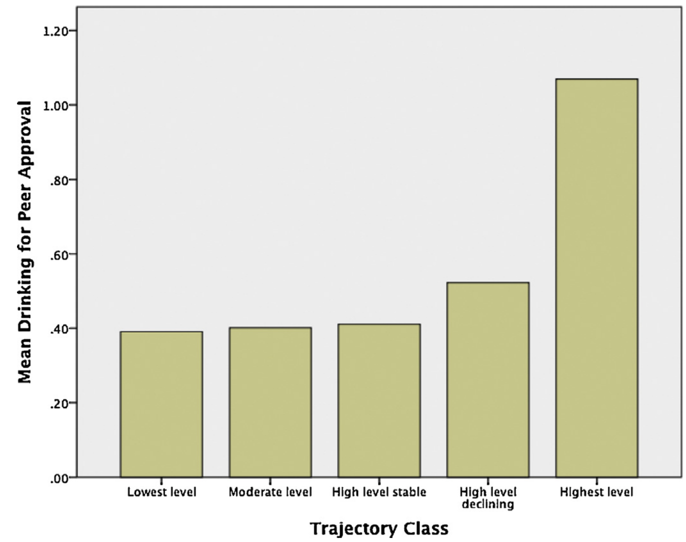

Fig. 6. Mean Drinking for Peer Approval Score for Each Class of the frequent binge drinkers.

by discussing some potential questions for stakeholder dialogue—the questions focus on consistency with experience, enhancing the program, reach of the program, learning from the evidence base and incorporating local context into program adaptations.

(i) *Consistency with experience of program implementers*: The dialogue can also encourage learning from the experiences of program implementers at the level of the "trajectory class" including: Did the pattern of differential program impacts match the expectations of program implementers? Do the risk profiles of the five groups correspond to the program implementers experience with such populations? Do the predictors of the risk groups (e.g. sensation seeking) agree with the practitioner's experience working with this population? Should different sets of program activities be planned to address the needs of individuals in different trajectory classes?

(ii) *Knowledge translation and enhancing the program*: An important focus of the stakeholder dialogue is to reflect on how future versions of the program can be enhanced. Potential questions for improvement include: Should specific program components be added to match the needs of the specific subgroups? Are there (other) ways in which the high-risk individuals (who probably need the program the most) could be identified? Should future versions of the program be localized to specific groups or should it be modified? Given the peer mechanisms involved in improved outcomes, it might be problematic to restrict the intervention to just the high risk group. Is the program doing enough to incorporate knowledge of the heterogeneous groups in its programming? In the practitioner's experience, are different program mechanisms needed to respond to the needs of the trajectory classes?

(iii) *Exploring the 'reach' of the program*: As illustrated by our case study, one of the ways the data analysis can aid dialogue is to get clarity if the program is actually serving the individuals it is meant to serve. Rather than an overall average, the analysis beings greater clarity about *who* is being served through a longitudinal profile of outcomes and in our experience can sometimes potentially serve as a 'healthy surprise' to program implementers. In the context of fraternity drinking behaviors, is the program really appropriate for low-level drinkers?[9] Should inclusion criteria be adjusted to ensure that the programs are reaching the individuals who need the program the most? In our experience, addressing such questions is often not trivial: responding to such questions can help shed light on critical aspects of the program mechanisms.

(iv) *Consistency and learning from the evidence base:* The dialogue needs to be informed *not* just by the results and stakeholder experiences, but also by the evidence base. How do the results match what is already known in the evidence? What is known about the mechanisms by which the program can work (Pawson, 2006)? For example, there is a rich body of evidence from trajectory analysis that can inform and guide such an analysis (Bartholow, Sher, & Krull, 2003); (Schulenberg, O'Malley, Bachman, Wadsworth, & Johnston, 1996); (Tucker, Orlando, & Ellickson, 2003).

(v) *Incorporating local context*: The dialogue can also incorporate features of the local context: Was the context 'complicit' in the observed pattern of results? As an example, we are presently using this methodology in a multisite analysis and exploring if site context is a predictor of trajectory class. An important contribution of the dialogue can be to highlight factors that are presumably part of the context that are not presently being measured.

The above points are only meant to be examples but our central point is that there needs to be structures that promote such dialogue towards action. The goal is to learn from addressing questions and to determine whether the program is doing enough to address the potential differential needs of multiple subpopulations. We think such a dialogue can lead to a system of continuous improvement (Morell, 2001) of programs and also an understanding of program mechanisms.

[9] In the context of this program reaching low-level drinkers is important given the mechanisms that guided the program were (Caudill et al., 2007): 1) knowledge of when an alcohol risk reduction intervention is needed, and 2) intervene effectively with peers to reduce risk in their friends, peers, or in drinkers they may not know, but whom may nevertheless need help.

5.3. Implications

In addition to the above points we see the following conceptual and methodological implications of this paper:

(i) First we hope to provoke discussion around what it means to have a 'good enough' program theory Greenhalgh et al., 2006. In our view, based on the discussions in this paper, a 'good enough' program theory (Sridharan & Nakaima, 2012) needs to address issues of heterogeneity and how the program plans to respond to such heterogeneity.

(ii) As noted earlier, these ideas are especially relevant in situations where the evaluation itself is intended to help develop the intervention over time. One key implication when faced with interventions in which knowledge of heterogeneous mechanisms is absent or incomplete, is that it might be useful to have multiple phases of the implementation of the interventions (see Fig. 7). The first phase would focus on learning about both the context and mechanisms (and the need for heterogeneous mechanisms) and the second phase would be a testing phase in which the revised and refined program that seeks to address the heterogeneous needs of individuals is tested. A subsequent phase can test the revised program theory in a more elaborate way. We note that such an iterative, evolutionary, adaptive stance to evaluation design has been the focus of a literature on sequential and adaptive designs (Chow & Chang, 2008; Chen et al., 2012; however, this literature has not been explicitly connected to the literature on theory of change and does not focus on learning about heterogeneous mechanisms.

5.4. Limitations

A limitation of the paper is that there is a danger that the broader set of ideas proposed in the paper may get lost in the statistical complexities of the developmental trajectory methodology. We reiterate that our interests are broader than a specific statistical technique. Our broader interest is in thinking about analytical frameworks that can help in learning and testing heterogeneous program mechanisms over time. In our own work we have applied other techniques such as qualitative longitudinal research (Saldana, 2003) and exploratory spatial data analysis (Anselin et al., 2007)) to learn about heterogeneous mechanisms.

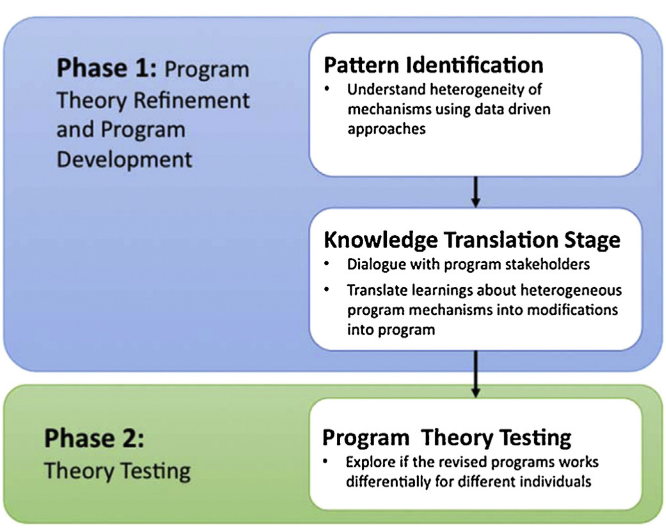

Fig. 7. A framework to learn about heterogeneous mechanisms over time.

We also appreciate that there are a number of interventions that already have well defined program theories with an understanding of heterogeneous mechanisms. We appreciate the methodology discussed in this paper might not apply to such evaluations.

In conducting the evaluation, our original intent was to conduct a summative evaluation. It was only midway through the evaluation that we became interested in the problem of heterogeneous impacts and heterogeneous mechanisms. As this was not originally planned for in the evaluation, we were not able to implement aspects of the stakeholder dialogue. We are however implementing similar types of stakeholder dialogues in other longitudinal evaluations.

We appreciate that many complex interventions are considerably more complex than the example presented in this paper; however its relative simplicity makes the point that we need to more explicitly consider issues of heterogeneity even for interventions that are not very complex.

Some of the other limitations relate to the challenge of the developmental trajectory methodology. Influential methodological critiques of developmental trajectories include (Muthén, 2006) and (Sampson & Laub, 2005). (Sampson & Laub, 2005) critique of developmental trajectories is especially important as it highlights that the search for heterogeneous mechanisms needs to be principled and informed by theory. As noted by (Sampson & Laub, 2005), we need to be careful not to reify methods themselves: "Many of the problems we note in this rejoinder stem from unreflective application and lack of attention to assumptions, a typical scenario when methods diffuse widely." We concur with this point of view.

Methodological concerns with the application of developmental trajectories in our case study might include that we have modeled program attendance and not program assignment. Given the self-selection issues involved in attendance, this paper does not model the selection issues directly. A relatively recent advance in developmental trajectories models such self-selection (Haviland et al., 2011).[10]

5.5. Heterogeneity and methods

The framework presented in this paper can include a number of other methods to understand heterogeneous mechanisms (Mark, 2006). Developmental trajectories are one of the many methods that focus explicitly on heterogeneous patterns. As noted, our interest in this paper is not to promote only one method, but to highlight the need to think more explicitly about heterogeneous mechanisms. Furthermore, as discussed earlier, a focus on developmental trajectories should not preclude other ways of learning about heterogeneous mechanisms including qualitative methods, other quantitative methods or evidence reviews (Pawson, 2006).

The approach described here fits with a realist focus on contexts, mechanisms and outcomes. The realist approach in itself is not a method, but highlights the need to think of a range of methods that can help in identifying and analyzing heterogeneous context-mechanism-outcomes configurations. As described by

[10] Another more advanced methodological critique is that the developmental trajectory approach provides an approximation to continuous heterogeneous differences. As an example, consider Nagin and Tremblay (2001, p. 10): "While there may be populations comprised of groups that are literally distinct, they are not the norm. Most populations are comprised of a collection of individual-level developmental trajectories that are continuously distributed across population members. The statistical question is how best to model the population heterogeneity of individual-level trajectories . . . The group-based approach should ideally be seen as an approximation of a "more complex reality." (Nagin & Tremblay, p. 10).

(Mark et al., 2000), a focus on such principled discovery methods should also be accompanied with other more confirmatory approaches. While our focus of this paper is on exploration, over time as heterogeneous mechanisms are understood and implemented, there may be a need to move towards confirmatory evaluation approaches.

6. Conclusion

Programs are dynamic systems and often change over time as key stakeholders learn more about the mechanisms by which they could work. As an example, consider (Pawson, Wong, & Owen, 2011): "Programs are active, not passive. Interventions do not work in and of themselves; they only have affect through the reasoning and reactions of their recipients." We have argued for a methodological strategy in which lessons learned from pattern recognition methods such as developmental trajectories can facilitate a more fruitful dialogue between program implementers to modify the program. Such modifications need to be based on an understanding of the heterogeneities of recipients 'reasoning and reactions' to interventions. The realist evaluation focus of 'what works for whom' (Pawson & Tilley, 1997) becomes even more potent if the program incorporates knowledge of "what *can* work for whom" in the first place. Applications of data-driven methods can be part of a dynamic learning process that can help one to learn about heterogeneous mechanisms by which programs work and also enhance the program to incorporate learnings about heterogeneous mechanisms.

References

Anselin, L., Sridharan, S., & Gholston, S. (2007). Using exploratory spatial data analysis to leverage social indicator databases: the discovery of interesting patterns. *Social Indicators*, *82*(2), 287–309.
Bartholow, B. D., Sher, K. J., & Krull, J. L. (2003). Changes in heavy drinking over the third decade of life as a function of collegiate fraternity and sorority involvement: a prospective, multilevel analysis. *Health Psychology*, *22*(6), 616–626.
Beck, K. H., Thombs, D. L., Mahoney, C. A., & Fingar, K. M. (1995). Social context and sensation seeking: gender differences in college student drinking motivations. *The International Journal of the Addictions*, *30*(9), 1101–1115.
Calvey, D. (2004). Longitudinal qualitative research: analyzing change through tim. *Contemporary Sociology*, *33*, 377–378.
Caudill, B. D., Luckey, B., Crosse, S. B., Blane, H. T., Ginexi, E. M., & Campbell, B. (2007). Alcohol risk reduction in a national fraternity: a randomized intervention trial with intent-to-treat analysis. *Journal of Studies on Alcohol and Drugs*, *68*(3), 399–409.
Chow, S.-C., & Chang, M. (2008). Adaptive design methods in clinical trials—a review. *Orphanet Journal of Rare Diseases*, *3*, 11.
Crosse, S. B., Ginexi, E. M., & Caudill, B. D. (2006). Examining the effects of a National alcohol-free fraternity housing policy? *The Journal of Primary Prevention*, *27*(5), 477–495.
D'Unger, A., Land, K., McCall, P., & Nagin, D. (1998). How many latent classes of deliquent/criminal careers? Results from mixed Poisson regression analyses of the London, Philadelphia, and Racine cohorts studies. *American Journal of Sociology*, *103*, 1593–1630.
Davidoff, F. (2009). Heterogeneity is not always noise: lessons from improvement. *Journal of the American Medical Association*, *302*(23), 2580–2586.
Djebbari, H., & Smith, J. A. (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, *145*(1–2), 64–80.
Elbers, C., Gunning, J. W., & de Hoop, K. (2008). Assessing sector-wide programs with statistical impact evaluation: a methodological proposal. *World Development*, *37*(2), 513–520.
Gawande, A. (2007). *Better: a surgeon's notes on performance*. New York: Picador, Henry Holt and Company.
Haviland, A., & Nagin, D. S. (2005). Causal inferences with group-based trajectory models. *Psychometrika*, *70*(3), 557–578.
Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory modeling in an observational study. *Psychological Methods*, *12*, 247–267.
Haviland, A. M., Jones, B. L., & Nagin, D. S. (2011). Group-based trajectory modeling extended to account for nonrandom participant attrition. *Sociological Methods & Research*, *40*(2), 367–390.
Hittner, J. B., & Swickert, R. (2006). Sensation seeking and alcohol use: a meta-analytic review. *Addictive Behaviors*, *31*, 1383–1401.
House, E. R. (2014). Origins of the ideas in evaluating with validity. In J. C. Griffith, & B. Montrosse-Moorhead (Eds.), *Revisiting truth, beauty, and justice: evaluating with validity in the 21st century* (pp. 9–15).142: New Directions for Evaluation.
Jones, B., & Nagin, D. S. (2007). Advances in group-based trajectory modeling and a SAS procedure for estimating them. *Sociological Research and Methods*, *35*, 542–571.
Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods and Research*, *29*(3), 374–393.
Julnes, G., & Mark, M. M. (1998). Evaluation as sensemaking: knowledge construction in a realist world. *New Directions for Evaluation*, *78*, 33–52.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, *90*, 928–934.
Mark, M. M., Henry, G. T., & Julnes, G. (1998). A realist theory of evaluation practice. In G. Henry, G. W. Julnes, & M. M. Mark (Eds.), *Realist evaluation: an emerging theory in support of practice* (pp. 3–32).San Francisco: Jossey Bass.
Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: an integrated framework for understanding, guiding, and improving policies and programs*. San Francisco, CA: Jossey-Bass.
Mark, M. (2006). *What works when: unraveling how context affects program effectiveness*. CDC Summer Evaluation Institute. http://www.eval.org/summerinstitute/06SIHandouts/SI06.Mark.BO12.Final.pdf.
Morell, J. A. ([118_TD$DIFF]2010). *Evaluation in the face of uncertainty: anticipating surprise and responding to the inevitable*. New York: Guilford Press.
Muthén, B. (2006). The potential of growth mixture modelling. *Infant and Child Development*, *15*(6), 623–625.
Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: a group-based method? *Psychological Methods*, *6*(1), 18–34.
Nagin, D. S., & Tremblay, R. E. (2005). Developmental trajectory groups: fact or a useful statistical fiction? *Criminology*, *43*, 873–904.
Nagin, D. S. (1999). Analyzing developing trajectories: a semi-parametric, group-based approach. *Psychological Methods*, *4*(2), 139–157.
Pawson, R., & Tilley, N. ([120_TD$DIFF]1997). *Realistic evaluation*. London: Sage.
Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2004). Realist synthesis: an introduction. *ESRC research methods programme*.
Pawson, R. D., Wong, G., & Owen, L. (2011). Known knowns known unknowns, unknown unknowns: the predicament of evidence-Based policy. *American Journal of Evaluation*, *32*(4), 518–546.
Pawson, R. (2006). *Evidence-based policy: a realist perspective*. London: Sage.
Pickett, S. T. A., & Cadenasso, M. L. (1995). Landscape ecology: spatial heterogeneity in ecological systems. *Science*, *269*, 331–334.
Roeder, K., Lynch, K. G., & Nagin, D. S. (1999). Modeling uncertainty in latent class membership: a case study in criminology. *Journal of the American Statistical Association*, *94*, 766–776.
Saldana, J. (2003). *Longitudinal qualitative research*. New York: Alta Mira Press.
Sampson, R. J., & Laub, J. (2005). Seductions of method: rejoinder to Nagin and Tremblay's developmental trajectory groups: fact or fiction? *Criminology*, *43*(4), 905–914.
Schlattman, P. (2009). *Medical applications of finite mixture models*. New York: Springer.
Schulenberg, J., O'Malley, P. M., Bachman, J. G., Wadsworth, K. N., & Johnston, L. D. (1996). Getting drunk and growing up: trajectories of frequent binge drinking in the transition to young adulthood? *Journal of Studies on Alcohol*, *57*(3), 289–304.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.
Sridharan, S., & Nakaima, A. (2011). Ten steps to making evaluations matter? *Evaluation and Program Planning*, *34*(2), 135–146.
Sridharan, S., & Nakaima, A. (2012). Towards an evidence base of theory-driven evaluations: some questions for proponents of theory-driven evaluation. *Evaluation*, *18*, 378–395.
Sterman, J. D. (2006). Learning from evidence in a complex world. *American Journal of Public Health*, *96*(3), 505–514.
Tucker, J. S., Orlando, M., & Ellickson, P. L. (2003). Patterns and correlates of binge drinking trajectories from early adolescence to young adulthood. *Health Psychology*, *22*, 79–87.
Zuckerman, M. (1979). *Sensation seeking: beyond the optimal level of arousal*. Hillsdale, NJ: Lawrence Erlbaum.