Journal of MultiDisciplinary Evaluation
Volume 12, Issue 27, 2016

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Are All Biases Bad? Collaborative Grounded Theory in Developmental Evaluation of Education Policy

Ross C. Anderson
University of Oregon

Meg Guerreiro
University of Oregon

Joanna Smith
University of Oregon

Background: Using two researchers as independent instruments for interpretation in education policy evaluation, this study applies a collaborative grounded theory approach to qualitative data analysis and theory generation.

Purpose: This study argues that varied perspectives should be a critical component in the methodological and analytical choices of education research, especially when the sought after outcome is deeper understanding of the impact, both positive and negative, of an education program or policy. In this study, rather than using one researcher to confirm the reliability of the other, the study explores the outcome of drawing on the positional reflexivity of two researchers, each with a distinct perspective, as a potential strength to co-generate themes and theory in the evaluation of complex policy or programs.

Setting: The data for this analysis originated from interviews of education leaders (*n* = 13) from two states with contrasting approaches to teacher evaluation: Kentucky and California.

Intervention: NA

Research Design: Qualitative developmental evaluation.

Data Collection and Analysis: Semi-structured interviews and grounded theory coding.

Findings: Results suggest that more robust theory and analysis may result from independent thematic development and converged theory generation when working in a research team, as opposed to early application of inter-rater reliability. The reflexivity of different perspectives was, in part, reflexive of the self—their own biased perspective and prior experience. When merged, their joint interpretation may have unearthed greater dimensionality. Findings from this study can inform future strategies for evaluating qualitative research data, especially within a developmental evaluation approach aimed at understanding system complexity in education policy and practice.

Keywords: *grounded theory; reflexivity; developmental evaluation; education policy; teacher evaluation.*

> Validity is not a property of a test or assessment as such, but rather of the meaning of the test scores. These scores are a function not only of the items or stimulus conditions, but also of the persons responding as well as the context of the assessment (Messick, p. 1, 1996).

As Messick articulated, in quantitative methodology, validity emerges not from statistical results but rather from meaning and interpretation applied by the researcher. In qualitative research, the one explicit role of the researcher is that of *interpreter*. As such, the concept of researcher-as-instrument is applicable across methodologies in educational research and evaluation. Quantitative assessments are instrumental in transforming an individual's orientation within a given construct into meaningful quantitative data. Still, to arrive at valid inference, the researcher plays an instrumental role by scripting items, choosing the measure, and modeling the results with other known variables to explain causality or association. Kane (1992) suggests that achieving validity requires a process of interpretive argumentation that starts with data as a premise and relies on warrants, backing, interpretation, and context within a field. From Kane's perspective, the researcher functions powerfully as an instrument that moves from data to conclusion through a rigorous process of interpretation and disconfirmation of alternative hypotheses. That description lends itself to certain qualitative approaches to evaluation and research in the educational context.

On the other hand, whereas the reliability that a technique will yield the same result in subsequent application addresses some validity concerns in quantitative research (Babbie, 2007), subjective interpretation that varies across contexts and researchers (i.e., instruments) is a hallmark of much qualitative research. By its very nature, education research is a complex sociocultural and political process bound by multiple perspectives each entitled to varying degrees of power. This study argues that varied perspectives should be a critical component in the methodological and analytical choices of education research, especially when the sought after outcome is deeper understanding of the impact, both positive and negative, of an education program or policy. Every perspective adds important detail to shape a complete and valid story, and the analytic process can warp or limit these perspectives depending on the inherent biases of the researcher-as-instrument. In the evaluation of education programs and policies, the perspectives selected and captured, the choice of analytic frame, and the explicit adoption of reflexivity each influence the value and validity of the information curated and the potential contribution to stakeholders.

A common approach in qualitative research and evaluation is to rely on inter-rater reliability and, therefore, force researchers into interpretative agreement. If valid theory generation is a goal, alternative approaches to qualitative data analyses (i.e., grounded theory) may provide richer interpretation and analysis through the use of a researcher-as-instrument approach that protects individual interpretation and biases to run their course in analysis. In this study, we examine the results of approaching researcher-as-instrument biases from an assets-based perspective to achieve meaningful analysis of and theory generation from qualitative data. The contrast between this approach and the paradigm of inter-rater reliability may be an important consideration for research endeavors that seek to evaluate the implications and impact of new policies in education.

Reliability in Qualitative Research and Evaluation

To some, qualitative data are coded to "identify, arrange, and systematize the ideas, concepts, and categories uncovered in the data" (Benaquisto, 2008, p.85). Ways of evaluating the quality of and driving purpose behind coding varies greatly, depending on the researcher's philosophical stance (Mays and Pope, 2000). At one extreme, the relativist position converges with a post-modernist perspective (Vidlich & Lyman, 1994) that challenges the notion of consistency on the grounds that there is no singular objective reality (Miller, 2008; Stenbacka, 2001). Any two accounts may parallel the multiplicities of reality, diverging and converging in significant ways, and still hold some degree of evidence regardless of the claims. The implications of this stance on program and policy evaluation are significant. This antirealist view of qualitative research rejects the idea that consistency should be an aim; to the contrary, this view holds that it is unrealistic and inappropriate to think that insights from data should be the same across individuals (Morse, 1994).

A more moderate view qualifies this extreme degree of reflexivity by upholding the scientific process of qualitative research demanding high quality argumentation and use of evidence (Armstrong et al., 1997), harkening Kane's (1995) perspective on quantitative methodology. When Glaser and Strauss (1967) developed their grounded theory approach, they underscored the hallmark of qualitative research: the special nature of the inductive process where theory remains close to the data and the special reality described by the data. Within grounded theory and other inductive approaches, the implicit quality of reliability emerges from transparency of technique, availability of data to replicate analysis, and a step-by-step explanation of the emergence of theory corresponding directly to the data (Maxwell, 2013; Merriam, 2009).

Leaning towards a more positivist stance, the approach towards subtle realism aims at representing a reality within a certain threshold of consistency, accuracy, and attention to subjectivity and reflexivity of the researcher as instrument (Mays & Pope, 2000). Researchers report different processes to achieve a level of agreement across a pair or group of researchers during the initial coding phase, including debriefing exercises to hash out disagreement (Olesen, Droes, Hatton, Chico, & Schatzman, 1994; Waitzkin, 1991). An approach to coding that requires multiple researchers to review individual interpretations may clarify alternative explanations and temper individuals' views. Yet, triangulation of different researchers can limit the inductive nature of qualitative research, exposing subjectivity, one of its signature qualities, to the same threats of spuriousness common to quantitative research. In education program or policy evaluation, a dominant aim for objectivity and consistency ignores the diverse assets brought by a team of researchers, each with unique biases relevant to the phenomena of interest. Indeed, a technically rigorous but constrained approach may undermine the development of original theory, new insight into critical issues, and connections across different contexts.

At the far end of the spectrum, the positivist paradigm subscribes to the notion "that it is desirable to cleanly differentiate between the perspectives of the informants and those of the researcher" (Davey, Gugiu, & Coryn, 2010, p.145). In other words, are the researcher's findings truly grounded in the data or do they reflect his or her personal ideological biases? This perspective posits that when multiple researchers code portions of the data, inter-rater reliability (IRR) should be used to measure agreement across coders to enhance the reliability across data and the validity of the interpretations drawn (Merriam, 2009; Maxwell, 2013; Patton, 2002). Under this paradigm, IRR also establishes a useful link between qualitative and quantitative fields (Davey, Gugiu, & Coryn, 2010), and to a degree, might be the most obvious complementarity of methodological paradigms (Symonds & Gorard, 2010). Pragmatically, achieving IRR also allows a large quantity of data to be coded and analyzed more efficiently across a research team without sacrificing consistency. Though this technique may achieve a standard of quantity, it may also diminish the quality of novel theoretical or practical contributions. A variety of approaches to calculating agreement across raters have emerged to challenge the commonly used Cohen's kappa (see, Grayson, 2001; Davey, Gugiu & Coryn, 2010). Still, there has been general agreement for nearly four decades among those who support the use of IRR in qualitative research that at least 70-80% agreement across coders shows an acceptable level of IRR (Davey, Gugiu & Coryn, 2010; Landis and Koch, 1977; Nunally, 1978; Schmitt, 1996). Researchers have applied this approach to address evaluation and research questions across a broad range of topics in the education field.

Are all Biases Bad?

Armstrong et al. (1997) explored the positivist versus relativist debate by empirically investigating whether different researchers do arrive at consistent accounts when analyzing the same unfamiliar transcripts. They found that six qualitative researchers arrived at similar themes, but within unique, diverse packaging, concluding that "all analysis is a form of interpretation" and the views of the researcher, inherently, accounted for critical parts of the interpretation (Armstrong et al., 1997, p. 605). Moving beyond an either-or purpose to qualitative research provides space to build the interpretation of data collaboratively without being anchored to a reliability coefficient or discarding the notion of an objective reality. Following the suggestions of Symonds and Gorard (2010), this current study proposes to blend relativist and positivist paradigms; thereby, cultivating a constructivist alternative that may suit education program and policy evaluation questions and engender richer, more meaningful findings and theory generation. In this study, rather than using one researcher to confirm the reliability of the other, the study explores the outcome of drawing on the positional reflexivity of two researchers, each with a distinct perspective, as a potential strength in the evaluation of a hotly debated and contested education policy.

Developmental Evaluation Meets Grounded Theory

If choice of research method depends on the questions posed, then perhaps the relationship between research methods and research purpose is too intertwined to separate into distinct entities. From this angle, Patton (2011) explains that developmental evaluation aims to embrace and unravel the complexity of systems rather than explain complexity away. In doing so, this approach provides a strong fit for social innovators by identifying "emergent processes and outcomes that accompany innovation" and acknowledging that adaptation in dynamic conditions, like education policy, contains critical learning opportunities (Patton, 2011, p. 46). Following a developmental approach to understanding complexity, our application of a grounded theory inductive analysis to understand the development and implementation of a challenging new educational policy cultivates a new methodology moment. Called for by Symonds and Gorard (2010), this type of methodological experimentation with divergent and convergent philosophies helps to move beyond "a tripartite view of research" (p. 1) and test a blending of techniques and perspectives to understand the reality formed around an education policy more deeply. The approach to research adopted in this current study suggests that to avoid the complexity, ultimately, would limit the validity of findings.

Overall, this study seeks to understand how different, openly reflexive interpretations of qualitative data, constrained by a common analytic process, diverge and contrast in ways that either limit or enhance a convergent composite interpretation and theory. Results from the study suggest that individual interpretations can fundamentally contribute to a robust grounded theory approach to data analysis; thereby, utilizing individual perspective, coding, and interpretation to create a complete focus and understanding.

The education program and policy chosen to evaluate within this study (high-stakes teacher evaluation) draws on the unique experience of each researcher described below and stretches our exploration of a collaborative constructivist methodology. The contrasting backgrounds of the two researchers participating in the data analysis—one a former classroom teacher with experience in assessment development and one a former district administrator with experience implementing strategic reform initiatives—offers insight that can inform methodological decisions in education evaluation and research. We believe that the approach we describe may be especially applicable to those who adopt a developmental evaluation approach and seek to reveal the complexity of the systems at work through multiple perspectives.

Methods

The methods detailed below outline the approach taken to both secondary data analysis by the two researchers of interest and the primary data analysis of the researcher's interpretations and theoretical development.

Context of Secondary Data

In recent years, state leaders across the United States have revamped their approach to teacher evaluation in order to receive a waiver from the *No Child Left Behind Act* (Baker, Oluwole, & Green, 2013). The design of some of these systems has been influenced by the federal *Race to the Top* grant competition and, in some states, uses controversial growth modeling of student achievement (Baker et al., 2013). While current evaluation systems differ, many have developed from the research of several experts in the field aimed at the dual purpose of supporting professional development and holding the profession accountable for student achievement (Danielson, 2011). The overarching emphasis of the new systems tends to include significantly more evaluation events for each teacher, more in-depth observation, the inclusion of other performance metrics such as student standardized test scores, and, in some cases, financial rewards and punishment through compensation formulas (Baker et al., 2013). As such, teacher evaluation provides a controversial topic useful for this study's goal of evaluating how two researchers with different backgrounds in the education field approach data coding and analysis.

Sample and setting. The data for this analysis originated from interviews of education leaders (*n* = 13) from two states with contrasting approaches to teacher evaluation: Kentucky (KY) and California (CA). To gather insights about the complexity of the system and from contrasting positions, interviews were conducted with leaders from different levels and positions including school district board members, district superintendents, State Board of Education members, Department of Education personnel, a senate education committee staffer, a non-profit advocacy organization, and a teachers union representative. This purposive sample included stakeholders with opposing perspectives. Snowball sampling (Patton, 2015) was used to solicit interview nominations of state level policy actors and education leaders who supported as well as opposed the state's teacher evaluation policy and of districts that varied along dimensions of performance, geographic location, and student demographics. The interviews ranged from 40-60 minutes and used a semi-structured protocol comprised of hypothetical scenarios with follow-up experience questions about a range of policy areas; the analysis focused on the portion of the interviews related to perceptions and experiences with teacher evaluation. So as to enable the use of grounded theory in the analysis rather than deductively fitting the data to theory already in circulation, the two researchers conducting the analysis did not conduct the interviews (Charmaz, 2014). This aspect provided ample space to draw out the different personal experiences and perspectives of the two coders in their interpretation of the data and to draw conclusions about whether this approach to qualitative coding appears to diminish IRR or strengthen results.

Research Questions

To understand how a collaborative constructivist grounded theory approach affects data coding, analysis, and theory development, this study aimed to answer the following research questions:

1. How do differences in researchers' professional experiences, personal characteristics, and perspective result in different interpretations of the data?
2. How can differences in interpretation be reconciled to arrive at a shared set of evaluative findings and synthesized development of theory?
3. Based on the results, does the blended constructivist paradigm appear to enrich or diminish the credibility of findings?

As noted above, the two researchers in this study brought differing personal and professional backgrounds: one as a female classroom teacher (Researcher A) and one as a male district-level administrator (Researcher B). Due to contrasting positions within education (e.g. teacher versus administrator) as well as gendered roles that can affect data analysis, the two researchers in the study decided on a common, yet individualized approach to evaluation. By evaluating the data individually and using a grounded theory approach, the two researchers would be able to provide an unbiased review and interpretation without the fear of stigma due to positionality. Additionally, the two researchers would feel ready to share, equally, in theory generation, regardless of personality traits (e.g. introversion), experience level (i.e. less experienced being more agreeable), or gender roles (i.e. males showing dominance of opinion).

In an effort to allow these different perspectives to influence the analytic process, the two researchers adhered closely to the coding, memos, and analysis schema endorsed by Charmaz (2014) and the grounded theory approach: begin with basic description, create conceptual ordering, then intuit an explanatory theory. The interviews were coded in Atlas.ti (Friese, 2013) in the same order and both used coding and analysis processes of open, line-by-line coding followed by focused coding and theory development as well as memoing techniques throughout analysis (Charmaz, 2014; Emerson, Fretz, &Shaw, 1995; Strauss & Corbin, 1990) in order to help outline theme identification and theory development.

After each researcher developed a full thematic interpretation of the data alongside an inductive, descriptive theoretical structure, the researchers sought to reveal and contrast their unique reflexivity of the *self* and the *other* through a comparative analysis. Accepting Pillow's (2003) challenge to the field to test out new reflexive practices in qualitative analysis, the approach used within this study sought to press distinct biases and perspectives against one another to find commonalities that might reveal a deeper more valid and complete interpretation and descriptive theoretical position.

Two guiding questions informed the individual approach to data analysis: (a) How do perspectives on teacher evaluation differ between stakeholders at the local school district level and state level? (b) What do these different perspectives reveal about the policy formulation and implementation process? After separate coding and analysis, thematic development and synthesis was conducted individually by each of the two researchers. Then, the two researchers shared their work and collaboratively analyzed the similarities and differences across each process, perspective, theoretical stance, and implications drawn. The two researchers made unified connections between the two separate code lists and themes in order to compare and contrast between individual syntheses.

Results

From the individual analyses, the different backgrounds of the two researchers became evident—one researcher with an emphasis in teaching and assessment and the other with an emphasis in district-level administration and strategic reform initiatives. To identify both common and contrasting interpretations, this section will draw from the independent coding, analysis, and theoretical development that each of the two researchers conducted.

Contrasts in Independent Coding

Each of the two researchers began from the same coding procedures to produce a unique set of codes (see Figure 1) in order to arrive at distinct coding development, theory, and analysis. Researcher A created a positive versus negative coding structure related to teacher evaluation. Many of the codes could have been negatively or positively coded depending on the framework (e.g. discord, control issues, inequality, and time consuming versus goal-setting). This categorization analyzed effect within themes and distinguished between positive, negative, or neutral association with a summary of focused codes within themes. Researcher A isolated some codes as neutral (neither positive or negative) and context dependent neutral (could be positive or negative depending on the context within the interview). For Researcher A, some codes were labeled as one category (i.e., negative); yet, during analysis, recoded as another category (i.e., neutral) dependent upon the interviewee's context. Researcher A generated two themes of power and teacher support; a third unidentified theme categorized context dependent codes, not directly associated with the other two. For example, the code *control issues* could have been negatively labeled with an interviewee who is a district administrator yet neutral with a non-district administrator; the district administrator may provoke the control issues while a non-district administrator may be constrained by the control issues.

In contrast, Researcher B assigned the codes to six thematic coding groups: Emotions-Engagement, Obstacles-Drivers, Partnership, Power, Structural Integrity, and Unit of Importance. From these themes, Researcher B used networking functions in Atlas.ti (Friese, 2013) to map codes into a hierarchy of associations, beginning with the overarching theoretical codes and working down through conditioning and descriptive codes. Mapping the codes allowed Researcher B to draw associations (e.g., negative, positive, dependent, etc.) between codes. This map helped to identify patterns, distinguish a hierarchy, generate theory, and test the theory against different cases in the evolving analysis.

Both researchers' coding structures identified distinct perspectives between states (CA and KY) as well as between the level and sphere of influence of the interviewee (local, internal state, or external state). Researcher A approached these differences from an angle of views and attitudes while Researcher B analyzed differences from an angle of individuals and systems, functionality and leverage. The two researchers drew on their individual experiences and positions to guide interpretations.

Throughout the coding process, Researcher A focused primarily on identifying expression of views and attitudes towards teacher evaluation. This framework supported the development of codes and connections between coding themes. Additionally, the positive-negative-neutral categorization helped contextualize the issues of power and teacher support within the framework of provision or opposition to teacher evaluation across levels. The interpretation and process of coding for Researcher A directly linked to her personal background as a teacher, especially within the themes of power and teacher support. Researcher A primarily generated focused codes highlighting the issues of power (claims of self-group greatness) and teacher support within the context of assessment types, school needs, student needs, data driven decisions, and specific levels of quality within the context of teaching. Researcher A noted that few transcripts focused on needs of schools-teachers as a focus of teacher evaluation with only one transcript mentioning student support. This coding signifies greater emphasis on ensuring that teacher evaluation meets the needs of student, school, and teacher appropriately.

In contrast, Researcher B focused on generating specific themes from focused codes to outline and work within an overall theoretical structure. By breaking down the two themes that were common across researchers—power and support—into specific areas such as structural integrity, emotions and engagement, obstacles, and drivers, Researcher B linked codes across common themes. Illustrated in Figure 1, this process identified focused codes within themes in order to create a more specific analysis and interpretation.

Researcher A		*Researcher B*	
Themes	Codes	Codes	Themes
Power	Claims of group greatness (positive v. negative)	Leadership levers	**Power**
	Claims of self greatness (positive v. negative)	Revealing advantage	
Negative vs. Positive Self vs. Non-self	Claims of greatness of ideas	Tightening reins	
	Control issues	Public perceptions	
	Reluctance to admit non-self greatness	Negotiated order	**Structural Integrity**
	Involvement by self/group	Institutionalized ego	
	Power	Getting rid of dead wood	
Support of Teachers	Collaboration	Complying	
	Competency of leaders	Building capacity	
	High quality of teachers	Purposeful partnering	**Partnership**
	Multiple modes of assessment	Setting the stage	
	Needs of schools	Recognizing competence	
Positive vs. Negative	Research-based	Sense-making	**Emotions & Engagement**
	Support of students	Hoping	
	Use of data	Frustration	
	Visionary	Hesitating trust	
	Dead wood	Rationalizing	
	Incompetency of leaders	Choosing loyalty	
	Poor quality of teachers	Barriers	**Obstacles**
Context Dependent	Compensation	Loose structure	
	Opposition	Opposing views	
	Time Consuming	Somebody keeping score	**Drivers**
	Transparency	Buy-in	

Figure 1. Themes and codes developed by each researcher

For Researcher B, the focus on political process and systems included analysis between themes to identify specific engendering of the policy development and implementation. The themes provided a coding structure that identified and spanned across administrative practices of policy production, collaboration between stakeholders, leadership, engagement, along with barriers and catalysts. Additionally, this specific categorization of codes helped to classify and qualify the facets of administrative discourse and policy implementation through themes focused on policy development, structural practices, administrative involvement, and collaboration. The interpretation and process of coding for Researcher B directly linked to his personal background as a district-level administrator, especially within the themes of policy development, leverage, and administrative leadership strategies for implementation.

Researcher A focused less on the development and implementation of teacher evaluation as a new policy or on the systemic context of power, perceptions, obstacles, and drivers. Rather, Researcher A's focus linked to personal background as a classroom teacher, identifying codes such as assessment types, student and school needs, data driven decisions, and definitions of quality. In contrast, Researcher B's thematic codes focused less on the impact of policy on teachers and students and, instead, targeted different dimensions of decision making, such as: revealing advantages, public perceptions, sense making, hesitation of trust, choosing loyalty, and loose structures. These themes emphasized policy development and implementation with a focus on production, collaboration, leadership, engagement, barriers and catalysts. Notably, Researcher B did not highlight assessment fidelity, needs of schools, or needs of students. In this way, one researcher filled the gaps left in the independent analysis of the other.

Thematic Development

The central themes threading across both independent analyses included power, teacher support, and partnership. In the draft of findings, Researcher A identified distinctions in the theme of power that existed between the member levels:

> The theme of power was evident throughout all transcripts; yet, represented very differently. Board members tended to portray power through control issues that represented self-greatness, specifically within California interviewees. District employees (superintendent, finance officers) portrayed power but did not focus on self-greatness; rather, they focused on power issues around involvement and control. State level interviewees (union leader, state board of education) also varied in their portrayal of power but tended to mention it minimally and include a more positive association. These analyses of power also correlated with the overall tone of the interviews as well as the numbers of negative codes; internal board members being most negative, district administrative employees being more neutral, external interviewees being most positive.

Researcher B highlighted the issue of power within the codes of leadership levers, revealing advantage, tightening reins, and public perception. In his draft of findings, Researcher B found that teacher evaluation was causing a power shift at different levels:

> One state level member in KY noticed that the specifics of the new teacher evaluation system may differentially 'reward' teachers who are 'progressive thinkers' compared to the 'old-school teachers that just want to stay isolated and work on their own.' A KY state leader noted that higher expectations of collaboration and professional growth, similar to the new standards also expected of students, challenged past ways of doing things. If this impression holds true it challenges the historically empowered veteran teachers who have more collective bargaining power, are protected by tenure, and have worked under the assumptions that steps on the pay scale follow sequentially with years of service.

These examples show a clear reflexive link between the different professional experiences and the different interpretations of each researcher when not using a common coding handbook. Themes developed around the *dis*empowerment that teachers absorb when administrators implement a new policy and the different perceptions held by administrative actors. Researcher A found that:

> CA district board members tended to portray teacher support with an association to claims of self-greatness and questions of quality. There was minimal, if any, support for teachers articulated without a negative association, with one

> interviewee referencing 'dead wood' in relation to poor teacher quality.... District administrators portrayed teacher support in a more positive light in comparison with board members. None of the district employee transcripts had negative references in regards to teacher quality. District employees had other positive codes focusing on needs of schools, multiple modes of assessment, and research-based practices. All of which were coded in support of teachers.

In contrast, Researcher B's interpretations explored the facets of institutional and policy drivers within the context of teacher evaluation—a highly politicized issue at state and district levels:

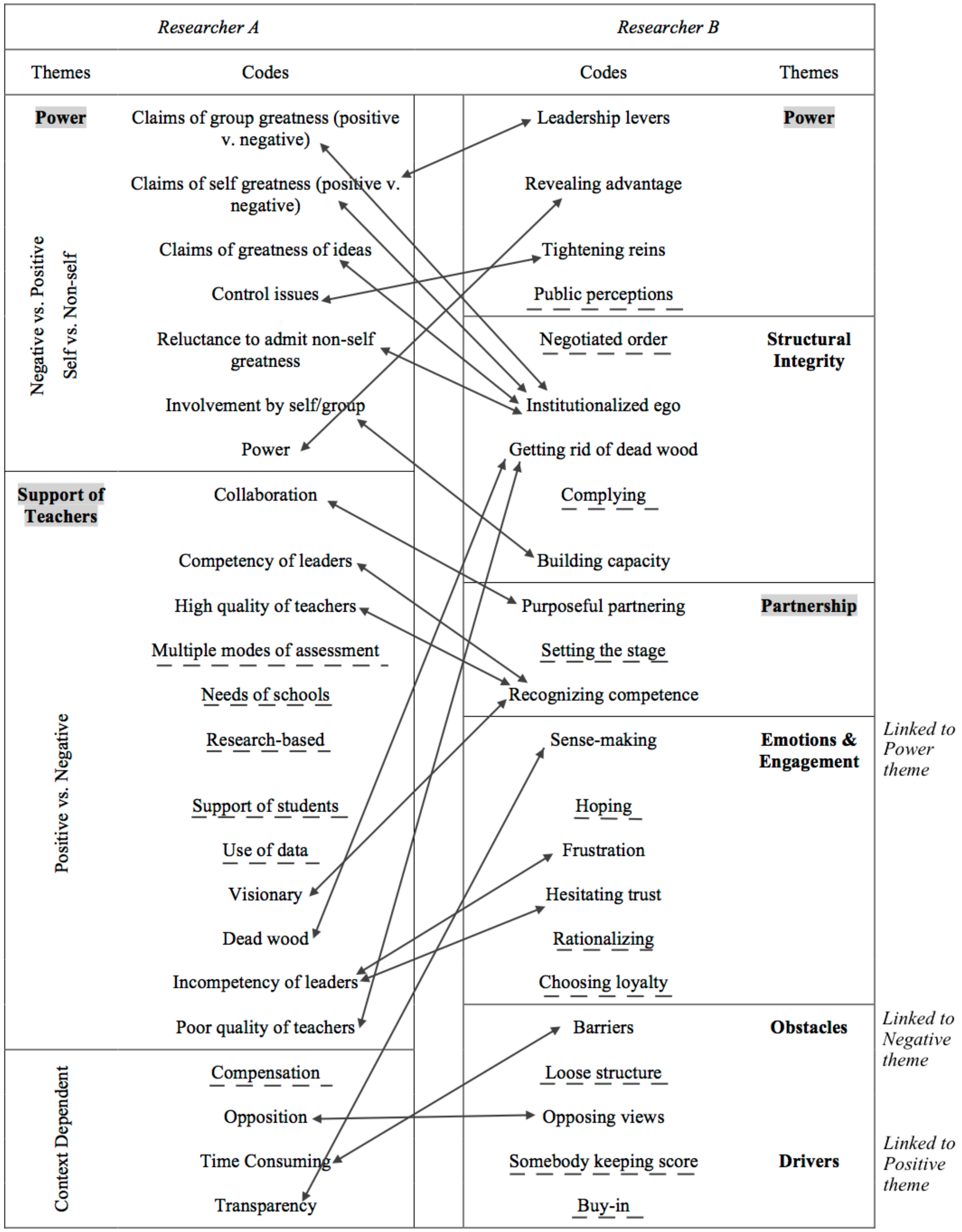
> Expressed by members at each system level, the evolving systems' architecture appears to have two dimensions. The first is seen as the externalized skin of the new and old policies converging into new and improved practice, voiced by members at each level. As one CA district board member mentions, 'in a perfect world' he would see each level of the system be evaluated by the same goals in a trickle down effect. Articulated through institutionalized ego, this 'perfect world' dimension shields the ongoing, messy structural renovations that attempt to fit new policy and practice into older systems. The second dimension is far less visible and only hinted at through the external skin. This negotiated order appears more clearly upon closer examination of the inner workings of policy structures from the state level down to individual classrooms. This dimension is mostly hidden from the public eye and only suggestive in the responses of members across levels.

Researcher B's focus links to personal background as a district-level strategist; codes such as negotiated order, setting the stage, and buy-in may signify the important factors of policy realization accounting for all stakeholder involvement and participation. Additional codes focusing on obstacles and drivers may also provide supplementary context for what is working, what is not working, as well as reasons for individual application. Specifically, the identification of codes around structural integrity, power, and partnership may indicate a level of power and focus on the structures in place and levels of involvement among participants.

Power and Partnership

Despite the two researchers' contrasting backgrounds, different coding and unique thematic development outlined above, similar themes were developed independently and uniquely within the individual analyses. Figure 2 outlines the comparisons and differences between researchers. Across all 47 codes created by Researchers A and B, 68% linked substantively. Both researchers focused on themes relating to power and partnership, which contributed to 83% of code categorization for Researcher A and 48% for Researcher B. This common theme represents a strong overlap; especially given the fact that Researcher B developed many additional themes. This similarity suggests that both researchers identified the importance of power, distribution of power, and effects of power within the context of educational policy and teacher evaluation. Many of the additional focused codes developed by Researcher B (e.g., institutionalized ego, emotions and engagement, and revealing advantage) relate directly to the theme of power as defined by the focused codes of Researcher A. Figure 2 synthesizes the comparative analysis of cross-referencing the codes and thematic grouping across researchers.

Partnership and teacher support demonstrates a similar parallel thread across the two analyses by illustrating values that emerged from the data for both researchers. According to Researcher A's definition of teacher support, many focused codes found in both researchers' coding sets helped to connect the theme of partnership by categorizing within various themes such as getting rid of dead wood (identified as a form of negative support), recognizing competence, and purposeful partnership. The shared importance of this theme in each researcher's thematic development was clear. Overall, the development of codes and themes aligned between researchers; yet, were complemented by unique distinctions. Many focused codes overlapped between themes within each set of codes and connected to themes between sets. Specifically, theses commonalities emerged from common themes such as self-ego (power), control (power), levers/capacity building (power), collaboration (partnership), "dead wood" (negative support), oppositions (negative coding), barriers (negative coding), and drivers (positive coding). Complementary distinctions from Researcher A included needs of schools, support of students, use of data, transparency, and multiple modes of assessment.



Researcher A		*Researcher B*		
Themes	Codes	Codes	Themes	
Power	Claims of group greatness (positive v. negative)	Leadership levers	**Power**	
	Claims of self greatness (positive v. negative)	Revealing advantage		
Negative vs. Positive Self vs. Non-self	Claims of greatness of ideas	Tightening reins		
	Control issues	Public perceptions		
	Reluctance to admit non-self greatness	Negotiated order	**Structural Integrity**	
	Involvement by self/group	Institutionalized ego		
	Power	Getting rid of dead wood		
Support of Teachers	Collaboration	Complying		
	Competency of leaders	Building capacity		
	High quality of teachers	Purposeful partnering	**Partnership**	
	Multiple modes of assessment	Setting the stage		
	Needs of schools	Recognizing competence		
Positive vs. Negative	Research-based	Sense-making	**Emotions & Engagement**	*Linked to Power theme*
	Support of students	Hoping		
	Use of data	Frustration		
	Visionary	Hesitating trust		
	Dead wood	Rationalizing		
	Incompetency of leaders	Choosing loyalty		
	Poor quality of teachers	Barriers	**Obstacles**	*Linked to Negative theme*
Context Dependent	Compensation	Loose structure		
	Opposition	Opposing views		
	Time Consuming	Somebody keeping score	**Drivers**	*Linked to Positive theme*
	Transparency	Buy-in		

Figure 2. Connections and distinctions made between the themes and codes of each researcher

Notes: Similarities in themes highlighted in grey inside figure. Connections between codes are identified by arrows; distinctions are underlined with dashed lines.

Table 1
Main Theories Developed by Each Researcher

Overall theory	Researcher A theory	Researcher B theory
Support for teachers within teacher evaluation	Support for teachers, school, and students should be made a priority when developing new (or updating existing) policies.	Purposeful partnering depends on leadership levers to set the stage for trust and ownership that are pivotal to make policy implementation meaningful.
Importance of teacher evaluation	Teacher evaluation is critical in training, assessment, and teaching and learning of students; therefore, the importance placed on the implementation and development of teacher evaluation should be made a priority.	Teacher evaluation is a lever of one critical element to an intact system's structural integrity. It is used as both a lever of professionalization for improvement and as a publicly visible demonstration of accountability.
Power distribution within teacher evaluation	When accounting for power, it is important to identify strengths and weaknesses among participants to avoid an unbalanced system.	Structural integrity depends on purposeful partnering to maintain a balanced fulcrum of power as well as a shared understanding and investment among stakeholders.

Complementary distinctions from Researcher B included revealing advantage, public perceptions, hesitation of trust, sense-making, loose structure, keeping score, and buy-in.

Synthesized Theory

The two independent theories induced from the same data easily converge into a fused synthesis. The following excerpts (annotated in Table 1) illustrate the separate, distinctive theories developed by each of the two researchers:

> (Researcher A) Teacher evaluation gets framed in the context of best practices and quality of teaching, increased learning opportunities for students, and collaboration between staff members; however, roles and responsibilities are skewed by different perspectives.... Although state-to-state differences exist in evaluation processes, understanding of the system differs most by position and perspective. The portrayal of support for teachers differs depending on position and perspective.

Researcher B identified a lack of shared investment as a driver of opposition to policy implementation and public perception of the teacher profession as a pillar of education's structural integrity.

> (Researcher B) Teacher evaluation is a lever of one critical element to an intact system's structural integrity: the continued professionalization of teaching and a positive public perception of education.... Structural integrity depends on...a balanced fulcrum of power.... When a lack of shared investment couples with insufficient leadership levers and exaggerated institutionalized ego, progress moves slowly, either from inaction or oppositional resistance. The drive towards structural integrity undergirds the rationale of new teacher evaluation systems...

As a convergent whole, the composite theory regards partnerships and support within the context of teacher evaluation critical to developing trust, identifying collaborative opportunities, and articulating levels of support. Teacher evaluation is a critical component in a professionalized education system and should include practices of fidelity and shared ownership. Across the different levels of the system, lack of consensus and collaboration produces greater division. Successful design and implementation of new teacher evaluation systems relies on a balance of power, optimized processes, and thoughtful implementation through strategic partnership. By identifying the levels of involvement, strengths, and weaknesses among stakeholders, leaders anticipate and prepare for the potential drivers and opposition expected along the way. The outline of each researcher's theory, as well as the overall theory developed between researchers, is displayed in Table 1.

Discussion

Using a methodical, inductive process, this study aimed to blend features of positivist and relativist approaches to ascribe validity to qualitative research and evaluation findings. This approach aimed to capitalize on researcher-as-instrument reflexivity to spark a new methodological moment and draw on collaborative grounded theory processes to evaluate education policy under development in a highly complex system—public education. The study used a two-stage process of qualitative data coding and analysis (separate then collaborative). The independent interpretation and construct development linked closely with the experiences in practice and scholarly work of each researcher. Notably, the analytic process included the questioning of whether and how differences in the construction of theory linked to dimensions of power hierarchies within the researcher partnership (e.g., teacher and administrator, male and female). By not adhering a priori to IRR, each researcher-as-instrument conducted a unique, in-depth analysis of the data, each outwardly reflexive of the *self* and of the *other*, in this case state and local policymakers, district administrators, and teacher advocates. By joining analyses after the initial independent stage, a more positivist paradigm emerged in which consistency between researchers was evident and produced a more valid, descriptive result. The evidence of a single robust synthesized theory through separate but parallel interpretations would not have occurred following only a divergent relativist or convergent positivist approach (e.g., adhering strictly to IRR).

The independent analysis and theoretical development shaped by two unique perspectives and constructed within different but converging theoretical frameworks mirrors Armstrong et al.'s (1997) finding that different researchers' "inherent subjectivity" was tempered by "a concordance at a level of situating themes within a wider framework" (p.605). The call from Mays and Pope (2000) for a high standard of rigor in qualitative research should not result in a singular target of 80% IRR between researchers. Rather, rigor should be complemented by Pillow's (2003) call for open "reflexivities of discomfort" (p. 187) and new alternatives to a tripartite of methodologies (Symonds & Gorard, 2010). Indeed, the interpretation of both researchers reflected personal background, perspective, and differences in positionality and power; yet, the shared correspondence to the data was clear in the converging analysis. With distinct individual backgrounds, their reflexivity of the *other* (teacher or administrator) was, in part, reflexive of the *self*—their own biased perspective and experience. When merged, their joint interpretation enlivened more dimensions.

Similarities and distinctions suggest that, despite independent code development and thematic categorization, two analyses from disparate perspectives yields one balanced, and more complete, common analysis. Yet, the choice of researchers is critical to achieve this balance. Moreover, it is important to note that through the development of a dual-researcher coding process, a more robust development of thematic awareness and more specific focused coding helped to portray a representation that may more soundly correspond to the complexities of teacher evaluation in policy and practice. As a result, without using IRR to code for similar themes and focused codes, researchers were free to integrate background and experience reflexively to identify a body of common predominant themes nested within unique perspectives.

Unbound by artificial parameters of quantitative reliability metrics, the researchers-as-instruments covered similar ground in coding and analysis. Findings suggest that the separate analyses may have developed independent interpretations into more nuanced theory than would have been developed if adhering to an IRR threshold during the coding, thematic development, and/or analysis. Driven by the subjective perspectives, experiences, and interests of each researcher, complementary theme development and coding drew connections unseen by the other researcher. Constricted by agreement thresholds of inter-rater coding and theme development, the alternative explanations and theoretical positions co-developed would have been lost and may have resulted in a less robust convergent interpretation. Findings suggest that this collaborative grounded theory approach suits the demands of a developmental evaluation context seeking to understand a complex system at work.

Theory development played an important role between researchers through the grounded theory approach, coding, theme development, and analysis. In future application of this approach, researchers might include the theory developed by researchers in the past tackling the same phenomena in a different context as a different form of *collaboration*—identified a priori but applied through the ad hoc comparison. The two researchers approached the coding process and theme development through different lenses (see Figure 1) with Researcher A developing a preponderance of focused codes of teacher support and power and Researcher B creating focused codes around policy development and implementation through facets of partnership and power. Still, common threads naturally emerged in the cross-researcher analysis. In fact, the common focus of power and partnership, within the context of teacher evaluation, took on more depth and dimensionality when the two analyses were layered together. The two researchers wove overlapping themes into different theoretical frames that clearly reflected the context of different backgrounds and experiences; yet, the overarching themes and insights between researchers followed a parallel process and analysis (as indicated in Table 1).

Overall, limitations of this study include the use of a small sample of participant interviews and the comparison of only two participating researchers. Though interviews were not randomly ordered, the order was consistent across researchers. As a grounded theory approach, it is important to acknowledge that the order of interviews coded could determine the development of focused and thematic codes. Future replications of this approach might consider the impact of a structured or randomized sequence for coding interviews.

Reliability Versus Complexity

Results from this study indicate that adherence to IRR at the coding and analysis stage may cause important interpretations and insights to go undetected by the unique perspectives (and biases) of different researchers, especially within complex systems like education. These reflexive perspectives become untapped sources of knowledge and understanding. Though certain positivist hypotheses may be tested using an IRR approach, the development of promising tangential interpretations and generation of new lines of inquiry may remain hidden in the data. By allowing separate inductive approaches by two or more researchers to proceed uninhibited, a more robust collective theory may surface. Findings from the study offer a reframing of the ongoing philosophical debate over the purpose of qualitative research to embrace opposing views. We respond to Symonds and Gorard's (2010) call to move beyond a simplified notion of *mixed methods* and construct new methodological moments. The empirical evidence presented in this study identifies a range of choices that qualitative researchers can make. Depending on the intentions and research questions, the separate but parallel approach may provide an important alternative path to explore themes and possibilities or purposefully develop a convergent theory from multiple and diverse perspectives. Through this approach, researchers do not suppress individual background and experiences; rather, background and experiences deeply connect the researcher to the data to develop a more substantive approach to theory refinement. In developmental evaluation, this feature may be advantageous.

Implications

The results of this study suggest that future qualitative data analyses in education research and evaluation should consider the use of a collaborative grounded theory approach with researchers who hold explicitly distinct biases. This approach provides a unique opportunity to evaluate individual background experiences and interpretations by using a cross-researcher comparative analysis to confront and capitalize on reflexivity productively, and enrich and enhance interpretation. Future methodological exploration might focus less on strictly applying a quantitative approach to achieve reliability and produce valid interpretation of qualitative data. In its place, researchers might emphasize quality of individual interpretation, analysis, and theory development within the complex systems represented by the data. Unless there is sound rationale within the research design, quantifying independent researcher's interpretations for the purpose of establishing early agreement may be to the detriment of the resulting analysis and theory. By postponing convergence until analysis and theory has undergone a full cycle independently, the often limiting implicit and explicit power dynamics within a research team can be attenuated. Future investigations of this type would engender a deeper understanding of how these power dynamics affect the work of a diverse research team.

By its inductive nature, a signature quality that sets qualitative research apart from quantitative methodology is the subjectivity of interpretation—though both paradigms exhibit reflexivity in unique ways. To take advantage of reflexivity in practice requires hybridizing philosophies and accepting that individual analysis of any data will undergo steps and stages that make identical interpretations and conclusions between two researchers, from start to finish, unlikely, and in a developmental evaluation context, potentially disadvantageous. This study suggests that a team approach to generating new theory may provide a more robust interpretation of system complexity by intentionally omitting an IRR approach to convergence. By challenging their own rationale for applying an IRR frame, researchers may discover a stronger synthesized theory and redefine rigor and reliability in terms of depth and breadth versus percentage agreement.

References

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. *Sociology, 31*, 597-606.

Babbie, E. (2007). *The practice of social research*. Belmont, CA: Thomson Wadsworth.

Baker, B. D., Oluwole, J. O., & Green III, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the Race-to-the-Top Era. *Education Policy Analysis Archives*, *21*, 1-64.

Burawoy, M., (1998). The extended case method. *Sociological Theory, 16*, 5-33.

Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). Thousand Oaks, CA: Sage.

Danielson, C. (2001). New trends in teacher evaluation. *Educational leadership*, *58*, 12-15.

Davey, J. W., Gugiu, P. C., & Coryn, C. L. S. (2010). Quantitative methods for estimating the reliability of qualitative data. *Journal of Multidisciplinary Evaluation*, *6*, 140-162.

Friese, D. S. (2013). *ATLAS.ti 7*. GmbH, Berlin: Scientific Software Development.

Glaser B. & Strauss, A. (1967) *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.

Grayson, K. (2001). Interrater reliability. *Journal of Consumer Psychology*, *10*, 71-73.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527.

Landis, J. R., & Koch, G. C. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*,159-174.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Mays, N. & Pope, C. (1995) Rigour and qualitative research. *British Medical Journal, 311*, 109-11.

Mays, N., & Pope, C. (2000). Assessing quality in qualitative research. *British Medical Journal*, *320*, 50-52.

Maxwell, J. A. (2013). *Qualitative research design: An interactive approach* (3rd ed.). Thousand Oaks, CA: Sage.

Merriam, S. B. (2009). *Qualitative research: A guide to design and implementation*. San Francisco, CA: Jossey-Bass.

Messick, S. (1996). Validity of performance assessments. *Technical issues in large-scale performance assessment*, 1-18.

Miller, P. (2008). Reliability. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 2, pp. 753-754). Thousand Oaks, CA: Sage.

Morse, J. M. (1994). Designing funded qualitative research. In N. K. Denzin and Y. S. Lincoln (Eds.) *Handbook of qualitative research*. London, UK: Sage.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Olesen, V., Droes, N., Hatton, D., Chico, N., & Schatzman, L. (1994). Analyzing together: Recollections of a team approach. In A. Bryman and R. G. Burgess (Eds.) *Analyzing qualitative data*. London, UK: Routledge.

Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York, NY: Guilford.

Pillow, P. (2003). Confession, catharsis, or cure? Rethinking the uses of reflexivity as methodological power in qualitative research. *International Journal of Qualitative Studies in Education*, *16*, 175-196.

Shenton, Andrew, K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information*, *22*, 63-75.

Symonds, J. E., & Gorard, S. (2010). Death of mixed methods? Or the rebirth of research as a craft. *Evaluation & Research in Education*, *23*, 121-136.

Stenbacka, C. (2001). Qualitative research requires quality concepts of its own. *Management Decision*, *39*, 551-555.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 81-84.

Strauss, A. L., & Corbin, J. M. (1990). *Basics of qualitative research* (Vol. 15). Newbury Park, CA: Sage.

Vidich, A. J., & Lyman, S. M. (1994). Qualitative methods: Their history in sociology and anthropology. *Handbook of qualitative research*, 23-59.

Waitzkin, H. (1991). *The politics of medical encounters*. New Haven, CT: Yale.