Scientific Research Publishing

Psychology, 2026, 17(1), 55-76
https://www.scirp.org/journal/psych
ISSN Online: 2152-7199
ISSN Print: 2152-7180

Rethinking Educational Evaluation: From Fourth Generation to an Integrated Model

Zhongfeng Hu[1,2*], Yifei Li[3*], Zhenzhen Zhang[4*], Guanghan Jia[2], Jiawei Li[5#]

[1]Center for Southern Xinjiang Education Development, Kashi University, Kashgar, China
[2]School of Educational Science, South China Normal University, Guangzhou, China
[3]School of Politics and Public Administration, South China Normal University, Guangzhou, China
[4]Kashi Municipal Bureau of Education, Kashgar, China
[5]Department of Counseling, Educational Psychology & Special Education, Michigan State University, East Lansing, MI, USA
Email: #lijiawei@msu.edu

How to cite this paper: Hu, Z. F., Li, Y. F., Zhang, Z. Z., Jia, G. H., & Li, J. W. (2026). Rethinking Educational Evaluation: From Fourth Generation to an Integrated Model. *Psychology, 17,* 55-76.
https://doi.org/10.4236/psych.2026.171003

Received: November 30, 2025
Accepted: January 12, 2026
Published: January 15, 2026

Abstract

Despite its widespread influence on educational evaluation reform, Fourth Generation Evaluation (FGE) has primarily been adopted without sufficient critical examination. This paper provides an in-depth analysis of this constructivist and negotiation-based model by highlighting its theoretical and practical limitations. There are classification flaws and historical gaps in the division of evaluation into generations, which creates ambiguity through the use of the term "generation" and blurs the distinction between measurement and evaluation. Moreover, FGE's foundation in the constructivist paradigm is seen as an insufficient theoretical basis, and its overreliance on naturalism risks leading to untenable outcomes. In practice, FGE is rarely implemented in its entirety, as its procedural complexity and overly idealistic conditions often hinder its application. Given the multifaceted and complex nature of modern educational contexts, FGE is increasingly ill-suited for contemporary evaluation needs. This paper concludes by advocating for an integrated educational evaluation framework with a specific focus on process integration to reconcile the perceived divide between constructivism and positivism.

Keywords

Fourth Generation Evaluation, Quantitative Evaluation, Qualitative Evaluation, Integrated Educational Evaluation

*Co-first authors.
#Corresponding author.

1. Introduction

Modern educational evaluation as a field originated in the 1930s when Ralph Tyler and his curriculum research team articulated its foundational principles through the Eight-Year Study (Tyler, 1949). This work paved the way for classic evaluation models, each reflecting a distinct understanding of the field. The objectives-oriented model (Tyler, 1949) focused on whether predetermined behavioral objectives were achieved. The responsive model (Stake & Fund, 1975) shifted attention to stakeholder concerns and the complexity of the process itself. The Context, Input, Process, and Product (CIPP; Stufflebeam, 1983) model framed evaluation as a comprehensive tool to inform decision-making. Despite these advances, evaluation practices remain overly standardized and methodologically narrow, often lacking responsiveness to the value of pluralism and staying heavily on dominant scientific paradigms of inquiry. In response to these oversights, Egon G. Guba and Yvonna S. Lincoln published *Fourth Generation Evaluation* (FGE) in 1989, aiming to define a new approach to evaluation that transcended the purely scientific paradigm. However, FGE has prompted debate over its conceptual robustness and utility in practice. Although some recent studies have contributed to further reflection on these concerns, a general tendency persists to adopt the model rather than critically examine it. This paper presents a comprehensive review of FGE from both theoretical and practical perspectives and proposes an alternative model tailored for increasingly complex and diverse evaluation contexts.

2. FGE: A Constructivist and Negotiation-Based Model

Guba and Lincoln (1989) described the evolution of educational evaluation in terms of four "generations": measurement, description, judgement, and constructivist. Grounded in the positivist paradigm, the first three generations shared key limitations, including a tendency toward managerialism, a failure to accommodate value pluralism, and overcommitment to the scientific paradigm of inquiry. The tendency toward managerialism refers to administrators wielding financial sponsorship to control evaluators, often creating unequal power dynamics in which some stakeholders' legitimate interests are compromised, and their input is suppressed throughout the evaluation process. In a pluralistic society, competing political parties and interest groups can fuel a perception of bias in educational evaluation, leading the public to question who conducts it and whose interests it serves. This dynamic often results in resistance and non-cooperation from those who feel their interests are at risk. Evaluation results that ignore value pluralism are unlikely to gain acceptance, even when presented as objective, standardized, and scientific. Similarly, overcommitment to the scientific paradigm of inquiry can saturate evaluation practice with excessive operationalization and overreliance on the power of data. Yet, data do not speak for themselves. Under the influence of managerialism, evaluators collect data using empirical and quantitative methods only after aligning with the predefined priorities and scope imposed by the managers (Zhang, 1995). As such, the resulting data often reflect managerial interests more than any

claim to objective truths. This challenge to objectivity is especially pronounced given that even pure natural science cannot claim to be entirely value-free or value-neutral, let alone education methods, which are just one of many ways humans can understand and evaluate the world.

Emerging from a constructivist paradigm, FGE directly responds to the limitations of the earlier generations by reframing evaluation as a socially constructed process centered on stakeholders' claims, concerns, and issues, which in turn informs both the inquiry and the information produced. Specifically, Guba and Lincoln identify seven characteristics that embody the new goals of evaluation: 1) a sociopolitical process, 2) a joint and collaborative process, 3) a teaching/learning process, 4) a continuous, recursive, and highly divergent process, 5) an emergent process, 6) a process with unpredictable outcomes, and 7) a process that creates reality (Guba & Lincoln, 1989: pp. 253-256). Within these processes, stakeholders are empowered as responsive focusing facilitates their active participation, ensuring they can voice and clarify their perspectives on the critical issues, concerns, and claims during negotiation. These interpretations become meaningful constructions (Koch, 2000), representing the core of how stakeholders make sense of their situations (Guba & Lincoln, 1989). FGE respects value pluralism by making negotiation its core driving force, achieved through a hermeneutic-dialectic process where divergent viewpoints are exchanged and continuously negotiated to construct a collective, shared understanding and, ultimately, a "common reality structure" (Rebien, 1996).

In summary, FGE is defined by three core principles: responsiveness, co-construction, and negotiation. Responsiveness, as a starting point, emphasizes a direct engagement with stakeholders' concerns. This engagement leads to co-construction, which represents the essence of evaluation as a joint, psychological construction of value judgments. Co-construction, then, is accomplished through negotiation, the central mechanism for reconciling divergent perspectives and forming these joint constructions. FGE leaves behind evaluation based on measurement, objectives, and professional judgement (Koch, 2000).

The paradigm shift is evident across the fundamental structure of evaluation. The first three generations were rooted in the positivist paradigm and aimed at objective discovery, treating the evaluator as an independent and value-neutral authority who always prioritizes empirical evidence and procedural control over participant voice. In contrast, FGE is grounded in the constructivist paradigm, asserting that realities are multiple and socially constructed, which requires the evaluator to act as a participatory facilitator. This structural move from expert-driven judgement toward collaborative meaning-making rejects the controlling approach of the past. Instead, FGE advocates for the hermeneutic dialectic method to promote active participation and reciprocal interactions among all participants and stakeholders, co-constructing a shared understanding that is "hermeneutic because it is interpretive in character, and dialectic because it represents a comparison and contrast of divergent views with a view to achieving a higher-level

synthesis of them all" (Guba & Lincoln, 1989: p. 149).

The divergence in the evaluation approach is also notable. The first three generations adhered to a preordinate approach where objectives and criteria were predetermined and set before the inquiry began. This approach primarily focused on discrepancies between actual outcomes and predetermined objectives, thereby serving a managerial agenda while neglecting the values and perspectives of the evaluated and other stakeholders. In contrast, FGE adopts a responsive approach that counters these managerial tendencies by not only attending to diverse stakeholder claims, concerns, and disagreements, but also adapting the subsequent design of information collection. As Stake and Fund (1975) argue, this approach allows the evaluation's scope and parameters to emerge through ongoing negotiation and feedback, which reflects the fundamental constructivist view that evaluation should aim for a co-constructed reality rather than simple verification.

Consequently, the evaluation process and the nature of the judgment itself also change with FGE. The first three generations employed a prescriptive, one-way process that lacked open communication and active participation of those being evaluated. FGE rejects the evaluator's exclusive authority, asserting instead that the process should be driven by interactive negotiation and coordination with the evaluator acting as the chief facilitator. Judgment is no longer an authoritative ruling by a single expert, but rather a co-constructed outcome that requires all stakeholders to collectively guide the negotiation agenda, navigate emerging issues, and achieve a shared consensus through the hermeneutic-dialectic process.

3. A Theoretical Critique of FGE

FGE, variously described as constructivist, naturalistic, hermeneutic, or interpretive (Guba & Lincoln, 1989), represents a fundamental departure from the assumptions and practices of the traditional positivist paradigm. FGE is uniquely positioned to address the claims, concerns, and issues of socially marginalized or politically vulnerable stakeholder groups. As a participatory and pluralistic process, it is highly responsive to shifting political and social contexts, as well as the needs of stakeholders. By promoting the expression of diverse values, this distinct model challenges the status quo and paves the way for meaningful social change. Moreover, its grounding in a hermeneutic paradigm, which necessitates comparing and contrasting disparate perspectives, captures the complex human factors often obscured by the positivist paradigm (Lay & Papadopoulos, 2007). In this regard, FGE provides a valuable framework that prioritizes equity, pluralism, and collective negotiation in complex evaluation settings.

Despite its promise, FGE has faced scholarly critiques from various perspectives. Through a political lens, the model suffers from "fundamental flaws" rooted in a serious inconsistency between its espoused political ideals and its operational reality (Reed, 1992). Guba and Lincoln (1989) correctly assert that evaluation is "fundamentally social, political, and value-oriented" (p. 7), positioning themselves in strong opposition to the positivist paradigm by declaring that "every act of science

was also a political act" (p. 118). However, Reed (1992) argues that the central paradox in their work is the reliance on conventional ideals of liberalism, pluralism, and participatory democracy as the means to achieve the stated radical goal of fundamental societal transformation in American society without outlining a clear strategy for achieving these aims within the proposed FGE framework.

Adding to existing theoretical concerns, a substantive critique from the perspective of evaluation participation argues that FGE is undermined by its reliance on an unrealistic assumption of all-inclusive stakeholder involvement. Gregory (2000), for example, suggests that the model's theoretical demand for comprehensive representation often fails to account for structural, administrative, and social barriers that constrain the composition of a truly representative stakeholder group. As a result, the claim that FGE fosters participatory practices is questionable and the stated ideal of promoting multiple stakeholder perspectives, particularly from marginalized or politically vulnerable groups, remains largely aspirational rather than fully realized in practice.

Further, the model's sustained influence across diverse contexts demands a renewed and detailed examination. While prior research has either offered theoretical insights based on close readings of FGE's core texts or identified practical limitations through case-based analysis, this paper extends those established strategies by offering a more detailed and integrated evaluation. Specifically, this paper focuses on FGE's framing of historical shifts in evaluation, its use of specific key terms, its underlying research paradigm, and obstacles encountered during its practical application.

3.1. Flawed Classification and Historical Gaps

Guba and Lincoln's (1989) four-generation schema, while useful for framing their argument, is fundamentally flawed. First, the schema misrepresents the history of evaluation as a sequence of replacements rather than the enduring landscape of methodological pluralism and coexistence. Specifically, the assertion that the first generation (measurement) ended in the 1930s is a historical inaccuracy that reflects untenable internal logic within the schema. Measurement has persisted and evolved, as evidenced by the proliferation of large-scale international assessments. For example, the Program for International Student Assessment (PISA) has shown consistent growth in both international participation (from 32 countries and 75,000 students in 2000 to 79 countries and 600,000 students in 2018) and the scope of its assessments, which now include innovative domains such as interdisciplinary literacy and 21st-century skills. This sustained expansion demonstrates that measurement remains a dominant, continually advancing methodology in educational evaluation.

Second, the schema itself is structurally flawed due to a category error resulting from inconsistent classifications. The proposed generations of "measurement", "description", "judgment", and "constructivism" are drawn from different conceptual domains and therefore do not represent a single, coherent framework. This

conceptual mixture is compounded by the fact that there is little scholarly justification for such a four-generation division of evaluation history (Wang, 2003). Furthermore, this scheme diverges sharply from established practice in the field, which, since Tyler's Eight-Year Study (Tyler, 1949), has signaled the start of modern evaluation, typically organized by model types rather than a linear, generational progression.

Third, the schema's internal chronology is riddled with significant, unexplained temporal gaps. These gaps are arbitrarily fixed, lacking any rationale for the decade-long gap between the end of the second generation (description) in the 1950s and the beginning of the third generation (judgment) in the 1960s, nor for the two-decade gap between the end of the third generation in the 1970s and the formal proposal of the fourth (constructivist) in 1989. This failure to define clear transitional criteria or a consistent timeline exposes the lack of a coherent theoretical basis.

3.2. Problematic Use of Term "Generation"

The use of "generation" to denote the development and paradigmatic shifts within education is conceptually flawed. A valid generational progression requires the dual presence of continuity and iteration. Specifically, a later generation either assimilates the functions of a prior one or builds upon them, while also introducing more advanced capabilities. However, the framework proposed by Guba and Lincoln (1989) does not adequately reflect this evolutionary coherence. Instead of illustrating an iterative succession, the divisions merely contextualize each period by describing its relationship to the evaluation needs and operational environment of its time. The first three generations are more accurately described as distinct, coexisting evaluation types rather than a linear succession in which one replaces the other; indeed, many of these evaluation models are still in use today.

Furthermore, the term "generation" inherently implies that each period represents a step forward, making the most recent generation the most advanced. In reality, however, the paradigm revolution from the first to the fourth generation of evaluation methods did not truly occur (Zhu & Cao, 2024). In fact, the supposed move to the fourth generation represents a fundamental step backward, as it completely rejects the positivist paradigm and forces a false dichotomy between quantitative and qualitative evaluation. The use of "generation" is therefore misguided, as science itself is a highly consensual and collaborative enterprise (Reed, 1992). Abandoning the entire positivist tradition is both unrealistic and unwise, given its significant role in social development.

3.3. Conflation of Measurement and Evaluation

As Guba and Lincoln (1989: p. 22) themselves acknowledge, "Evaluation as we know it did not simply appear one day". In early formal education, teachers used examinations to assess student mastery, and over time, a variety of scales and testing formats were developed to capture other aspects of student development. From a

historical perspective, advances in measurement and the accumulation of empirical experience provided a solid foundation for systematic evaluation.

It is crucial to recognize that measurement and evaluation are not the same. While early measurement practices may have served as a tool or a means to an evaluative aim, the distinction between the two is fundamental. More precisely, measurement is a factual judgement that assigns numerical values to represent the subject attributes, whereas evaluation involves a value judgment that interprets data against an established set of criteria or values to render a conclusion of worth. Given that these practices reside in fundamentally distinct conceptual and disciplinary domains, treating measurement as either the initial step or the first generation of evaluation is conceptually untenable. As Guba and Lincoln (1989) argue, the emergence of evaluation resulted from the complex construction and reconstruction of various social, political, and historical forces, of which the contribution of measurement was merely a partial factor in this process, rather than a holistic one. Hence, equating the two is an unwarranted oversimplification of conceptual boundaries.

3.4. Insufficient Theoretical Basis of Constructivism

While FGE represents a radical departure from the positivist assumptions of its predecessors, its reliance on a constructivist paradigm is conceptually problematic. Unlike the stable realist ontology and positivist assumptions that underpinned the first three generations, FGE is grounded in a relativist ontology, a subjectivist epistemology, and a hermeneutic-dialectical methodology. This framework views facts as human constructs rather than objective realities. That is, truth is not "out there" to be discovered; rather, it is shaped by individuals involved and influenced by broader social and cultural factors. As Patton (2002: p. 44) explains, "Truth is a matter of consensus among informed and sophisticated constructors, not of correspondence with an objective reality". Consequently, FGE advocates for an interactive mode of evaluation in which consensus is achieved through the active, shared, and negotiated participation of evaluators and stakeholders. In this regard, FGE is grounded in naturalism (Liu, 2018), which draws heavily from traditions such as phenomenology, hermeneutics, ordinary language analysis, and symbolic interactionism. This model emphasizes that evaluators and the evaluated interact within a genuine natural setting, and that the necessary data are not discovered but are actively generated through this process of construction.

Although Guba and Lincoln prefer to term it "constructivism" as a guiding paradigm, many of its core assumptions remain difficult to operationalize in the real world. Constructivists reject the notion of an objective, scientifically verifiable reality and the possibility of finding impartial or objective answers. Instead, they argue that multiple realities are socially constructed and negotiated, requiring every stakeholder to be open, trustworthy, and able to communicate effectively for FGE to succeed (Guba & Lincoln, 1989). However, the hermeneutic circle overestimates both the willingness and ability of most stakeholders to engage in meaningful dia-

lectical dialogue, as the model assumes (Reed, 1992). This idealism fails to account for the reality that stakeholder groups often prioritize their own interests. While negotiation can help align competing interests, the concept of relative interest is problematic, as it tends to reinforce the power of dominant groups at the expense of third- or fourth-level beneficiaries whose claims are equally legitimate (Laughlin & Broadbent, 1996). However, FGE offers little practical guidance for addressing these concerns. Its reliance on constructivism as a solution to ontological and epistemological concerns reveals a major theoretical limitation (Russell & Willinsky, 1997). Hence, we maintain that naturalism is merely a stream within modern philosophy and is insufficient to provide a robust and reliable foundation for evaluation practice.

3.5. Untenable Outcomes of Overreliance on Naturalism

Guba and Lincoln agree with Biklen and Bogdan's distinction between two approaches to naturalistic inquiry: one at a methods level, which is simply "a collection of tools and techniques", and another at a paradigm level, which is "a wholly different way of viewing the world" (Guba & Lincoln, 1989: pp. 159-160). Despite this distinction, they note shared features, including "collecting descriptive data in the natural setting, with the evaluation serving as the inquiry instrument, focusing on 'educational issues as they are perceived and experienced by people' (p. 95); and utilizing an inductive process that focuses and narrows as the evaluation proceeds" (p. 160). As the second paradigm level approaches central to constructivism, FGE is characterized by conducting inquiry in natural settings, adopting the role of learners who enter the context without prior assumptions, using qualitative methods, and incorporating tacit knowledge (pp. 174-176).

Departing from the rational (or scientific) paradigm that has traditionally guided the disciplinary research (Stufflebeam, 2007: p. 425), Guba and Lincoln (1989) established themselves as strong proponents of naturalistic inquiry. They argue that this approach effectively avoids the weaknesses inherent in the rational model, providing contextual relevance, rich entry points, and process sensitivity, as well as the ability to build theory from evidence and offering distinct standards to replace the rational model's emphasis on objectivity (Stufflebeam, 2007: p. 426). They conclude that these unique benefits make FGE the ideal realization of the naturalistic paradigm. However, FGE's extreme emphasis on purely qualitative research risks falling into a trap of extreme subjective relativism by establishing a rigid dichotomy between science and the humanities. This approach is fundamentally contradictory to the principle that value, while interpreted subjectively, derives from the inherent properties of external objects and thus possesses an objective foundation (Zhang, 1995). Consequently, this methodological extreme denies the existence of objective data and automatically discards objectivity. The evaluation process is therefore disproportionately governed by the amplified values of the evaluative subjects.

Furthermore, Guba and Lincoln's prediction that the naturalistic paradigm would

become the dominant approach in the 21st century has not materialized. Over the past three decades, this paradigm has not only failed to become mainstream but has also been overshadowed by the emergence of mixed-methods research, which integrates both quantitative and qualitative methods. Ultimately, FGE is more accurately described as a specific evaluation model or a new evaluation theory rather than a new (fourth) generation of evaluation (Zhang, 1995). Its reliance on a single philosophical position prevents it from providing the robust, comprehensive framework it claims.

3.6. Procedural Limitations

FGE, built on a constructivist inquiry and centered on a hermeneutic-dialectical process, represents a core departure from the traditional positivist paradigm. The validity of this process rests upon a demanding and sometimes impractical set of preconditions: diverse stakeholders not only possess genuine disputes over "facts" or issues, but also willingly commit to a cooperative relationship with the evaluator rooted in sincerity, trust, and ethics. Since virtually all "facts" are socially constructed differently based on individual value systems, FGE provides the essential framework for practice.

However, establishing a truly harmonious evaluative relationship is not easily achieved. On the one hand, stakeholders often possess conflicting interests, and any agreement reached through negotiation frequently results from a compromise of interests reached through weighing costs and benefits, rather than a genuine change in values or acceptance from the compromiser. On the other hand, the nature of the evaluative relationship varies with the evaluation's purpose and intended use of its results. High-stakes evaluations, such as those used for selecting and grading, tend to evolve into highly competitive structures. When the demands of direct stakeholders (from whom the outcomes are life-changing) outweigh those of indirect stakeholders, this power imbalance creates significant evaluation distortion and leads to tense, uncooperative relationships. Examples like China's national college entrance exam (*Gaokao*) and its school accreditation systems illustrate evaluation activities where the adoption of FGE fails to achieve its desired effect of consensus through negotiation. Therefore, FGE is not an appropriate primary evaluation paradigm for all activities, but rather is better suited for, and should focus primarily on, formative evaluation. Blindly extending FGE to evaluation activities of a different nature can cause its inherent strengths to transform critical weaknesses, making its distinct rationality an impediment (Zhang, 1995). This limitation is further highlighted by the contradictory reality: programs in highly authoritarian or hierarchical systems could benefit most from FGE's empowering effects, yet constructivist evaluation is unlikely to succeed unless its demanding preconditions are met (Lay & Papadopoulos, 2007).

4. A Practical Critique of FGE

Although Guba and Lincoln's seminal work on FGE detailed its constructivist par-

adigm and implementation principles (Guba & Lincoln, 1989), the complexity and intensity of this design render it cumbersome and challenging to implement in practice. The experience of conducting FGE evaluations reveals practical challenges and dilemmas that are not always apparent in theory, particularly due to the model's reliance on the evaluator's demanding role as a politically mandated social change agent without providing adequate guidance for its exercise (Fishman, 1992).

The inherent methodological limitations of FGE become consistently evident in practice. FGE is widely criticized for its reliance on naïve and idealistic assumptions regarding the structural, social, and managerial aspects of an evaluation. Specifically, the model presumes that stakeholders will naturally align their values through dialogue and that fundamental conflicts of interest can be easily resolved through negotiation, a premise that often fails in real-world political contexts. Furthermore, the consensus central to FGE is often elusive, as the model's idealized communication environment relies on uncritical assumptions of democratic consultation and the principle of procedural justice that are rarely, if ever, easily established in real-world settings (Lay & Papadopoulos, 2007). These failures stem from FGE's lack of operational clarity in several crucial areas, including the under-theorized role of the evaluator, the necessity for rigor in interpreting complex social backgrounds and customs, and the managerial factors essential for controlling a highly participatory environment (Gregory, 2000; Laughlin & Broadbent, 1996). Moreover, FGE suffers from practical deficiencies in stakeholder management, including the difficulty in identifying all relevant stakeholders, a failure to establish concrete metrics and guiding procedures for ensuring mutual understanding among diverse and often conflicting interests, and the absence of detailed operational measures for genuinely empowering stakeholders within the evaluation process (Du, 2010). In sum, FGE's theoretical simplicity and resulting lack of operational clarity restrict its operational capacity and, consequently, limit the likelihood of achieving its stated participatory goals in complex, real-world settings.

4.1. The Lack of Full Implementation

Many purported applications of FGE in project evaluation fail to follow the rigorous hermeneutic dialectic negotiation process outlined by Guba and Lincoln, which is the methodological anchor distinguishing FGE from the conventional positivist paradigm. Instead, practitioners often employ a partial application, selectively using constructivist techniques focused narrowly on stakeholder engagement, which in turn overstates FGE's theoretical contributions and underestimates the complexity of its genuine implementation, often resulting in an incomplete and potentially ineffective evaluation.

In evaluating a game and information project, Lay and Papadopoulous (2007) claimed to apply FGE but focused primarily on its more accessible stages (e.g., stakeholder identification and initial data gathering). Specifically, they identified diverse stakeholder groups, including project managers, coordinators, volunteers

(who were also parents), program staff, and agency personnel, and collected qualitative data primarily through observation. The evaluation team subsequently convened multiple meetings to operationalize the collaborative negotiation and power-sharing principles central to FGE. In these sessions, stakeholders negotiated data collection processes, reviewed tools and methods, deliberated on constructions, and reconciled interpretive discrepancies related to claims, concerns, and issues. These activities affirmed the goal of empowering all stakeholders in the design and implementation of the evaluation. However, the study represents only a partial application of FGE because it omitted the critical, defining stage: the hermeneutic dialectic cycle required to synthesize various stakeholder constructions and achieve a final negotiated resolution. The study failed to demonstrate how structural conflicts between gatekeepers and responders were resolved, nor did it clearly present the multi-round response cycle and its resulting changes. Furthermore, the role of external evidence (e.g., literature, documents, or observation data) in shaping constructions remained unclear, as were the precise procedural responsibilities of both evaluators and stakeholders. Although the researchers assert FGE holds undeniable advantages in project evaluation, the existing omissions make it difficult to determine whether its success is attributable to FGE or other contributing factors, a distinction crucial for understanding the model's true utility.

4.2. Operational Impracticalities of Full Stakeholder Inclusion

A core dilemma in FGE application lies in stakeholder representation. While the model is recognized for its potential to empower stakeholders and amplify the voices of those often overlooked by the traditional positivist paradigm, achieving these equity-focused ideals requires the evaluator to precisely identify all relevant stakeholders, including agents, beneficiaries, and those negatively impacted. This task becomes particularly challenging when an evaluation involves thousands of beneficiaries and hundreds of employees, making comprehensive inclusion methodologically impractical. The model offers no clear operational guidance on several core issues, including the responsibility for conducting the stakeholder analysis, the criteria for ensuring identification of all relevant stakeholders, or the process for guaranteeing genuine participation across all stakeholder groups. This lack of concrete guidance points to an inherent and unresolved issue with the method itself (Rebien, 1996).

The evaluation of a university-community youth development partnership (Huebner & Betts, 1999) clearly demonstrated the complexity of reconciling FGE's methodological demands with operational realities. In this study, FGE was primarily utilized during the initial collaborative stage to listen and identify critical factors contributing to positive youth development. The evaluation team engaged 20 stakeholder representatives (including youth, parents, school staff, and community members from diverse backgrounds) and employed the constant comparative method as the main data analysis approach to simultaneously collect and process data. Following FGE's guidance for generating constructions, the inquiry began with a

series of open-ended questions posed to the initial stakeholders. Each subsequent interviewee was then asked to respond to the emerging constructs of the previous participant to form a collective understanding of the concept of positive youth development. This iterative process continued across multiple rounds of interviews, ensuring every participant provided their perspectives on the evolving concepts. While this case demonstrated FGE's strength in building consensus on evaluation content by integrating diverse stakeholder views, the model struggled to engage stakeholders with broad, abstract topics such as "positive youth development". This difficulty necessitated the incorporation of external theories to justify the inclusion of specific stakeholders. This reliance on external theory, however, often introduced inconsistency in stakeholder definition and ultimately compromised the reliability of the evaluation outcomes.

Another critical factor to consider is the evaluation capacity of the stakeholders themselves. The hermeneutic-dialectic process central to FGE demands a high level of cognitive and communicative competence, requiring participants to possess both sufficient background knowledge and strong expressive skills. Effective participation, therefore, requires individuals to be able to clearly articulate their own views and remain intellectually open to the perspectives of others. Nevertheless, the model provides no clear guidance on selecting stakeholders who possess these requisite capacities. The controversial nature of the topic further undermines the collective construction process, making it less probable that FGE can convene a diverse group of stakeholders who are equally knowledgeable and genuinely open to opposing viewpoints (Huebner & Betts, 1999).

In sum, FGE is seemingly straightforward in theory but difficult to implement in practice. The model requires a significant investment of an evaluator's time and energy and demands a very high level of professional expertise. For large-scale school evaluations, such as those conducted in school systems, questions regarding stakeholder selection and the construction of the hermeneutic cycle present a substantial and often unresolved challenge (Lu, 2009).

4.3. Overly Idealistic Conditions for Implementation

Documented cases of using FGE for project evaluation remain limited, yet they consistently reveal a set of actual practical issues in the field. For instance, an evaluation of a hospice care project identified several internal barriers directly attributable to its use of FGE, including ineffective communication, a lack of specialized training for professionals, and interpersonal difficulties among team members (Almeida et al., 2019). Similarly, challenges in the practical application and stakeholder engagement of FGE have been directly linked to the model's core collaborative and collective expectations (Pavani et al., 2023). The necessity of specific evaluator expertise further compounds these issues, as demonstrated by the study of an Australian youth movement that integrated FGE with participatory and empowerment models. Promoting positive multi-stakeholder interaction becomes "an uphill battle" if the evaluator lacks proactive strategic skills, decisive judgement,

and the ability to facilitate collaborative problem-solving among stakeholders (McDonald, 2008). The need for this high level of expertise is magnified because when evaluation results are presented and debated, conflict, rather than consensus, often emerges, demanding that the evaluator possess sophisticated skills in contention management and enabling resolution. Collectively, these cases demonstrate that the practical application of FGE is frequently constrained by contextual barriers and inadequate evaluator capacity, thus failing to achieve the model's idealized participatory outcomes.

5. Beyond FGE: Toward an Integrated Educational Evaluation

Drawing on the preceding review and reflection, educational evaluation should actively challenge the monistic tendency in traditional evaluation practice. The field has reached a critical juncture that requires strengthening the integration of evaluation paradigms grounded in theoretical and methodological reflection to ensure the evaluation process remains pluralistic and open. Inspired by the pioneering efforts of Chinese scholars who have developed evaluation models that integrate quantitative and qualitative methods (Hu et al., 2006), despite their work having received little attention, we proposed the Integrated Education Evaluation (IEE) framework. This approach seeks explicitly to reconcile constructivist and positivist paradigms, offering a more adaptive and inclusive approach to evaluation.

5.1. The Core Principles of IEE

IEE is fundamentally a process integration evaluation paradigm, defined by the systematic synthesis of diverse evaluation methods across the entire evaluation process. The framework is grounded in the triadic philosophical tenets of dialectical materialism, systems theory, and scientific humanism. From dialectical materialism, IEE draws on the laws of the transformation of quantity into quality and the unity and conflict of opposites. The former emphasizes that quantitative and qualitative changes represent distinct yet interconnected stages of development, where accumulated quantitative changes eventually precipitate a fundamental shift in qualitative insight. The latter acknowledges that contradiction is inherent to all processes, with unity and struggle forming the inseparable core of dialectical motion. Consequently, IEE views quantitative and qualitative dimensions not as a binary opposition, but as a symbiotic duality. Rather than prioritizing one over the other, the framework recognizes that the pendulum-like development between them is precisely what drives evaluative theory and practice toward maturity. Complementing this dialectic, systems theory provides a holistic lens, asserting that systems are universal structures and that wholeness is their fundamental property. Because any phenomenon can be viewed as a system, its essence lies in the organic interaction of its elements rather than their mechanical summation. In educational evaluation, where single-method approaches often fail to capture the full complexity, IEE integrates diverse methodologies to generate emergent

properties. This approach ensures that the evaluative whole is substantively greater than the sum of its parts, yielding more accurate value judgments of complex phenomena. Finally, this integration is guided by the tenets of scientific humanism. IEE advocates for the mutual permeation of scientific and humanistic culture, reconciling the friction between technical rationality and intrinsic value. While scientific inquiry provides the clarity needed to navigate complex systems, humanistic culture identifies the ultimate purpose of the inquiry and provides the momentum for scientific growth. This integration ensures that IEE remains rigorous yet human-centered. By establishing these philosophical foundations centered on integration, IEE establishes methodological parity by giving equal standing to both quantitative and qualitative evaluations, recognizing their unique strengths and weaknesses. IEE thus promotes the flexible integration of both sets of methods to effectively address the complex and dynamic nature of modern evaluation tasks.

The essence of integration within the IEE framework moves beyond the general definition of combining elements into a new whole (*Xinhua Chinese Dictionary*, 2013: p. 1246); instead, this process reflects the core philosophical tenet that true integration yields emergent properties, which are unattainable when examining components in isolation. Accordingly, the process of integration is not a simple aggregation, but a dynamic, organic fusion designed to create a more comprehensive system and to profoundly enhance the effectiveness of each contributing component.

Drawing on this foundation, integration within IEE is defined by two specific, interconnected meanings that guide its application. First, integration transcends simple methodological aggregation, necessitating the synthesis of underlying philosophies and approaches throughout the entire evaluation process. Second, integration is goal-driven, establishing the primacy of purpose over methods. Methods serve only as tools to achieve specific goals, and their selection should always be dictated by the evaluation task at hand. The framework discourages unnecessary methodological combinations when a single quantitative or qualitative method can achieve goals effectively. Similarly, forcing any unsuited method for a specific evaluation task is deemed inappropriate.

Within the IEE framework, integration is not a mandatory requirement for every project, but as a philosophical guiding principle. This integration process is instead governed by efficiency, rationality, and scientific rigor. Notably, the integration of quantitative and qualitative methods is not only limited to the final analysis stage, but is best realized by beginning at the initial stage and continuing dynamically throughout the research process (Perez, 2019). Building on this nascent form of process integration, IEE expands the scope by advocating for the integration of all evaluation methods across the entire evaluation process.

In summary, IEE represents an educational evaluation philosophy in which the idea of integration permeates every stage and aspect of the evaluation process. The following three dimensions of integration capture its foundational strengths.

5.1.1. Integration across All Evaluation Stages

Integration in IEE is a continuous, holistic process that requires the strategic and flexible use of all relevant methods as needed at every stage of the evaluation process. This emphasis arises from the recognition that each evaluation stage presents unique complexities, which in turn demand an adaptive, multi-method approach for effective resolution. For instance, the preparation stage involves critical initial steps, including analyzing background information, defining key questions, and designing the comprehensive plan. The inherent diversity of these steps and their related activities, ranging from conducting preliminary interviews and field investigations to reviewing extensive documentation and performing diagnostic assessments, necessitates a combined, integrated methodological approach. Through the coordinated use of quantitative and qualitative data collection, complex problems are triangulated and probed from multiple dimensions. This integrative approach to problem framing ensures that the evaluation remains dynamic, open, and adaptive from the outset. Similarly, the implementation stage necessitates continuous methodological integration to uphold the responsiveness. IEE requires that methods to be dynamically adjusted and articulated in real time to address the specific purposes, target populations, and unforeseen challenges that arise. This real-time adaptability is crucial to ensuring efficacy and seamless implementation.

5.1.2. Integration of Diverse Methods

In educational evaluation, a common approach to generating complete and reliable findings often involves the simple combination of quantitative and qualitative methods. For example, during implementation, the evaluator simultaneously captures rich qualitative context through subject observation and records precise quantitative data on the frequency of specific behaviors. Furthermore, IEE recognizes that integration is not limited to combining methods between paradigms, but is also encouraged to occur within a single paradigm, such as combining different quantitative methods or different qualitative methods to afford a richer perspective than any single approach could offer. This orientation rests on the core principle that integration is not achieved through arbitrary accumulation, but through rigorously reasoned and theoretically grounded alignment aimed at constructing a more robust, multi-dimensional triangulation of evidence. A prime example is the rigorous use of triangulation, where evidence from diverse sources, methods, and perspectives is systematically synthesized. This process not only facilitates the continuous testing and refinement of conclusions but also helps uncover new insights and reduce bias, ultimately leading to more robust and valid outcomes.

5.1.3. Integration of Philosophies

IEE rejects the notion of a single, universally superior method, maintaining that no evaluation method is inherently good or bad. Instead, the framework recognizes that every method has specific strengths and limitations, and that no single

method is universally sufficient across contexts. Therefore, IEE views all methods as complementary tools, rather than ranking them. Methodological value is determined by utility and alignment with specific goals and contexts of a given evaluation. Because evaluation methods and purposes dynamically interact throughout the process, reliance on any fixed hierarchy or predetermined weighting becomes unhelpful and ineffective for comprehensive evaluation. Accordingly, IEE moves beyond a narrow methodological orientation focused on method selection to emphasize the deliberate construction of a responsive methodological system grounded in problem-solving and sensitive to the complexity of authentic educational problems and the evolving nature of evaluation contexts.

5.2. Key Considerations of IEE Application

To facilitate a better understanding of IEE's structure and process, this section provides a brief explanation of the three key stages: Evaluation Preparation, Evaluation Implementation, and Evaluation Analysis. An illustrative application of IEE in teaching evaluation and its conceptual differences compared with mixed-methods evaluation are also presented.

5.2.1. Evaluation Preparation

The preparation stage focuses on clarifying the evaluation's purpose and developing the comprehensive plan. IEE begins with an extensive preliminary investigation that is critical to establishing its purpose. Specifically, the team conducts a thorough analysis of internal and external environments, aimed at identifying key issues and understanding the distinct concerns, values, and attitudes of all relevant stakeholders. This preparatory work enables the evaluation team to delineate the individuals, events, behaviors, processes, and meanings most central to the inquiry. Rather than relying solely on the deductive logic of quantitative research, defining the evaluation purpose is often inherently inductive or involves a strategic integration of both approaches, which helps evaluators construct a more holistic understanding of the issues at hand.

The development of a strong indicator system is a critical component of evaluation design. While quantitative methods are essential for establishing validity, they are insufficient on their own. The evaluator's own understanding of the content, cumulative experiences, professional knowledge, and tactical expertise are equally vital, as these factors inform the value judgements involved in selecting or discarding specific indicators. Consequently, the central issue is less about whether qualitative and quantitative methods can be used simultaneously, and more about the often-overlooked dimension of evaluator's methodological awareness and critical reflection. Rather than opposing different evaluation methods as mutually exclusive choices, a more productive approach is to redefine them as a comprehensive toolbox. This means that the selection process is best guided by employing the most advanced and appropriate methods, while preserving the flexibility to adjust them according to evolving practice and real-time demands.

5.2.2. Evaluation Implementation

The successful implementation of IEE necessitates the rigorous collection of both quantitative and qualitative data, and the subsequent integration of the resulting findings to advance the evaluation activity. The central challenge, however, is that integration is neither formulaic nor straightforward. Since the specific forms of integration are too diverse for exhaustive listing, evaluators need to rely on the core principles to determine the most effective approach for any given context. Notably, the integration of evaluation methods is not a mere stacking of techniques, but a process of abstracting their underlying philosophies. Evaluators need to transcend the labels of "quantitative" and "qualitative" to discern the inherent value within each method. Consideration of a method's methodological, ontological, and epistemological foundations is thus the critical starting point for fully understanding its unique value and inherent differences.

Subsequently, the philosophical groundwork allows evaluators to adopt a higher-level perspective and act as reconstructive agents. By using disaggregated elements as building blocks, evaluators can leverage each method's strengths while mitigating its weaknesses, creating an integrated outcome that is greater than the simple sum of its parts. Ultimately, this integrated outcome is not a fixed or predetermined evaluation model, but a flexible and rational framework for examining the complexity of evaluation subjects and content.

5.2.3. Evaluation Analysis

The analysis stage begins by systematically sorting and categorizing all evaluation materials to fit a designated framework. In a mixed-methods approach, evaluators apply various statistical techniques to the quantitative data while simultaneously interpreting the qualitative data through diverse theoretical lenses, such as hermeneutics, constructivism, psychology, and sociology.

While quantitative results can be generalized from a sample to an entire population through mathematical probability, generalizing qualitative results requires considerable caution due to their inherently unique and context-specific nature. The synthesis of these diverse findings into a comprehensive judgment requires applying relevance theories and methods of relevance from fields like pedagogy, statistics, and fuzzy mathematics. This rigorous process, guided by specific standards and decision rules, leads to a final, well-supported conclusion.

Crucially, IEE calls on evaluators to resist the premature dismissal of divergent findings or minority perspectives, particularly during the early phases of the evaluation. By maintaining procedural independence, IEE allows distinct methodological strands to unfold independently, ensuring that alternative narratives are fully articulated before integration occurs. In instances where quantitative and qualitative results prove incompatible in the later stages, IEE treats such discrepancies as heuristic opportunities rather than as methodological deficiencies. Specifically, IEE adopts a dual strategy. It employs side-by-side comparisons to grant each finding an equal voice and preserve complexity, while simultaneously using these contrasts as a catalyst for higher-order synthesis. Rather than settling for one

finding over the other, evaluators are prompted to construct a third interpretation that reconciles the tension by situating the competing explanations within a more comprehensive and sophisticated theoretical lens.

5.2.4. Illustrative Application

A brief illustration of IEE can be drawn from its application to teaching evaluation. The process begins with the construction of a comprehensive situational profile, synthesizing data from surveys, interviews, and pedagogical artifacts to capture the complexities of teaching processes, teaching quality, student performance, and institutional functioning. Within this context, evaluators employ a blend of inductive and deductive reasoning to sharpen the evaluative focus, ensuring the inquiry's objectives align with the diverse value propositions of students, teachers, and administrators. This alignment is operationalized through an integrated indicator system that pairs quantitative metrics (e.g., measurable learning outcomes and teaching quality) with qualitative indicators that capture process data (e.g., behavior changes and classroom dynamics). While quantitative methods provide the structural rigor necessary for scientific validity, qualitative approaches facilitate the critical value-based selection of indicators and the assignment of their relative weights. Implementation involves multi-stakeholder participation, drawing on the perspectives of administrators, peers, students, and teachers to collect data through summative, formative, and comprehensive assessments. This multi-modal approach integrates observation records and peer review to document the teaching process alongside performance data, ensuring that the collected evidence reflects both measurable outcomes and the nuances of the teaching process. During the final analytical stage, evaluators employ triangulation to cross-reference quantitative results, such as standardized scores, with qualitative evidence, such as classroom observations, to maintain objectivity and minimize bias. This process produces an integrated evaluative conclusion that offers both numerical rankings and rich, descriptive narratives.

Notably, while FGE similarly advocates for the inclusion of multiple stakeholders, it prioritizes contextualism and the interplay between teaching and learning. Unlike IEE, FGE places little emphasis on quantitative measures, preferring qualitative methods to assess student learning and teacher behavior, and relying heavily on process-oriented analysis to form value judgments.

5.2.5. Conceptual Distinctiveness

While IEE shares features with mixed-methods evaluation, it differs in several key aspects. First, it moves beyond procedural mixing toward substantive integration. A standard mixed-methods approach often emphasizes the phased use of different methods, employing them additively or sequentially to capitalize on respective strengths. In contrast, IEE prioritizes a deep synthesis of evaluative reasoning, ensuring diverse methodologies are organically embedded throughout the entire evaluation lifecycle, including conceptualization, design, implementation, and feedback. Second, IEE advocates for a parsimonious approach to integration. While traditional

mixed-methods evaluation often expects the simultaneous collection of quantitative and qualitative data, IEE is informed by efficiency and rational necessity. It maintains that not every evaluation requires integration, nor must every instance involve a combination of both quantitative and qualitative data. Instead, the decision to integrate remains purposeful and driven by specific evaluative needs. Third, IEE prioritizes process-wide integration over stage-specific mixing. While mixed-methods evaluation typically coordinates methods at predetermined intervals or across two consecutive stages, IEE considers integration a dynamic process not restricted to any single phase. Instead, integration can occur at any juncture of the evaluative activity. Consequently, IEE is less a fixed model and more of an integrative mindset. Because the form of integration evolves alongside the progress of the evaluation, it remains distinct from the relatively stable patterns found in traditional mixed-methods evaluation.

The philosophical cornerstone of IEE lies in its integrative logic, which serves as the primary mechanism for mitigating managerialism's constraints and facilitating equitable stakeholder engagement. By adopting an integrative logic, IEE repositions evaluation from a technical routine toward a democratic paradigm. This shift requires administrators to actively align organizational priorities with the distinct interests of all stakeholders. Ultimately, integration facilitates a collaborative space where evaluators and participants negotiate shared objectives as equals. By balancing the administrative drive for measurement with the participants' need for equity and agency, IEE effectively transforms evaluation into a shared governance of power. Furthermore, the process-based integration advocated by IEE creates the necessary space to reconcile stakeholder conflicts. This integrative orientation fosters an open, pluralistic framework in which diverse methods generate a multi-dimensional evidence chain. Within this framework, diverse data streams both supplement and cross-examine one another, ensuring the claims of various interest groups are given equal weight. This prevents any single party from dominating the evaluative discourse and preserves the possibility of mediated dialogue. In this way, the interpretive process shifts from a unilateral administrative judgment to collective decision-making. Integration thus becomes a process of constructing shared values, rendering the final results a true vehicle for the common interest.

Even as IEE attempts to bridge the methodological and axiological divides, it maintains a reflective stance regarding the limitations of positivism in social sciences. Complex social phenomena, particularly in education, cannot be fully reduced to mere natural occurrences. Empirical conclusions often fall short of capturing the intentionality and cultural contexts of human behavior. Moreover, the longstanding methodological divisions within empirical research can result in fragmented understandings of social reality. As previously discussed, IEE addresses these limitations through a fundamental shift in conceptual mindset. By promoting a systemic and holistic perspective, it encourages evaluators to remain inquisitive and critical of different methodologies. IEE prompts evaluators to conduct

practice based on the specific problem and purpose rather than a rigid adherence to methodological dogma. Through this rational reflection and methodological bridging, the inherent limitations of studying complex social systems can be effectively mitigated.

6. Conclusion

Evaluation methodologies must continually adapt in response to shifting political and social contexts and evolving needs of diverse stakeholders (McCoy & Hargie, 2001). While the interpretivist paradigm underlying FGE remains relevant in certain settings, its limitations have become increasingly apparent. First, FGE's reliance on naturalistic inquiry and its stark opposition to the rationalist paradigm do not fully demonstrate its theoretical superiority or innovative value. The binary opposition inherent in this theoretical critique is, to a degree, its Achilles' heel, as it easily traps FGE in a circular reasoning process that undermines its overall coherence and rigor. Second, FGE's promotion of naturalistic evaluation dangerously privileges qualitative methods and the hermeneutic-dialectic cycle, creating a methodological schism that elevates qualitative evaluation to a singular position of authority while simultaneously marginalizing essential quantitative methods. This bias forces a wedge between human insights sought through interpretation and scientific rigor required for comprehensive accountability, thereby leaving the entire evaluation process more vulnerable to subjectivity. Third, given the rapid pace of social development and the increasing demands of project complexity, FGE's singular methodological orientation clearly lacks the necessary adaptability, necessitating that the field move beyond these limitations and actively pursue a new evaluation framework capable of achieving necessary integration and iteration. From this perspective, the IEE framework proposed in this paper presents a valuable initial attempt to address this need, although it requires further refinement and empirical validation to fully realize its potential.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Almeida, C. S. L. d., Marcon, S. S., Matsuda, L. M., Kantorski, L. P., Paiva, B. S. R., & Sales, C. A. (2019). Operation of a Hospital Palliative Care Service: A Fourth-Generation Evaluation. *Revista Brasileira de Enfermagem, 72,* 383-390. https://doi.org/10.1590/0034-7167-2017-0848

Du, Y. (2010). Negotiation and Consensus: the Realistic Choice of Improving Evaluation Effectiveness: Based on the Analysis of Implementation Cases of the Fourth Generation Evaluation. *Research in Educational Development, 30,* 47-51. (In Chinese)

Fishman, D. B. (1992). Postmodernism comes to program evaluation: A Critical Review of Guba and Lincoln's Fourth Generation Evaluation. *Evaluation and Program Planning, 15,* 263-270. https://doi.org/10.1016/0149-7189(92)90090-h

Gregory, A. (2000). Problematizing Participation: A Critical Review of Approaches to Par-

ticipation in Evaluation Theory. *Evaluation, 6,* 179-199. https://doi.org/10.1177/13563890022209208

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth Generation Evaluation.* SAGE.

Hu, Z. F., Li, X. Q., & Liu, G. Q. (2006). On the Integration of Qualitative Evaluation and Quantitative Evaluation. *Curriculum, Teaching Material and Method, 26,* 3-7. (In Chinese) https://doi.org/10.19877/j.cnki.kcjcjf.2006.02.001

Huebner, A. J., & Betts, S. C. (1999). Examining Fourth Generation Evaluation: Application to Positive Youth Development. *Evaluation, 5,* 340-358. https://doi.org/10.1177/13563899922209020

Koch, T. (2000). 'Having a Say': Negotiation in Fourth-Generation Evaluation. *Journal of Advanced Nursing, 31,* 117-125. https://doi.org/10.1046/j.1365-2648.2000.01253.x

Laughlin, R., & Broadbent, J. (1996). Redesigning Fourth Generation Evaluation: An Evaluation Model for the Public-sector Reforms in the UK? *Evaluation, 2,* 431-451. https://doi.org/10.1177/135638909600200406

Lay, M., & Papadopoulos, I. (2007). An Exploration of Fourth Generation Evaluation in Practice. *Evaluation, 13,* 495-504. https://doi.org/10.1177/1356389007082135

Liu, Z. J. (2018). *Educational Evaluation.* Beijing Normal University Press.

Lu, L. T. (2009). Measurement, Description, Judgment, and Construction: A Review of the Four Generations of Educational Evaluation Theories. *Educational Measurement and Evaluation, 3,* 4-7, 17. (In Chinese) https://doi.org/10.16518/j.cnki.emae.2009.03.002

McCoy, M., & Hargie, O. D. W. (2001). Evaluating Evaluation: Implications for Assessing Quality. *International Journal of Health Care Quality Assurance, 14,* 317-327. https://doi.org/10.1108/09526860110409081

McDonald, D. (2008). Revisiting a Theory of Negotiation: The Utility of Markiewicz (2005) Proposed Six Principles. *Evaluation and Program Planning, 31,* 259-265. https://doi.org/10.1016/j.evalprogplan.2008.03.003

Patton, M. Q. (2002). *Qualitative Research & Evaluation Methods.* SAGE.

Pavani, F. M., Pires, A. U. B., Wetzel, C., Olschowsky, A., & Duarte, M. D. L. C. (2023). Fourth Generation Evaluation as a Path for Knowledge Translation in Mental Health. *Revista Gaúcha de Enfermagem, 44,* e20220226. https://doi.org/10.1590/1983-1447.2023.20220226.en

Perez, A. (2019). *Assessing Quality in Mixed Methods Research: A Case Study Operationalizing the Legitimation Typology.* Ph.D. Thesis, University of Nebraska-Lincoln. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1335&context=cehsdiss

Rebien, C. C. (1996). Participatory Evaluation of Development Assistance: Dealing with Power and Facilitative Learning. *Evaluation, 2,* 151-171. https://doi.org/10.1177/135638909600200203

Reed, M. C. (1992). Book Reviews: Fourth Generation Evaluation by Egon Guba and Yvonne Lincoln. Newburg Park, CA: Sage, 1989, 294 pp. *Evaluation Practice, 13,* 197-203. https://doi.org/10.1177/109821409201300308

Russell, N., & Willinsky, J. (1997). Fourth Generation Educational Evaluation: The Impact of a Post-Modern Paradigm on School Based Evaluation. *Studies in Educational Evaluation, 23,* 187-199. https://doi.org/10.1016/s0191-491x(97)00012-6

Stake, R. E., & Fund, J. (1975). *Evaluating the Arts in Education: A Responsive Approach.* Bell & Howell.

Stufflebeam, D. L. (1983). The CIPP Model for Program Evaluation. In G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation Models: Viewpoints on Educational and Human Services Evaluation* (pp. 117-141). Springer.

https://doi.org/10.1007/978-94-009-6669-7_7

Stufflebeam, D. L. (2007). *Evaluation Models.* Peking University Press.

Tyler, R. W. (1949). *Basic Principles of Curriculum and Instruction.* University of Chicago Press.

Wang, X. L. (2003). An Introduction to Several Typical Models of Educational Evaluation. *Early Education, 2,* 8-9. (In Chinese) https://doi.org/10.3969/j.issn.1005-6017-B.2003.02.003

Xinhua Chinese Dictionary (2013). The Commercial Press.

Zhang, M. X. (1995). Response, Negotiation, and Co-Construction: A Review of the "Fourth-Generation Evaluation Theory". *Journal of Foreign Education Studies, 3,* 53-59. (In Chinese)

Zhu, D. Q., & Cao, D. F. (2024). Prospects for the Reform of Educational Evaluation in the New Era: The Fifth-Generation Educational Evaluation. *Global Education, 53,* 58-73. (In Chinese) https://doi.org/10.3969/j.issn.1009-9670.2024.12.005