ABSTRACT

Software Engineering (SE) problems are—from both practical and theoretical standpoints—immensely complex, involving interactions between technical, behavioral, and social forces. In an effort to dissect this complexity, SE researchers have incorporated a variety of research methods. Recently, the field has entered a paradigm shift—a broad awakening to the social aspects of software development. As a result, and in concert with an ongoing struggle to establish SE research as an empirical discipline, SE researchers are increasingly appropriating methodologies from other fields. In the wake of this self-discovery, the field is entering a period of methodological flux, during which it must establish for itself effective research practices. We present a unifying framework for organizing research methods in SE. In the process of elucidating this framework, we dissect the current literature on replication methods and place replication appropriately within the framework. We also further clarify, from a high level and with respect to SE, the mechanisms through which science builds usable knowledge.

1. INTRODUCTION

Even though a scientific explanation may appear to be a model of rational order, we should not infer from that order that the genesis of the explanation was itself orderly. Science is only orderly after the fact; in process, and especially at the advancing edge of some field, it is chaotic and fiercely controversial.

— William Ruckelshaus [36, p. 108]

Nancy Leveson introduced her keynote address at the 1992 International Conference on Software Engineering with this quote on the chaotic and controversial processes of science [26]. The observation that science is disorderly and that scientific explanations are not born in coherence connects deeply with researchers and practitioners of Software Engineering (SE), who for decades have struggled to study the immensely complex phenomena by which software practitioners create some of the most complex systems ever conceived by humankind.

In the wake of the tragic radiation therapy overdoses by the Therac-25 (in the late 1980s) [27] Leveson’s 1992 observations foreshadowed the catastrophic failure of the Ariane 5 rocket in 1996 [13]. In both cases, software development processes were shown to have been the cause. These events, among others, helped to bring about a broad awakening in the SE research community—a general recognition of the immature scientific foundations on which practitioners were building software. Referring to those foundations, Leveson further noted that “we may be straining at the limits of what we can do effectively without better inventions based on known scientific and engineering principles” [26, p. 7].

Although several researchers, particularly those at the University of Maryland [5, 42], had at that point been working for several decades to establish sound research methodologies within SE, as a field, SE had not yet broadly understood, nor incorporated their work. Consequently, SE practitioners were, prior to 1990, operating on theoretical principles grounded primarily in anecdotal evidence.

In response to growing concerns regarding the state of research practices in SE, advocates for empirical science began educating the SE community, making the case for analyzing and improving the field’s core research methods [4, 6, 22, 23, 26, 33, 44]. Numerous forums were organized to advance the state of empiricism in SE and to incentivize researchers to adopt more rigorous practices [42]. These efforts have significantly improved the empirical foundations of our field.

1.1 A Unified Framework for Research Methodologies

Although the SE discipline has progressed significantly over the past fifteen years, it has not yet fully synthesized, understood, established, and embodied research principles by which to produce usable knowledge [6, 20, 42]. Many researchers have contributed to the topic of knowledge building [6, 12, 19, 20, 39, 42], but the numerous frameworks, processes, observations, and insights that have been put forth have not been adequately synthesized and unified under a common understanding [42].

Objective 1. Establish a unified framework for SE research methodologies.

1.2 A Context-Specific Interpretation of Replication

In an effort to separate “what is actually true” from “what is only believed to be true,” SE has, for more than two decades, been pushing to establish its empirical foundations [6, p. 456]. The idea behind the move to empiricism is to separate truth from belief, and in so doing to build knowledge. However, we seem to have applied much of our empiricism to an endless set of new problems, such that after several decades, “the balance between evaluation of results and development of new models is still skewed in favor of unverified proposals” [6, p. 456] [19, 44].
To help address this problem, Brooks, Roper, Wood, Daly, and Miller refined for SE a set of principles from other disciplines for conducting scientific replications [8, 9, 12]. They presented the first formal embodiment of the concept of replication to the SE community in the mid-1990s. Since that time the topic has been analyzed by several researchers and various groups have made efforts to incorporate replication practices into SE [20, 21, 28, 38, 39]. Nevertheless, the low frequency of replications (originally noted by Brooks et al. [8]) seems to have only marginally improved [16, 20].

We believe, as Juristo and Vegas have indicated, that “we might be dealing with the issue of SE experiment replication from too naïve a perspective” [20, p. 356]. The traditional purpose of replication (i.e., to validate results in case of experimental flaws or fabrication) needs to be analyzed and adapted to the SE context [20].

Objective 2. Refine the concept of replication as it applies to SE and place it appropriately within the unified framework of research methodologies.

1.3 An Understanding of the Knowledge Building Process

When effectively practiced, replication has the power to build knowledge [9]. Juristo and Vegas note that “after several replications have increased the credibility of the results, the small fragment of knowledge that the experiment was trying to ascertain is more mature” [20, p. 356]. However, there is little discussion in the literature on the actual mechanisms and processes by which replication matures knowledge, and the obvious function of replication as a guardian against experimental mistakes and fabrication does not really satisfy that question. Further, attempts at “exact” replication, in order to validate results, have thus far proven to be difficult and (in isolation) relatively ineffective [20].

Incorporating the recent work of Juristo and Vegas [20], we attempt to answer the question of how replication can be used effectively within SE to build knowledge. We posit that replication—properly interpreted—is the key to building knowledge and, consequently, the primary mechanism by which theoretical paradigms [25] are created, evolved, broken, and replaced.

Objective 3. Clarify the mechanisms through which science builds usable knowledge and identify the role that replication plays in the knowledge building process.

In pursuing these three objectives, we must be clear that none of the component ideas in this paper are new. Our contribution occurs in the synthesis of methods and insights from numerous other works, which we attempt to juxtapose within a governing framework. Therefore, we provide synthesis, structure, and refinement to concepts that have previously been presented, largely independent of one another.

In the next four sections we lay the groundwork necessary for addressing the outlined objectives. Section 2 discusses the complex nature of SE and a theoretical perspective with which to manage that complexity. Section 3 identifies three fundamental levels on which we need to simplify research results and discusses the problem of premature generalization. Section 4 outlines the primary SE research methods and argues the necessity of a multi-method approach to SE. And finally, section 5 explores the concept of fundamental patterns, the process of discovering those patterns, and the role of replication in that process. Following those four sections, we present the Cycle of Maturing Knowledge, a unified framework for research methodologies in SE.

2. SOFTWARE ENGINEERING IS COMPLEX

Modern software systems are immensely complex, but SE has formulated metrics and techniques by which to measure, handle, and dissect that complexity, thus enabling practitioners to produce large-scale systems that were previously infeasible. However, despite advances in software systems design, software production processes have not yielded as completely to dissection and understanding. Researchers have studied these processes for nearly half a century, but still struggle to synthesize observations into general theories that hold consistently across environments. Consequently, during the question-answer session following Steve McConnell’s keynote address at the 31st International Conference on Software Engineering [30], when asked for his opinion regarding the implications of specific development techniques, McConnell responded with the oft-repeated SE phrase, “It depends.” McConnell then elaborated on the implications of various development contexts and the limitations of SE theory to predict context-specific interactions.

The complexity of software development environments, which arguably exceeds that of the largest software systems (especially when considering the human and social elements), is a significant contributor to the context-specific limitations of SE theory. SE environments—and thus SE experiments—involve deep interactions between technical, behavioral, and social forces. On this point, Juristo and Vegas note that the context of a single experiment is subject to literally hundreds of variables [20]. That claim is easily supported, for example, by the seemingly simple concept of programmer productivity, which has been linked by numerous studies to upwards of 250 contributing factors [18, 31].

2.1 Common Theoretical Perspectives for Managing Complexity

In an effort to understand the complexity of software environments, researchers have adopted a number of theoretical perspectives. In most cases, these perspectives have not been explicitly articulated. In fact, for most researchers, the choice of perspective is likely made unconsciously. Nevertheless, maintaining an accurate theoretical perspective is vital to the experimental dissection, synthesis, and interpretation processes. An inaccurate perspective can lead to research bias and general misalignment of evidence.

Two of the most common theoretical perspectives adopted by researchers in fields that straddle both technical and social domains are determinism and social constructivism.

2.1.1 Determinism

The deterministic approach to managing complexity depends on the belief that the phenomena of interest exist independent of human factors; if the phenomena do intersect with people, then they are only subject to the most basic mechanical forces inherent in the human condition, which are, for the most part, beyond the control of conscious thought. Under this assumption, determinists focus on modeling finite relationships between measurable variables, the underlying goal of which is to understand the cause-effect
interactions at play in a system. Understanding cause-effect interactions enables system administrators to optimally configure the system [29, 32].

Not surprisingly, computer scientists consistently espouse a deterministic perspective in their research. However, since computer science is predominated by highly technical problems which are generally sterile of human contact, the deterministic perspective is fairly effective. Modern software development, on the other hand, is a team activity, deeply embedded in social structure and human consciousness. Despite that fact, SE researchers often turn to a deterministic perspective when confronted with the complexity that arises from the social aspects of software development. Responding to this tendency toward a deterministic view of research, Leveson stated, “[W]e need to avoid equating humans with machines and ignoring the cognitive and human aspects of our field” [26, p. 9].

2.1.2 Social Constructivism

From a socially constructed perspective, phenomena are said to be produced in consequence of human decisions and social mechanisms. Therefore, social constructivists seek to explain phenomena through an interaction of human factors. Creating predictive theories requires understanding the motivations, objectives, interests, limitations, and interactions of the human participants [32, 45].

Since most SE researchers come from highly technical backgrounds, as a field, SE is partially blind to the socially constructed perspective of software development. In making that statement, however, we do not want to denigrate the excellent work that many of our colleagues undertake to study social and cognitive factors. Nevertheless, as a discipline, our knowledge and skills in sociological and psychological research are still immature, and the tools that we have borrowed from other disciplines have not yet fully disseminated within our science. Despite these difficulties, we are making progress. The most difficult step to take is, in fact, the first—to see the world in a new way.

2.2 An Enacted View of SE Complexity

Both determinism and social constructivism are applicable to SE—each contributes fundamental insight into the mechanisms governing software development environments and project ecosystems. However, the two perspectives are diametrically opposed, and as such, they are each only able to capture particular facets of “the big picture.” Since both views oversimplify SE phenomena, we propose adapting and incorporating into SE a theoretical perspective originally formulated by Orlikowski and Iacono for studying “the digital economy,” termed an enacted view [32].

An enacted view of SE incorporates both deterministic and socially constructed forces into the same paradigm. This perspective maintains that software production is neither a balancing of technical and logistical trade-offs, nor is it simply a matter of managing people. In the following quote by Orlikowski and Iacono we substitute “software engineering” in place of “the digital economy” and note the profound applicability of the enacted view in examining SE:

This view suggests that [software engineering] is neither an exogenous nor a completely controllable phenomenon, but an ongoing social product, shaped and produced by humans and organizations, and having both intended and unintended consequences. It is our individual and institutional actions in developing, constructing, funding, using, regulating, managing, supporting, amplifying, and modifying the phenomenon we refer to as [software engineering] that enacts it over time. [32, pp. 357–358, italics added]

Assertion 1. SE represents the culmination of a rich array of interactions between technical, behavioral, and social forces, and only when considering it from that full perspective will an accurate “big picture” begin to emerge from the complexity.

3. THE NEED TO SIMPLIFY: COGNITIVE, SCIENTIFIC, AND PRACTICAL

Innately, we all have a need to simplify the world around us. At its most fundamental level, the need to simplify is motivated by core cognitive processes. These processes seek to categorize the physical and abstract elements of our environments into generalized schemas, which are critical to the storage and retrieval of information in memory. The categorization mechanism has been shown, for instance, to be a driving force behind social stereotyping, in which we mentally classify each other based on prominent and identifiable features, such as race, gender, and social status [3].

Research in organizational behavior has also observed that, as participants in an organization, we generalize strategic issues into a hierarchy of opportunities and threats [14]. Thus as humans, we maintain numerous taxonomies by which we organize the elements of our environments.

In addition to taxonomies, which represent elements and element-relationships, human cognition also relies on operational rules to govern the interactions between the elements in our numerous taxonomies [43]. These rule sets, which enable us to make judgments about the world (e.g., to predict the consequences of adding manpower to a late project), are generalizations of cascades of interactions that occur between elements of the real environment.

As previously discussed, SE researchers sometimes oversimplify software environments—for instance, by espousing a deterministic perspective of software production mechanisms. In a notable case, recent evidence [7] suggests that Eric Raymond’s The Cathedral and the Bazaar [35], oversimplifies the organizational structures of productive open source software projects. However, generalization is not necessarily a detriment to science. In fact, scientific inquiry relies on the distillation of information in order to be productive. Since knowledge is an abstract representation of elements and the relationships between those elements, building knowledge requires constructing taxonomies and rule sets. Science, therefore, is designed to be a conscious and formalized embodiment of the natural cognitive processes by which we come to “know” the world, and as such, relies on the digestion and generalization of environmental complexity. Indeed, the need for generalization in science is reflected throughout the meta-theoretical principles on which it operates—consider, for example, Occam’s razor.

In addition to cognitive and scientific needs for generalization, daily life also depends on simplified taxonomies and rule sets. Day-to-day living generally requires us to make judgments quickly—for instance, do I walk or drive to work today? In the practical application of SE we often have more time to make strategic organizational decisions than when
we’re late for work, but those decisions still must be made in relatively short time frames. Simplified environmental models are the mechanism by which we make efficient decisions.

Assertion 2. The abstract models by which we interpret and manage environmental complexity must be simplified on at least three levels: cognitive, scientific, and practical.

3.1 The Problem of Premature Generalization

As Leveson explains, intuition is a necessary component of formulating hypotheses [26]. If intuition is an informal theoretical synthesis of prior observations, then hypotheses are the formalized, testable embodiments of that intuition. However, intuition can be misleading, and thus hypotheses must be tested. As such, Leveson boldly states that we “need to recognize the unproven assumptions and hypotheses underlying our current software engineering techniques and tools and evaluate them in the context of what has actually been demonstrated about these hypotheses instead of what we would like to believe” [26, p. 8].

In response to this and similar arguments, SE researchers have conducted numerous empirical studies. However, Basili et al. [6]—and more recently Jørgensen and Sjøberg [19]—caution that the expense of new proposals extends far beyond the breadth and depth of our theories. With so much data now in massive repositories and with computer networks bringing it to our fingertips, it’s no wonder that published research is often reduced to a sophisticated game of trivial pursuit. At a certain point, we risk drowning in a sea of “new proposals” and unsynthesized results. Because our studies have become highly disconnected [19, 42]—and since clearly no single study has the independent power to produce definitive results [19, 20, 38, 42]—by definition, all generalizations we can make will be shallow and immature. Without greater research synthesis, therefore, we will continue to ignore the most interesting phenomena of our field, and our theories will remain shallow and narrow in scope.

Assertion 3. Disconnected research necessarily leads to premature generalization and narrow, immature theories.

4. NO SILVER METHOD

According to Popper [34], theories are reliable only after researchers have made numerous attempts to falsify them. Thus, a theory must ultimately be grounded in observation, or it is of little value. Conversely, empirical—that is, observational, experiential, or experimental—work is useless unless the resulting observations are synthesized into governing theories. Empirical and theoretical methods are therefore complementary. Theoretical work involves logic, deduction, inferences, and thought experiments, whereas empirical work requires observation and, when formalized, measurement (qualitative and/or quantitative). In response to the question, “How will we build our [scientific] foundation?” Leveson states, “It will require both building mathematical models and theories and performing carefully-designed experiments” [26, p. 7]. Thus both theoretical and empirical methods are necessary to capture and dissect the complex problems of SE [4, 6, 19, 22, 23, 42].

Empiricism comprises numerous research methods (e.g., controlled studies, ethnographic studies, and data mining and analysis), which can be grouped into two categories: observational and experimental. Observational methods are conducted “in the wild”—that is, the phenomena are observed in their natural context. Experimental methods, on the other hand, examine phenomena under controlled conditions, in which the researcher selects subjects to receive “treatments” and then observes the consequences. If properly designed, experimental methods enable the researcher to conclude with a degree of confidence that the variation in the applied treatment caused the observed variation in the resulting experimental groups. The key difference, therefore, between observational and experimental methods is whether the researcher controls the context of the study.

If experimental methods enable researchers to make causal inferences, then why perform observational studies at all? In SE the most common answer to this question is that SE experiments are expensive or impractical to conduct. Although that response is generally accurate, it risks mischaracterizing observational methods.

The goal of empirical work is to observe both the elements of a system and the relationships between those elements in order to build an understanding of the dynamics of that system. Creating an accurate and complete understanding of a system requires that two conditions be met: 1) the understanding must capture and incorporate all elements and element-relationships in the system, and 2) the understanding must completely dissect and digest those elements into their constituent parts. We have already observed that experimental methods are more useful than observational methods for dissecting complexity. However, experimental methods are limited in their scope, since they cannot explicitly address all important variables, many of which are not readily measurable. Experimental science relies on random assignment of subjects to treatment groups to compensate for missing information. If the samples are large enough and truly random, then cause-effect relationships can be confidently identified—but only those relationships for which the researcher is specifically testing. Any elements not explicitly identified and designed into the experiment cannot be objectively dissected from the system. Thus at the extremes, research methods either fail to capture the full complexity of the system, or they fail to objectively dissect that complexity. In reality, the more complexity a method preserves, the less power it has to dissect that complexity.

Consider, for example, case studies The primary criticism levied against case studies is that they tend to be anecdotal. How is one to know that a researcher’s intuition is accurate? To this effect, Basili, Shull, and Lanubile state, “Experimentation in software engineering is necessary. Common wisdom, intuition, speculation, and proofs of concepts are not reliable sources of credible knowledge” [6, p. 456]. We agree with this criticism, but note that it simply echoes the limitation of observational methods, that they are unable to objectively dissect complexity.

Utilizing experimental methods, however, we still fail to satisfactorily generalize our results. Whenever the “it depends on context” qualification is applied to conclusions, then the methodology has abstracted away complexity in favor of dissection. In order to meet both requirements for accurately understanding a system, breadth as well as depth, we must learn how to effectively blend diverse research methods into synthesized analyses. As Frederick Brooks might say, there is no “silver bullet”—or method [10].

Assertion 4. No single research method can fully reveal the complex mechanics of SE.
4.1 The Multi-Method Approach to SE: Embodies an enacted view

At present, the best solution to the conflicting limitations of observational and experimental research methods is, as Daly terms it [12], the multi-method approach. According to Daly, the “multi-method approach involves using two or more different empirical techniques to investigate the same phenomenon” [12, p. 65]. Daly notes that corroborating results (across multiple differentiated studies) provides confirmatory power and represents a form of validation. However, we propose to extend Daly’s concept of the multi-method approach by recognizing that its most fundamental and powerful contribution, when applied consistently, is to knowledge building. In fact, Daly hints at this contribution when discussing an evolutionary embodiment of the approach, in which researchers leverage the strengths of various research methods at specific stages in a research program in order to incrementally build their understanding. Each successive stage is designed around the knowledge gained from the previous stages—thus the preliminary results not only impact the final conclusions, but they configure the overall methodology as the study proceeds.

In sections 5 and 6 of this paper, we discuss the power of fundamental patterns and employ the multi-method approach as a fundamental pattern under which we synthesize a unified methodology for SE research. In the process of elucidating the overall research framework, we dissect the current literature on replication methods and place replication within that research framework. As a consequence of these efforts, we also further clarify, from a high level, the mechanisms identified by Basili et al. [6] through which science builds usable knowledge.

5. PATTERNS

Although finding interesting phenomena is relatively easy in SE, examining them deeply [37] and placing them appropriately within the context of a plethora of other interacting elements is as difficult as ever. Sometimes, amid the chaos of interactions, finding useful generalizations seems impossible. The difficulty arises from the complexity of SE environments and ecosystems, which are enacted over time by technical, behavioral, and social forces and are therefore continually evolving. Drawing again from Orlikowski and Iacono’s work on “the digital economy,” we suggest that accurately generalizing SE complexity requires a shift in perspective to fully embrace an enacted view, and to embody that view in our research. Using the words of Orlikowski and Iacono,

[Software] is a phenomenon that is embedded in a variety of different social and temporal contexts, that relies on an evolving technological infrastructure, and whose uses are multiple and emergent. As a result, research studies will yield not precise predictions—because that is not possible in an unprecedented and enacted world—but underlying patterns . . . Similarly, research studies will offer not crisp prescriptions—because these

are unhelpful in a dynamic, variable, and emergent world—but general principles to guide our ongoing and collective shaping of [software]. [32, p. 375]

Thus, we need to structure research methods in SE to enable us to discover the fundamental patterns underlying and interconnecting the observable phenomena. But what are fundamental patterns? And how do we distinguish universal truths from shallow, context-specific poses? Our discussion leads us now to Christopher Alexander’s “The Timeless Way of Building” [1] and “A Pattern Language” [2], and from thence to Thomas Kuhn’s “The Structure of Scientific Revolutions” [25].

5.1 In Search of “Double-Star” Patterns

In Christopher Alexander’s original work on architectural patterns [2], he identifies three broad classes of patterns: Patterns associated with two asterisks are those that the authors believe to have captured a true invariant, “that the solution we have stated summarizes a property common to all possible ways of solving the stated problem,” and “that it is not possible to solve the stated problem properly, without shaping the environment in one way or another according to the pattern that we have given—and that, in these cases, the pattern describes a deep and inescapable property of a well-formed environment.” Patterns with one asterisk are those believed to “have made some progress towards identifying such an invariant.” Those with no asterisks “have not succeeded in defining a true invariant,” meaning that their pattern represents one particular solution, but does not capture the essence of the problem. [2, p. xiv]

Essential invariants involve both the structure of the architectural element in question, as well as the way in which such an element interacts with human behavior. As you read the following architectural example from Alexander, consider the implications to SE:

In New York, a sidewalk is mainly a place for walking, jostling, moving fast. And by comparison, in Jamaica, or India, a sidewalk is a place to sit, to talk, perhaps to play music, even to sleep. It is not correct to interpret this by saying that the two sidewalks are the same. . . . Each sidewalk is a unitary system, which includes both the field of geometrical relationships which define its concrete geometry, and the field of human actions and events, which are associated with it. So when we see that a sidewalk in Bombay is used by people sleeping, or for parking cars . . . and that in New York it is used only for walking—we cannot interpret this correctly as a single sidewalk pattern, with two different uses. The Bombay sidewalk (space + events) is one pattern; the New York sidewalk (space + events) is another pattern. They are two entirely different patterns. [1, p. 73]

Alexander’s notion of double-star patterns embodies the idea of context-free solutions, principles and patterns that are not invalidated by an “it depends” clause. If the solution to a problem is, “it depends,” then the applicable patterns must, per force be narrowed to the context of the specific
problem. As researchers we err when we stretch our narrow findings so broadly that we misperceive a solution in a specific context as having captured a broader principle. Although there is nothing wrong with context-specific solutions to context-specific problems, in order to build knowledge, we must understand the broader, higher-level patterns that operate. Again, an example from Alexander:

Since every church is different, the so-called element we call “church” is not constant at all. Giving it a name only deepens the puzzle. If every church is different, what is it that remains the same, from church to church that we call “church”? [1, p. 84–85]

Such a discussion brings to mind broad-brush appellations such as “open source” versus “closed source” software development processes, as if we really understand what each expression means or why exactly they matter when it comes to engineers cutting code. “A pattern only works, fully, when it deals with all the forces that are actually present in the situation” [1, p. 285].

Our experience suggests that the quest for fundamental, Alexandrian double-star patterns provides a mechanism for encapsulating chunks of knowledge at various levels of abstraction (from narrow and specific to broad and general), as well as a mechanism whereby meta-level patterns may be discovered and understood that describe the generation and interaction of lower-level patterns [24]. Such tools provide the means for better understanding the scientific models that always underlie context-specific empirical research, but which are often either unidentified or unrecognized.

**Assertion 5.** *Scientific models are more generalizable when based upon fundamental (Alexandrian) patterns; therefore, it is the job of SE researchers to discover and develop fundamental context-free patterns from which practitioners can compose an endless variety of software organizations and practices, each to meet specific, context-sensitive needs.*

### 5.2 Seeing “The Big Picture”

Finding fundamental patterns requires seeing and understanding, within the same context, a broad range of empirical observations. As any puzzle enthusiast knows, it is only when viewing the pieces in context with one another, correctly fitted, that the overall picture becomes obvious.

In SE we have come to broadly recognize the need for empirical methodologies in order to build knowledge (hypotheses are based on intuition, and so must be tested through observation). However, it seems that our efforts to become empiricists have lead us away from deep observational synthesis and theory building—that is, we have oversimplified our methodologies. Indeed, for many of the most important questions in SE, we have failed to establish any relevant formalized theories [19]. Consequently, our studies have become disconnected, thus leading to shallow, narrow conclusions that do not generalize well [19, 42]. It is unfortunate that while most academic papers include an empirical study of some sort, we find almost no papers that are purely (or even primarily) theoretical. Where are the papers synthesizing these empirical studies into a unified, cohesive whole? We agree with Jørgensen and Sjøberg who caution that “there are much too few review papers in software engineering trying to summarize relevant research” and “[w]ithout a much stronger focus on [the] theory-development step, we probably will continue to produce isolated, exploratory studies with limited ability to aggregate knowledge” [19, p. 33].

In addition to fragmented observations and premature generalization, a lack of theory building in science leads to an even more subtle, degenerative problem. According to Kuhn, mature scientific disciplines pass through periods of intense community focus, knit together by deeply shared beliefs about how the world works (i.e., paradigms). These periods of focus are followed by periods of broad self-assessment and theoretical upheaval (i.e., scientific revolutions) [25]. Therefore, science is a two-fold process of evolving theories based on a continual stream of new observations, followed by replacing theories when they inevitably outlive their usefulness. In the words of Greenberg, we must allow for periods of both “getting the design right” (science under a paradigm) and “getting the right design” (scientific revolution) [17, p. 115]. The two are separate and distinct activities, and in the absence of either, scientific inquiry is severely handicapped.

Since theories are the embodiment of a way of thinking, and paradigms form around a common way of thinking, then without theory building, paradigms of broad reach and significant impact cannot occur. As a result, the research community is not cohesive, communication between researchers is difficult, and research findings are disjoint and disconnected. Thus, theory building is necessary for establishing paradigms, and paradigms are vitally important to building knowledge. Without theory building, SE will struggle to maintain deep community focus on a particular set of ideas [6, 25, 42].

To this point, we have identified three issues that prevent researchers from seeing “the big picture”—fragmented observations, premature generalization, and an inability to sufficiently focus community efforts—and have hinted that theory building helps to resolve all three. However, we further observe that theory building is not only necessary to discover fundamental patterns, but it is also not sufficient. Theory building is the glue that binds observations together into a cohesive and meaningful context (without which empirical work remains fragmented and context-specific), and that binds a community of researchers together so that theoretical proposals can be deeply probed for limitations. Nevertheless, theories rely on observation for validation and evolution, and so theory building without a rigorous empirical counterpart generally produces beautifully useless “knowledge.” Therefore, the presence of both theoretical [19, 40, 41, 42] and empirical [15, 22, 23, 33, 42, 44] methodologies—with an effective dialogue between—is necessary in any science in order to discover fundamental patterns.

**Assertion 6.** *Discovering fundamental patterns requires seeing “the big picture,” which is accomplished in science by an alternating process of observation and theoretical synthesis.*

### 5.3 A Terminological Aside

Before we proceed with the discussion at hand, we must clarify some terminology relative to replication methods in SE. In the literature, researchers have employed various terms to describe replication studies, including: *exact versus non-exact* [20, 28, 39, 9, 20, 28], *close* [20], *conceptual* [39], and *literal versus theoretical* [28].

In selecting our terminology, we consider the two fundamental goals of replication, as outlined by Shull, Carver,
5.4 Replication: The Key to the Knowledge Building Process

As discussed earlier, the complexity of SE necessitates a multi-method approach to research, which, by definition, requires multiple studies to repeatedly examine the same questions and hypotheses. The idea is (as Basili et al. describe, in terms of parsimony [6]) to tackle threats to validity through many studies, each with different threats, rather than trying to hopelessly control all threats within a single study. If we accept the “parsimony” approach to research, then the alternating process of observation and theoretical synthesis ultimately lies in the domain of replication.

However, replication has traditionally been defined in the literature in terms of strict replication, and so researchers have invested significant resources attempting to exactly replicate studies—a practice which has thus far proven to be infeasible in SE. For this reason, Juristo and Vegas suggest that researchers in SE should relax the demand for strict replication, suggesting that “opening the door to non-identical replications could in actual fact encourage researchers to do more replications . . . and, at the same time, [turn] up new knowledge” [20, pp. 365–366]. In fact, since it is unlikely that a replication can be exact, strict replications generally end up being treated like differentiated replications anyway, except that traceability is more difficult because the changes are unplanned.

Although strict replication is important for validating observations—that is, confirming that we accurately understand the conditions under which a set of observations occur—it does not test theory, and therefore, it does not build usable (i.e., practical) knowledge. In making this claim, we do not argue that attempting strict replication cannot uncover new information which can in turn be used to build knowledge. We simply observe that strict replication is designed to verify that a study’s reported conditions do in fact produce the reported observations. Testing the correlation between conditions and observations to detect experimental flaws or fabrication is not the same as testing the validity of theoretical axioms. Differentiated replication is the mechanism that tests the limits of theoretical axioms by altering the conditions or the context of prior investigations. Strict replication simply tells the researcher whether or not the latest observations should be accepted and allowed to affect the current theory. As an example, consider the work of Lung, Aranda, Easterbrook, and Wilson, who after performing a strict replication state, “On reflection, the literal replication was much more complicated than we expected, and told us very little about the underlying theory. On the plus side, we identified a number of flaws in the researcher’s experimental design, which we were able to correct” [28, p. 200]. Therefore, we highlight a somewhat subtle distinction between the type of knowledge created by the practice of validating observations versus that of refining theories. The former practice generates knowledge useful only to the researcher/theoretician, whereas the latter generates knowledge for building theories, and is, consequently, ultimately useful to the practitioner. In case these two assertions appear to be reversed, we further note that in mature scientific domains, practitioners operate on well-established theories, not the results of individual studies.

**Assertion 7:** Strict replication is the process by which researchers test a study’s methods and procedures to validate that the reported conditions produce the reported observations.

**Assertion 8:** Differentiated replication, founded on the multi-method approach to research, is an embodiment of the alternating process of observation and theoretical synthesis, and is, therefore, the process by which theoretical paradigms are created, evolved, broken, and replaced; consequently, differentiated replication is the primary mechanism by which fun-
damental patterns can be discovered and is the key to making the knowledge building process productive.

6. THE CYCLE OF MATURING KNOWLEDGE

Having laid the necessary groundwork, in this section we present a unified framework for research methodologies in SE, which we refer to as the Cycle of Maturing Knowledge (CMK). We first summarize the philosophical constructs presented thus far, after which we discuss the CMK in detail.

In SE we struggle to produce usable knowledge that generalizes across software environments. We struggle in part because we have not yet developed sufficient methods to manage the interacting technical, behavioral, and social complexities of modern software development. Currently, our empirical studies are disconnected and our theories are generally shallow. In order to overcome the problem of premature generalization, we must embrace more fully the complexity of SE. Guided by an appropriate theoretical perspective (e.g., an enacted view), and through an alternating process of observation and theoretical synthesis, we can learn how to perceive “the big picture” amidst the complexity. With the requisite community focus, we can further discover fundamental patterns, the essence of which must form the structure of our theories. Based on fundamental patterns, our theories will be more generalizable, enabling us to distill usable knowledge.

The CMK (depicted in Figure 2) incorporates each of the elements discussed in this paper and represents a SE-specific adaptation of the general knowledge building process discussed in the literature [6]. The process begins with a common experience: an epiphany or insight gained from preliminary observations of some set of phenomena. This experience generally progresses through four phases: 1) we observe an aspect of the world; 2) our observation triggers a new insight; 3) we begin “seeing” the concept embodied by our insight everywhere; and 4) as we revisit and re-experience the concept in different contexts, we refine our understanding of it. In the spirit of Alexander, these “ah-ha” moments, followed by the iterative process portrayed here, represent a fundamental pattern of learning, which the CMK attempts to model. Therefore, an initial observation (or set of observations), followed by initial insight and initial questions, leads us to a deeper and more methodical analysis.

Within the cycle, each of the boxed items represents an element that inevitably occurs, independent of whether it is formally or consciously expressed. For instance, a researcher cannot perform an empirical study without considering at some level the possible outcomes of that study. Accordingly, a researcher cannot help but ponder the implications of new observations. Each of the boxed items, therefore, expresses a feature of the natural human pattern of learning, by which we develop simplified models of complex environments. Conversely, the bracketed items represent features that require formalized methods and conscious effort if they are to occur.\(^2\)

Examining the CMK, it becomes clear that strict replication is one method for validating empirical observations, and although essential to the cycle, it serves a supporting role (rather than a fundamental or preeminent role) in the process of building knowledge. Conversely, differentiated replication is the primary mechanism driving the knowledge building process, without which the cycle ceases to productively generate new theories and underlying conceptual foundations that may then be formally analyzed.

\(^2\)Note that we use the terms internal and external in this diagram in the same context as Brooks et al. [9], to denote validation that is conducted by the original researchers as opposed to a third party. This usage differs from that of Campbell and Stanley, who use the terms to talk about the degree of confidence in cause-effect conclusions, as opposed to the generalizability of results [11].
In producing the CMK, we explicitly incorporate the multi-method research approach [12], which stipulates that differentiated replications must be performed using diverse methods in order to prevent methodology bias [17]. The CMK also expresses key elements of Basili, Shull, and Lanubile’s knowledge-building framework, “Families of Experiments” [6], which describes the need to iterate the cycle of analysis numerous times—thus performing a “family” of related differentiated replications—in order to study a particular question or theory.

Basili et al. also describe methods for synthesizing replications [6]. Synthesis methods are represented in the center of the cycle, and incorporate the processes by which usable knowledge is distilled from a “family” of differentiated replications. Juristo and Vegas have contributed significantly to these processes [20]. However, only recently have we begun to recognize the importance of differentiated replication (as opposed to strict replication) in the process of knowledge building. These processes of synthesis are, therefore, still immature and demand further analysis to refine appropriate and effective methods.

**Assertion 9.** *The Cycle of Maturing Knowledge unifies the current SE research methodologies and represents a SE-specific embodiment of the general knowledge building process.*

7. **CONCLUSIONS**

In an attempt to clarify research methodologies in SE, we have synthesized the contributions of numerous scientists and researchers. We are indebted to them for their conceptual, theoretical, and philosophical contributions derived from extensive experience with empirical research.

From this synthesis process, we identify two general principles: 1) In order to discover fundamental patterns on which to build generalizable theories, SE must embrace and account for its true complexity, rather than ignore or abstract it away. 2) In order to avoid premature generalization in the search for fundamental patterns, SE must develop and motivate the use of processes that appropriately and effectively blend theoretical and empirical work. These processes must be founded upon deep synthesis of methodologically diverse studies.

As a reflection of these principles, the CMK identifies three areas for process improvement in SE research: 1) In addition to strict replication, we must explore additional methods for internal and external validation of observations. 2) We need to continue developing methods for conducting differentiated replications, such that results can be synthesized (e.g., improving traceability [20]). 3) We must spend more time synthesizing—that is, iterating the knowledge-building cycle, conducting “families” of differentiated replications [6], and unifying our observations through theory building.

Alas, the search space is vast and sometimes feels daunting. We resonate with the anonymous author who wrote the following inspired summary:

> We have not succeeded in answering all our problems. The answers we have found only serve to raise a whole set of new questions. In some ways we feel we are as confused as ever, but we believe we are confused on a higher level, and about more important things.

8. **REFERENCES**


[15] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. Selecting empirical


