修士学位論文

題目

ファンクションポイント研究の近年動向と有効性評価に関する系統的レビュー ―2019－2025 年の文献分析―

指導教員

楠本 真二 教授

報告者

WU ZIJING

令和 8 年 2 月 1 日

大阪大学 大学院情報科学研究科

コンピュータサイエンス専攻

令和 7 年度 修士学位論文

ファンクションポイント研究の近年動向と有効性評価に関する
系統的レビュー —2019 – 2025 年の文献分析—

WU ZIJING

内容梗概

ソフトウェア開発における見積もりは，計画立案や資源配分の根拠となる重要な活動であり，その前提として開発規模を適切に把握する必要がある．ファンクションポイント法（FPA）は，実装技術に依存せずユーザ視点で機能規模を定量化できる尺度として広く利用されてきた一方で，計測に専門的判断と工数を要し，結果のばらつきや適用条件の不透明さが課題として指摘されている．さらに近年は AI・機械学習の活用が進み，FP が見積りや意思決定の入力情報として再評価される場面も増えているため，研究成果の全体像と実務適用可能性を改めて整理する必要がある．

本研究では，2019 - 2025 年に公表された FPA 関連研究を対象に系統的文献レビュー（SLR）を実施し，研究トピックの分布と位置づけ変化，研究の成熟度（進展段階），提案手法の有効性評価の厳密性，ならびに 2018 年時点で指摘された未解決課題の解消状況を整理した．その結果，研究関心は「計測補助」および「計測結果の活用」に集中し，提案や実装が増加している一方で，業務プロセスへの定常統合が明示される実証適用は限定的であり，提案・実装から運用定着への移行にギャップが残ることが確認された．また，有効性評価は広く実施されているものの，データ・比較・統計の整備にばらつきがあり，とりわけ再現性が顕著に不足していた．さらに，コンテキストの記載不足と教育（普及）手法の不在は依然として未解消であり，FP で計測しにくい要素への対応も，本レビュー範囲では体系的な蓄積が乏しいことが確認された．

以上の結果を踏まえ，本研究では，近年の FPA 研究が示す傾向と課題を横断的に考察し，知見の蓄積が進む領域とエビデンスの不足が課題として残る領域を対比しつつ，研究と実務の接続に向けた論点を整理した．加えて，知見として累積させるために，優先順位と依存関係を明確にした統合的ロードマップを提示した．

主な用語

ファンクションポイント法，系統的レビュー，ソフトウェア見積もり，ソフトウェア規模計測

目次

図目次

表目次

1　はじめに

ソフトウェア開発における見積もりは，計画立案や資源配分，コスト管理の根拠となる重要な活動である [1]．見積もりの前提として，開発対象規模の適切な把握が求められるが，その規模を定量化する手法の一つにファンクションポイント（Function Point; FP）法がある [2]．FP 法は 1979 年に Albrecht によって提案された手法であり，ソフトウェアが提供する機能に着目して開発規模を測定する手法である．ユーザ視点で機能を数値化できるため開発者と利用者の合意形成に役立ち，また実装技術や開発言語に左右されにくい汎用的な尺度であることから，FP 法は現在も実務上重要な位置を占めている．実際に情報システム開発では，要求内容に設計または開発工程が含まれる場合に FP に基づく見積もりが求められる場面が増えている [3, 4]．

一方で，FP には計測者の判断に依存する工程が含まれるため，計測結果の一貫性や信頼性に懸念が指摘されている [5]．また，FP の計測作業自体に時間と労力を要することも課題として指摘されてきた [6]．このような課題に対して，これまでに計測結果の信頼性評価，計測作業の支援・自動化，計測された規模と工数・生産性等との関係分析など，多様な研究が蓄積されている [7, 8]．

こうした研究蓄積を俯瞰する試みとして，山田ら [9] は，1979 - 2018 年に公表された FP 関連研究を対象に系統的文献レビューを実施し，研究を複数のトピックに整理するとともに，研究の重心が FP の利点評価や活用方法から，計測補助を志向した研究へと移行していることを示した．また，同レビューでは，開発言語・規模・業種といったコンテキスト報告の不足や，教育・普及手法の体系化が不十分である点など，実務適用を妨げうる未解決課題も指摘されている．

しかし近年では，FP 研究を取り巻く技術・実務環境そのものが変化しつつある．具体的には，見積もりの場面で AI や機械学習が広く用いられるようになり，FP が「人手で算定する成果物」であるだけでなく，「学習・予測モデルへの入力特徴量」として扱われるケースが増えている．すなわち，FP は「算定結果」から「推定プロセスを構成する入力情報」へと役割が拡張している．その結果，FP に関する研究課題も，計測手順の改善や自動化にとどまらず，AI や機械学習を用いる見積もりプロセスの中で FP をどのように位置づけ，どのような条件下で有効に利用できるかへと関心が広がっている．

このような変化を踏まえると，2018 年までの先行研究 [9] で得られた結論や指摘事項を単に踏襲するだけでは不十分である．先行研究で示された傾向や課題が，2019 年以降の研究でも維持されているのか，あるいは新たな前提の下で変質しているのかを明確にする必要がある．

さらに，研究対象が多岐にわたり論文数も増加しているため，新規参入の研究者や実務者が当該分野の全体像を把握することは容易ではない [3]．加えて，FP 法が現在も見積もり実務で使用される以上，研究で得られた知見が現場で適用可能かどうかを検討することが重要である．とりわけ，提案手法がどの程度の根拠で検証されているかが見えにくい，すなわち実証的研究や評価の厳密性が一般に不足して

いるという指摘 [9] もある．そのため，実務者（新しい手法・ツールの導入を検討する担当者）にとっては，提案手法・ツールを自組織の見積り業務に導入すべきか（適用可否，前提条件，期待効果，必要コスト・工数）といった採用判断の材料が得にくい状況が生じうる．

特に，評価データの性質や比較対象，統計的検証，再現に必要な情報が十分に報告されない場合，提案の有効性を客観的に判断しにくい．この点は，研究と実務のニーズの乖離が指摘されるソフトウェア工学分野の課題とも整合し，FP 法の研究分野においても同様のギャップが存在するかを確認し，現場適用の観点から知見を整理し直す必要がある．

以上の背景を踏まえ，本研究では FP 法に関する近年の研究を体系的に整理し，トピックの変化と研究の成熟度（進展段階），および有効性評価の厳密性と未解決課題の解消状況を明らかにすることを目的として，系統的文献レビュー（Systematic Literature Review; SLR）を実施する [10]．SLR は，網羅性と再現性を重視して文献を収集・選別・統合する手法であり，定式化した問いに対して既存研究に基づき客観的な回答を導出するために用いられる．本研究では 2018 年までの先行研究で整理されたトピック構造と指摘事項を出発点とし，2019 年以降に公表された FP 関連研究を対象として，実務における適用可能性および信頼性，ならびに先行研究で示された課題の解消状況を検証する．

本論文の構成を述べる．第 2 節では，本研究の前提となるソフトウェア見積もり論およびソフトウェア規模計測手法の基礎知識を整理し，特に FP 法の特性と手順について詳説する．また，SLR 手法および関連する先行研究の知見を概観する．第 3 節では，SLR の具体的な手順として文献の検索・選別・情報抽出の方法を詳細に説明する．第 4 節では，収集した文献の分析結果に基づき，各 RQ に対する回答を示す．第 5 節では，得られた知見について考察し，研究動向の含意や残存課題を議論する．第 6 節では，本研究の結論をまとめ，研究の限界を示した上で，今後の課題と展望を述べる．

2　準備

2.1　ソフトウェア見積もり

ソフトウェア見積もりは，開発計画の立案，コスト管理，資源配分および進捗管理の基盤となる重要な活動である．見積もりの不確実性が高い場合，納期遅延や予算超過を引き起こし得るため，定量的かつ説明可能な根拠に基づく見積もり手法が求められる．また，見積もりの入力として用いる「規模」の定義と計測方法が適切でなければ，見積もり結果の妥当性に大きく影響する [11]．したがって，ソフトウェア見積もりを論じる上で，規模計測の代表的手法とその特性を整理しておくことが不可欠である．

2.2　ソフトウェア規模計測

ソフトウェアの規模計測には，ソースコード行数（Source Lines of Code; SLOC），ファンクションポイント（Function Point; FP），ユースケースポイント（Use Case Points; UCP）など複数の手法が存在し，さらに機械学習などのアルゴリズムと組み合わせたアプローチも提案されている [12, 13]．

SLOC はソースコードの行数に基づいて規模を見積もる手法であるが，実装前の上流工程では適用が難しく，またプログラミング言語やコーディング規約に依存しやすいため，開発環境が異なると結果が一貫しないという課題がある．

一方，FP 法は実装される機能に着目して規模を算定する枠組みであり，特定の開発プロセスや技術に対する依存度が低い [2]．このため，FP 法は要件定義など上流工程から詳細設計・実装後まで複数の段階で適用でき，実務で広く用いられてきた．

UCP は FP 法を拡張する形で Karner により提案された手法で，抽出したユースケースとアクタを複雑度ごとに分類してポイントを算出する [14]．オブジェクト指向開発プロジェクトで広く利用されているが，詳細なユースケース記述が整うのは開発後期になることも多く，上流段階での適用は難しいと指摘されている．

以上のように，ソフトウェア規模の見積もり手法は多種多様であり，上流工程への適用可能性，技術への依存度，および説明性・客観性といった観点でそれぞれ異なる特徴を持つ．本研究では，機能に基づいて技術非依存的に規模を評価できる点に着目し，FP 法を中心に議論を進める．そこで，次節では FP 法の標準化動向と代表的な手法について整理する．

2.3　ファンクションポイント法

FP 法（Function Point Analysis; FPA）は，ソフトウェアが持つ機能量を定量化するための手法であり，比較的成熟した実用的手法として広く普及している．

なお，FP 法の国際標準化は，当初，単一方式としての統一を目指して検討が開始されたが，地域ごと

表 1: ISO/IEC における主なファンクションポイント関連手法規格

手法規格	手法	発行年
ISO/IEC 20926	IFPUG 法	2009
ISO/IEC 20968	Mk II 法	2002
ISO/IEC 24570	NESMA 法	2005
ISO/IEC 19761	COSMIC	2011

に複数手法が定着していたことや，手法間で測定方法の差異が大きいことなどから，単一方式への統合は困難であると整理された．そこで ISO/IEC では，個別手法を直接統一するのではなく，まず「機能規模」および「機能規模測定」に関する概念と用語，ならびに機能規模測定手法が満たすべき要求事項を定める概念規格（ISO/IEC 14143-1）を整備し，複数手法の併存を許容する共通基盤を確立した．さらに，概念規格に基づく手法の適合性評価や検証，参照モデル，機能領域の定義，利用上の指針といった周辺要素を支援規格（ISO/IEC 14143-2〜6）として段階的に整備し，手法規格を策定・維持するための枠組みを補強した．

その上で, この概念規格･支援規格の枠組みに準拠することを前提として, 各 FP 派生手法が ISO/IEC [15, 16, 17, 18] により手法規格として国際標準化される流れが形成されている．代表例として IFPUG Function Points（IFPUG）[19]，Mark II Function Points（Mark II）[20]，NESMA Function Points（NESMA）[21]，および COSMIC-FFP（COSMIC）[22] が挙げられる．これらの派生手法が併存する理由は，(i) 何を「機能」とみなして数えるか，(ii) どの成果物・どの開発段階を入力として想定するか，(iii) どの種別のシステムやプロジェクト（新規開発・保守・拡張）に適合させるか，といった前提が一様ではないためである．例えば，IFPUG や NESMA は，データ機能とトランザクション機能の分類と複雑度判定に基づく枠組みを中心に据える一方，Mark II はトランザクション側の計測をより重視する設計思想を持つ．また，FP 法は適用範囲の広さが利点である反面，計測者判断への依存や非機能要件の扱いなどの課題が指摘されてきたため，より明確な定義でデータ移動量を数える COSMIC のように，別の機能概念や計測単位を採用するアプローチも提案・標準化されている．そのため，ISO/IEC では単一方式に統一するのではなく，目的・対象・入力情報の違いに応じて複数方式を国際標準として整備している．ISO/IEC における主なファンクションポイント関連手法規格を表 1 に示す．

FP 法の利点として，要求定義段階の早期から適用できること，ユーザ視点で機能を定量化するため開発者と利用者双方の合意形成に役立つこと，開発言語や実装技術に左右されにくい汎用的な尺度であること，そして標準化された客観指標としてプロジェクト横断的な規模比較や進捗管理，過去実績との工数比較に利用できることが挙げられる．一方で，FP 法には計測者の主観的判断に依存する工程が含まれるため結果にばらつきが生じやすいこと，計測作業に専門知識と多大な工数を要すること，および

性能やセキュリティなど非機能要件の規模への影響を直接には反映しにくいことが課題として指摘されてきた [4, 23].

FP 法にはいくつかの手法的バリエーションが存在するが，最も広く採用されているのは IFPUG 法である [24]．1986 年にアメリカに設立された国際ファンクションポイントユーザグループが定義し，ISO/IEC 20926：2009[15] として国際標準となった.

以下では，代表的手法である IFPUG 法における FP 算出手順について詳述する．なお，本論文では「機能規模としての FP」を Step5 の結果（未調整 FP）として扱う．Step6 および Step7 は，値調整係数（VAF）に基づく調整済 FP を算出するための手順であり，IFPUG 法が ISO 標準となって以降はオプショナルとして扱われる．また，本研究のレビュー対象論文では VAF への言及は確認されず，主として未調整 FP（UFP）を対象としていた.

1. **Step1：算出種類の選択**

   計測目的に応じて算出種類を選択する．具体的には，アプリケーションソフトウェアが提供する機能規模を表す「アプリケーション FP」，アプリケーション FP にデータ移行機能分を加えた「新規開発プロジェクト FP」，および追加・変更・削除・データ移行を含めて算出する「機能改良プロジェクト FP」を区別する.

2. **Step2：計測境界の設定**

   計測対象アプリケーションを境界の内部とし，他のアプリケーションおよびユーザを外部として計測境界を設定する．これにより，どのデータおよび入出力を外部とのやり取りとして扱うかが明確となる.

3. **Step3：データファンクションの計測**

   まず，アプリケーション中に存在し，ユーザが認識できる論理的な意味でのデータのまとまりを同定する．次に，更新対象となる内部論理ファイル（Internal Logical File; ILF）と，参照のみされる外部インタフェースファイル（External Interface File; EIF）に分類する.

   複雑さ評価として，レコード種類数（Record Element Type; RET）およびデータ項目数（Data Element Type; DET）を計数する．RET は異なる意味合いを持つデータのまとまりの個数であり，DET はユーザが認識できる論理的なデータ項目数である．これらの組合せに基づく判定表を用いて，ILF・EIF の複雑さ（低・中・高）を決定する.

4. **Step4：トランザクションファンクションの計測**

   トランザクションファンクションは，アプリケーションに対するデータの入出力を伴う処理（登録・出力・検索等）として同定する．ここでは，(1) 計測境界外と可変なデータの入出力を伴い，(2) ユーザに対する入出力が完結し，データの整合性が保たれる最小単位となるように分割する．トランザクションファンクションの種類は，外部入力，外部出力，外部照会に分類する．外

部入力（External Input; EI）は外部からのデータ入力による ILF の更新である．外部出力（External Output; EO）は何らかの処理ロジックを通した外部へのデータ提供処理である．外部照会（External Inquiry; EQ）は単純なデータ検索による外部へのデータ提供処理である．

複雑さ評価では，関連ファイル数（File Types Referenced; FTR）および DET を計数し，判定表に基づいて複雑さ（低・中・高）を決定する．関連ファイル数は，当該トランザクションファンクションの処理中に更新または参照されるデータファンクションの個数である．なお DET は，入力項目数と出力項目数を合算した上で，共通項目の重複を除外し，さらに実行のきっかけやメッセージ等の扱いについても規定に従って整理する．

5. **Step5：FP（機能規模）の算出（未調整 FP）**

   各ファンクションタイプ（ILF・EIF・EI・EO・EQ）について，複雑さ（低・中・高）に応じた重み付けを適用し，合算して未調整の FP（Unadjusted FP）を得る．

6. **Step6：調整係数（VAF）の算出（オプション）**

   システム全般の特性を評価する 14 の項目ごとに影響度（0〜5）のランクを付与し，各項目の影響度の合計（Total Degree of Influence; TDI）を求めたうえで，値調整係数（Value Adjustment Factor; VAF）を算出する．ここで用いる 14 項目の名称と概要を表 2 に示す．

   以上より，TDI を各項目の影響度の合計値として，調整係数は次式で与えられる．

$$\mathrm{TDI} = \sum_{i=1}^{14} \mathrm{DI}_i, \quad \mathrm{VAF} = 0.65 + 0.01 \times \mathrm{TDI}$$

   ここで，TDI は表 2 に示した 14 項目に対して付与した影響度の合計である．

7. **Step7：調整済 FP の算出（オプション）**

   算出した VAF を未調整 FP に乗じることで調整後ファンクションポイント（Adjusted FP）を得る．

以上が IFPUG 法における FP 計測の基本的な手順である．機能規模としての FP は Step5 の結果であり，Step6 および Step7 は必要に応じて実施するオプショナルな手順である．

FP 法は上記のような標準化された手順が整備された一方で，その実務的有用性を高めるために種々の改良研究も行われてきた [4]．近年では，FP 値による工数見積もりの精度向上や計測作業の省力化を目的として，多様なアプローチが提案されている．例えば，要求記述から深層学習に基づく固有表現抽出（Named Entity Recognition; NER）によって手作業の機能抽出を代替する試みなどがある [25, 26, 27]．このように，FP 法の研究は手法そのものの標準化・普及だけでなく，計測の支援・自動化や精度向上を目指す発展的な提案へと広がっている．

以上の背景から，本研究では FP 法に関する近年の研究知見を体系的に整理し，信頼性の高い形で統合することを試みる．そのためのアプローチとして，次節では本研究で採用する系統的文献レビューの

表 2: システム特性の 14 項目と概要

#	項目名	説明
01	データ通信	アプリケーションやシステムと制御情報やデータを送受信しているか
02	分散データ処理	分散処理しているか
03	パフォーマンス	レスポンスやスループットの性能目標はあるか
04	高負荷構成	システムの利用頻度は高いか
05	トランザクション量	トランザクション量は多いか
06	オンラインデータ入力	オンライン入力する方法は多いか
07	エンドユーザー効率性	アプリケーションはエンドユーザーの効率を考慮して設計されているか
08	オンライン更新	オンラインで論理内部ファイルを更新する機能はあるか
09	複雑な処理	アプリケーションに広範な論理的または数学的処理はあるか
10	再利用可能性	再利用を考慮して開発されているか
11	インストール容易性	インストールが簡単になるように開発されているか
12	操作容易性	起動，バックアップ，およびリカバリ手順等が効率的，また自動化されているか
13	複数サイト	複数のサイトで使用することを考慮して開発されているか
14	変更容易性	システム改修を考慮して開発されているか

概要と手順について述べる．

2.4 系統的レビュー

Kitchenham らが提案したガイドラインに基づき，系統的文献レビュー（Systematic Literature Review; SLR）とは，網羅性と再現性を重視して文献を収集・選別・統合する調査手法である [10]．SLR では事前に明確化した手順に従って文献調査を行うため，手順設計が適切であれば実施者が異なっても同等の結果に到達しやすく，レビュー結果の信頼性を高められる．医療分野ではエビデンス収集の標準的手法として広く用いられており，ソフトウェア工学分野においても，仮説の検証，既存研究の要約，研究空白の特定などを目的に数多く実施されてきた．ソフトウェア工学における SLR の手順は Kitchenham らにより体系化されており，その枠組みに従ったレビュー事例が多数報告されている．本研究で実施する SLR も，基本的には Kitchenham らの提案手順に準拠する．

SLR は単に研究動向を整理するだけでなく，提案手法の評価の厳密性や未解決課題の残存状況を構

造的に把握する上でも有効な手段である [28]．本研究では，SLR を通じて 2019 年以降の FP 法研究を体系的に再整理し，実務への適用可能性と信頼性を検討するための基盤知識を構築する．

2.5 先行研究

先行研究 [9] は 2018 年までの FP 法研究について SLR を実施し，FP 法に関する研究テーマが複数のトピックに分類可能であることを示した．本研究でもこれに倣い，FP 法研究の対象を 6 つのトピック（「FP の利点や欠点の評価」,「計測補助」,「計測結果の活用」,「計測ルールの変更」,「計測が難しいとされている対象への適用」,「その他」）に分類して整理する．

先行研究では，FP 法研究の重心が FP の利点評価や活用方法から，計測補助を志向した研究へと移行している点が指摘されている．一方で，未解決課題として少なくとも次の 3 点が挙げられていた．

1. 適用コンテキストの違い（組織，ドメイン，開発プロセス等）によって FP 計測結果や工数見積もり精度が変動し得るにもかかわらず，その取扱いが十分体系化されていない点．
2. FP 計測の信頼性を担保するための教育手法・訓練プロセスの開発が進んでいない点（なお，調査対象論文には教育を主題として扱った研究は確認されていない）．
3. FP 法で直接計測しにくい要素（例：非機能要求や品質要件）をどのように扱うかが未整理である点．

以上の知見は，FP 法研究のトピック構造と動向を示すと同時に，FP 法の実務適用上の課題や研究の信頼性に関わる問題が依然残存していることを示唆する．したがって，2019 年以降に発表された研究成果を対象に，これらの課題がどの程度解消されているかを改めて検証する必要がある．また，本研究では，先行研究の整理枠組みを踏まえつつ，研究の成熟度（進展段階）と有効性評価の厳密性を新たな観点として加え，研究成果の実務適用可能性とエビデンスの信頼性を併せて評価する．

3　実施した系統的レビュー

本節では，本研究で実施した系統的文献レビュー（Systematic Literature Review: SLR）の手順を詳細に説明する．SLR の実施手順は，Kitchenham らが提案したガイドラインに基づき，以下の 5 手順で構成する．

1. Research Question（RQ）の設定：レビューで明らかにしたい問いを設定する．
2. 研究論文の収集：検索対象データベース，キーワード，期間を事前に定めて論文候補を収集する．
3. 研究論文の選別：包含基準・除外基準および品質基準に基づき，調査対象論文を決定する．
4. 情報の抽出：RQ への回答に必要な情報を対象論文から抽出し，統一形式で整理する．
5. 結論の導出：抽出情報を集計・分析し，RQ への回答を導出する．

以降では，まず本研究の研究目的を述べ，その上で各手順について順に説明する．

3.1　研究目的

ファンクションポイント法（FPA）に関する研究は，計測手順の改善，自動化支援の導入，計測結果の活用方法の拡張など，多様な方向へ展開している．一方で，FPA は実務で広く用いられているにもかかわらず，計測に要する作業負荷や判断のばらつき，適用コンテキストの差異による結果の変動などが継続的に課題として指摘されてきた．先行研究 [9] では，これらの課題認識に加え，研究動向として重心が FP の利点評価や活用方法から，計測補助を志向した研究へと移行している点，ならびに未解決課題（コンテキスト，教育，FPA で計測しにくい要素への対応等）が整理されている．しかし，近年の技術・実務環境の変化により，FPA は手作業の計測対象にとどまらず，AI・機械学習を用いた見積りの入力として扱われる場面も増えており，2019 年以降の研究成果を改めて体系的に捉え直す必要がある．

以上を踏まえ，本研究の目的は，2019 - 2025 年に公表された FPA 関連研究を体系的に整理し，(1) トピックの変化と研究の成熟度（進展段階），(2) 提案手法の有効性評価の厳密性，(3) 先行研究で示された未解決課題の解消状況を明らかにすることである．具体的には，2018 年までの先行研究で整理されたトピック構造と指摘事項を出発点として，2019 年以降の研究を同一の観点で分類し，どのトピックが中心となっているか，また各トピックが手法提案・ツール開発・実証適用といった進展段階のどこに位置づくかを整理する．さらに，有効性評価については，評価データの多様性・規模，比較対象の妥当性，統計的検証の有無，再現に必要な情報の公開状況といった観点から，提案の根拠がどの程度厳密に示されているかを整理する．加えて，コンテキスト記述の不足，教育・判断支援の体系化，FPA で直接計測しにくい要素（例：非機能的側面）への対応といった未解決課題が，2019 年以降の研究でどの程度扱われ，どの課題が依然として残存しているかを明らかにする．

3.2 Research Question の設定

本手順では，本研究が SLR により解明したい問いを Research Question（RQ）として設定する．先行研究では，FP 法研究の重心が FP の利点評価や活用方法から，計測補助を志向した研究へと移行している点が指摘されている．一方で，改良手法がどの程度厳密に評価され，どの課題が解消されつつあるかは，2019 年以降の成果に対して体系的に再確認する必要がある．そこで本研究では，研究動向の把握（RQ1），研究の進展段階の整理（RQ2），有効性評価の厳密性の把握（RQ3），および未解決課題への応答状況の確認（RQ4）という 4 つの観点から，次の RQ を設定した．なお，RQ1 で得たトピック分類を以降の分析の基盤とし，RQ2〜RQ4 ではトピック別・進展段階別の比較を行う．

- **RQ1**：2019 年以降の FP 研究はどの研究トピックに分類され，各トピックの研究上の位置づけはどのように変化したか．
- **RQ2**：各トピックに含まれる研究は，成熟度（進展段階：手法無し・手法提案・ツール開発・実証適用）の観点からどのように構成されているか．
- **RQ3**：提案手法の有効性評価はどの程度厳密に行われているか．
- **RQ4**：2018 年時点で指摘された未解決課題は，2019 年以降の研究において，どの程度解消されているか．

以降，3.2.1 節で RQ1，3.2.2 節で RQ2，3.2.3 節で RQ3，3.2.4 節で RQ4 の意図と分析観点を述べる．

3.2.1 RQ1

2019 年以降の FPA 改良研究は多様化しており，個別研究を断片的に読むだけでは全体像を把握しにくい．そこで RQ1 では，対象論文をトピックに体系分類し，2019 年以降の研究動向を俯瞰できる形で整理することを目的とする．

トピック分類の設計　本研究では，対象論文を 6 トピック（評価，計測補助，活用，ルール変更，適用，その他）に分類する．各トピックは「研究の主目的（何を明らかにする・実現するか）」と「主要な貢献（何を提案・実装・検証するか）」に基づいて定義し，分類基準の解釈が揺れないように整理した上で適用する．例えば，既存手法間の比較や精度・有用性の検証を主目的とする研究は「評価」，計測作業の簡略化・自動化・支援を主目的とする研究は「計測補助」，FP を入力として別目的（見積り・テスト・計画・分析等）に利用する研究は「活用」とする．計測ルールや定義の変更・拡張を提案する研究は「ルール変更」，特定ドメイン・特定形態への適用・適用手順の整理を主目的とする研究は「適用」とし，いずれにも当てはまらないものを「その他」とする．

分類手順　分類は，(i) タイトル・要旨・結論に加え，(ii) 研究課題，(iii) 提案内容，(iv) 評価対象（対象物・データ・事例）を確認したうえで一次分類を付与する．複数トピックに跨る場合は，主要内容（最も大きい貢献・結論が置かれている部分）に基づき一次分類を行い，補助的な要素は備考として記録する．これにより，1 本の論文が複数トピックに属することによる二重計上を避けつつ，研究内容の多面性も失わない形で整理する．

集計と分析　分類結果に基づき，2019 年以降のトピック別件数と構成比を算出し，トピック間の相対的な存在感（主要トピック・周辺トピック）を把握する．また，先行研究（1979 - 2018）で報告されたトピック分布と比較し，構成比の増減や順位の変化を通じて「研究上の位置づけ」の変化を議論する．ここで「位置づけ」は，単なる件数だけでなく，構成比の変化と，後続の RQ2〜RQ3 で明らかにする成熟度・エビデンスの厚みとの関係も含めて解釈する．

3.2.2　RQ2

2019 年以降の FPA 関連研究は，計測補助（自動化）や活用を中心に多様化しているが，研究の進み具合（どこまで実装・適用されているか）は一様ではない．例えば，概念や手順の提案に留まる研究，ツール・システムとして実装まで行う研究，実プロジェクトや実データでの実証適用まで行う研究では，実務への距離や得られる根拠の強さが異なる．

このような「提案から実装，実環境での検証へ」という段階的な進展は，設計科学研究における成果物の構築と評価の考え方 [29] や，方法論としての設計・開発，デモンストレーション，評価という流れ [30] と整合する．また，ソフトウェア工学における研究成果の実務への移転に関しても，解決案の提示から段階的検証を経て実運用に至るモデルが提示されている [31]．さらに，要求工学分野では，解決策の設計・提案とその検証，実装・適用後の評価を区別して整理する枠組みが議論されている [32]．加えて，研究成果の妥当性は分析や例示，実運用での経験など多様な検証形態によって支えられることが指摘されている [33]．

進展段階（成熟度）の定義　そこで RQ2 では，対象論文を各トピックに分類した上で，研究の成熟度（進展段階）の観点から，研究がどの段階に位置づくかを整理する．本研究では進展段階を（i）手法無し，（ii）手法提案，（iii）ツール開発，（iv）実証適用の 4 区分として定義する．ここで（i）手法無しは，新規の手順・モデル・ツール等を提案せず，整理・議論・観察・比較枠組みの提示に留まる研究を指す．（ii）手法提案は，新規の手順・モデル・ルール・アプローチ等を明示的に提示するが，実装としてのツール提供を必須としない研究を指す．（iii）ツール開発は，提案内容がツール・システム・プロトタイプとして実装され，手順の再実行が可能な形で示されている研究を指す．（iv）実証適用は，提案手

法・ツールが組織の業務プロセスに組み込まれ，日常的・継続的に利用されていることが本文中で明示される研究を指す．

判定手順　進展段階の判定は，(a) 新規提案の有無，(b) 実装物の有無，(c) 実データ・実プロジェクトでの適用の有無，の 3 点を中心に行う．例えば，実装があるが業務プロセスへの統合が明示されない場合は（iii）に留め，日常的・継続的利用が明示される場合は（iv）とする．また，提案内容が示されていても実装・適用の記述が無い場合は（ii）とする．このように判定規則を明文化することで，分類の一貫性を確保する．

集計と分析　全対象論文について進展段階を付与し，トピック別に（i）〜（iv）の分布を集計する．これにより，どのトピックでツール開発や実証適用が進んでいるか，あるいは提案段階に留まる研究が多いかを明らかにする．さらに，RQ1 のトピック分布と合わせて解釈することで，「件数として増えている領域」と「成熟している領域」が一致するか否かも検討する．加えて，進展段階 2〜4（手法提案・ツール開発・実証適用）に該当する研究は，後続の RQ3 における有効性評価の厳密性分析（提案の根拠の強さ評価）の調査対象の候補として位置づける．

3.2.3　RQ3

FPA は実務での利用を前提とする尺度であるため，研究成果が現場で参照可能な根拠として提示されているか，すなわち提案手法の有効性評価がどの程度厳密に行われているかが重要となる．しかし，有効性評価の記述粒度は論文により異なり，評価が特定データ・特定条件に依存していないか，比較対象の設定が公正か，統計的検証が行われているか，再現に必要な情報が開示されているかを横断的に把握することは容易ではない．

特に比較対象の妥当性については，提案手法が有利になるように「実験しなくとも提案手法に有利な比較対象」を選ぶ，すなわち恣意的に弱いベースラインや不十分な既存法を比較対象として採用することで，改善効果が過大に見積もられ得る点に注意が必要である．この種の不公正な比較や手続きの不一致は，比較研究の結論を不安定・不正確にし得ることが指摘されている [34, 35]．そのため，本研究ではベースライン比較の重要性や，多手法比較における統計的検証の推奨，および成果物公開による再現性確保と整合する形で，有効性評価の厳密性を定義する [36, 37, 38]．

調査対象　RQ3 の調査対象は，RQ2（成熟度調査）において進展段階 2〜4（手法提案・ツール開発・実証適用）に該当し，かつ提案（手法・アプローチ・ツール等）を含む研究（30 本）とする．ただし，提案を含むにもかかわらず有効性評価を実施していない研究が存在するため，まず評価の実施有無を判定し，評価段階（評価無し〜評価上）を付与する．

評価段階（評価無し〜評価上）の定義　本研究では，対象論文について以下の手順で評価段階を決定する．まず，提案に対する有効性評価（実験，ケーススタディ，比較評価等）が実施されていない場合は，評価無しとする（評価観点の判定は行わない）．一方，評価が実施されている場合は，後述の 4 観点をそれぞれ Y・N（Yes・No）で判定し，満足（Y）数により次の段階を付与する：

- **評価下**: 4 観点のうち 満足数が 1 個以下（0 または 1）
- **評価中**: 4 観点のうち 満足数が 2〜3 個
- **評価上**: 4 観点のうち 4 個すべて満足

この定義により，「評価を実施したかどうか」と「評価の厳密さ（満足数）」を分離し，提案効果の根拠の強さを段階的に整理する．

4 観点（Y・N）と判定基準　評価を実施している論文について，次の 4 観点を Y・N で記録する．判定は，本文・付録・補足資料に記載された情報のうち，第三者が追跡可能な具体的記述に基づく．

- **Q1（データセットの多様性・規模）**：評価データについて，(i) データの出所（企業データ・公開データ等）または収集方法が明記され，(ii) 規模数（プロジェクト数，要求文数，事例数など）が明記され，(iii) 複数プロジェクト（または複数データセット）を含む，のすべてを満たす場合を Y とする．いずれかが欠ける場合（例：出所のみで規模不明，単一事例のみ等）は N とする．
- **Q2（比較対象の妥当性：公正比較か）**：提案手法が，少なくとも 1 つの比較対象（既存法または明示的ベースライン）と同一データ・同一評価指標・同一評価手順で比較されており，かつ比較対象の選定理由が説明されている場合を Y とする．ここで「妥当な比較対象」とは，実務・研究で広く用いられる代表的手法，または先行研究により標準的ベースラインとして位置づくものを含み，「提案手法に有利な比較対象」を恣意的に選ぶ（弱すぎるベースラインのみ，主要ベースラインの欠落，条件を揃えない比較など）ことを避けた比較を指す．比較が無い，条件不一致，選定理由の欠如，または上記のような恣意性が疑われる場合は N とする．
- **Q3（統計的検証の有無）**：手法間の差について，(i) 有意性検定（例：Wilcoxon 検定等）を実施している，または (ii) 信頼区間や効果量を提示している，のいずれかを満たす場合を Y とする．評価指標（平均誤差等）のみで，不確実性・差の根拠が示されない場合は N とする．
- **Q4（再現性：データ・手順の公開）**：第三者が追試可能となるように，(i) データ，(ii) コード（実装・スクリプト），(iii) 手順（前処理，設定，パラメータ等の再実行に必要な詳細）のうち，少なくとも 1 つが公開され，入手可能である場合を Y とする [38]．公開が一切無い場合や，入手不能（URL 不通・非公開等）の場合は N とする．

集計と分析　各論文について，評価実施の有無（評価無し・それ以外）と，Q1 - Q4 の満足数（0 - 4）を記録し，評価段階（評価無し・評価下・評価中・評価上）の件数を算出する．さらに，トピック別および進展段階別に評価段階の分布を比較し，エビデンスの不足・偏りを可視化するマップ（Evidence Gap Map; EGM）として可視化する．これにより，どのトピック・どの進展段階の研究でエビデンスが厚い・薄いかを把握し，実務適用可能性を判断するための根拠の整備状況を明らかにする．

3.2.4　RQ4

先行研究では，少なくとも次の未解決課題が指摘されている．すなわち，(1) 適用コンテキスト（組織，ドメイン，開発プロセス等）の整理不足，(2) 教育への対応（教育を扱う研究の不足），(3) FP で計測しにくい要素への対応である．そこで RQ4 では，2019 年以降の研究がこれらの課題にどの程度応答しているかを，トピック分類および個別研究の主張に基づき整理する．

調査観点（未解決課題の定義）　RQ4 では，先行研究指摘された 3 点を「未解決課題」として固定し，2019 年以降の研究における扱いを確認する．(1) コンテキストは，研究の対象（ドメイン，規模，言語等）が明示され，結果の適用範囲や前提条件を理解できる形で整理されているかに着目する．(2) 教育への対応は，FPA の学習・普及・訓練・指導（教材，教育手順，支援環境等）を研究対象として扱い，方法の提案または効果の検証が行われているかに着目する．(3) FP で計測しにくい要素への対応は，特に非機能要件や品質特性など，FP の枠組みで捉えにくい要素をどのように補うか（補助尺度の導入，拡張ルール，別手法との併用等）に着目する．

判定と整理の方法　各論文について，上記 (1) 〜 (3) に関する記述を確認し，(i) 課題を明示しているか，(ii) 何らかの対応（手法・枠組み・運用方法）を提案しているか，(iii) 対応の効果や妥当性に関する根拠（事例，評価，議論）が示されているかを抽出する．これにより，「言及のみ」で終わる研究と，「具体的な対応」を示す研究とを区別し，未解決課題に対する応答の深さを整理する．

集計と分析　整理結果は，RQ1 のトピック分類および RQ2 の進展段階と対応づけて集計する．例えば，計測補助・活用が増加している一方でコンテキスト報告が依然として不足しているか，教育研究がどのトピック・どの進展段階に現れるか，非機能要件等への対応が特定領域に偏っているかを確認する．

3.3　研究論文の収集

本手順では，調査対象となる研究論文の候補集合を収集する．収集に先立ち，検索期間，検索対象データベース，および検索キーワードを設定する．

3.3.1　検索する期間

本研究は，先行研究が対象とした期間（〜2018 年）の後続研究を整理することを目的とするため，検索期間を 2019 年〜2025 年とした．

3.3.2　検索対象データベース

検索対象データベースは，ソフトウェア工学分野の主要文献を広く包含し，論文を体系的に収集できる点を重視して選定した．本研究では，以下のデータベースを検索対象とした．

- ACM Digital Library
- Google Scholar
- ScienceDirect
- Scopus
- IEEE Xplore

3.3.3　検索に用いるキーワード

検索キーワードは，FPA および関連する機能規模計測文脈を取りこぼさないことを意図して設定した．具体的には，以下を基本キーワードとし，DB ごとの検索仕様に合わせてクエリを構成した．

- "function point"，"function points"，"function-point"，"function-points"

IEEE Xplore 等の主要 DB では，タイトル，アブストラクト，キーワードのいずれかに上記語が含まれる論文を対象とした．一方で Google Scholar は，抄録・本文まで含めると無関係論文が大量に混入するため，タイトルに上記語が含まれる論文を対象とした．なお，検索語を「function point」系列に限定しているため，当該語を含まない関連研究は取りこぼし得る．

3.4　研究論文の選別

図 1 に対象論文の収集および選別のフローを示す．

本手順では，収集した論文候補から調査対象論文を選別する．選別は，重複除去，適合性スクリーニング，品質基準チェックの順で行った．手順を以下に示す．

1. 各 DB から得た候補を統合し，タイトル・著者・出版年等に基づき重複論文を除外する．
2. タイトル，アブストラクト，キーワードを確認し，ソフトウェア工学分野に属さない論文，および FPA と無関係な論文を除外する．

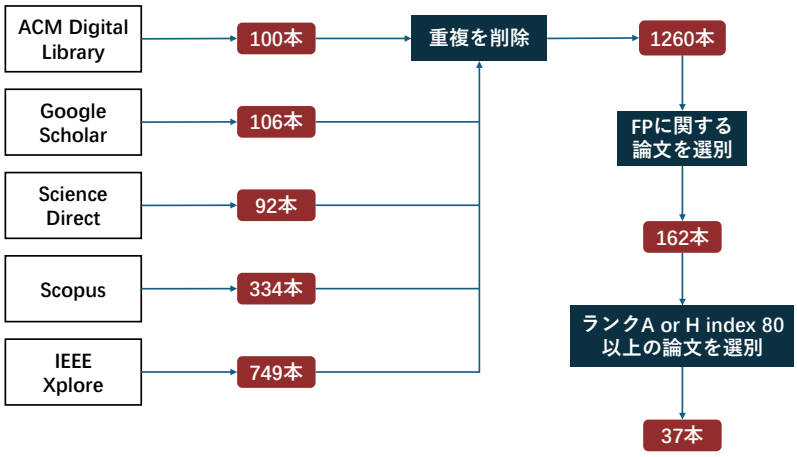


図 1: 論文収集と選別の結果

3. 各論文の出典を確認し，CORE Rankings Portal においてランクが A 以上である論文誌または国際会議の論文集および h-index が 80 以上である論文誌または国際会議の論文集に採録されている論文以外を除外する．
4. 残った論文を調査対象論文として確定する．

本研究では，先行研究に倣い，品質基準を出典（ジャーナル・国際会議）に基づいて運用した．

具体的には，会議論文については，査読・採択の厳格さを出典レベルで担保するため，CORE Rankings Portal を用いた．CORE Rankings Portal は主要会議を対象に，学術委員会が提出情報と客観的データに基づいて評価・更新するランキングであり，A は「当該分野で高く尊敬される優れた会議」，A* は「旗艦会議」と位置づけられている．したがって，会議論文の出典を A 以上に限定することで，採択競争性と査読品質が一定以上と見なされる会議に絞り込み，質が不明確な会議録の混入リスクを低減できる．

一方で，CORE Rankings Portal は会議を主対象とするため，ジャーナル論文の出典評価には，被引用に基づく h-index を用いた．h-index は「少なくとも h 本の論文がそれぞれ h 回以上引用されている」状態を表し，研究成果の蓄積量と学術的影響の双方を反映する指標である [39]．本研究では，ジャーナルの出典として一定の影響力が継続して確認される論文を優先する目的で，h-index 80 以上を閾値として設定した．以上の方針により，会議論文は CORE Rankings Portal（A 以上），ジャーナル論文は h-index（80 以上）というように，出典種別に応じて適切な基準を適用し，調査対象の質を担保した．

以上の選別を通じて，最終的に 37 本を調査対象論文として選定した．

3.5 情報の抽出

本手順では，対象論文の本文を確認し，RQ への回答に必要な情報を抽出し，統一形式で整理する．抽出項目は，(i) 論文の基本情報と，(ii) RQ 別に必要となる分析情報から構成する．

3.5.1 論文の基本情報

- タイトル
- 著者名
- 出典名
- 出版年

出版年は RQ1 の年代別傾向分析や RQ2〜RQ4 の時系列比較に用いる．

3.5.2 RQ 別の抽出情報

- 研究背景（RQ2，RQ4 の原因分析・文脈整理に使用）
- 研究目的（RQ1 のトピック分類，RQ2 の成熟度判定に使用）
- 提案内容・手法（RQ1 の分類根拠，RQ2 の成熟度判定に使用）
- 有効性評価の有無と方法（RQ3 の評価に使用）
- 結果（RQ4 の未解決課題の解消程度に使用）
- 結論（RQ4 の未解決課題との対応付けに使用）

3.6 結論の導出

本手順では，抽出情報を集計・分析し，RQ への回答を導出する．RQ1 では，トピック分類に基づき，2019 年以降の研究の中心領域と周辺領域を明確化する．RQ2 では，各トピックに含まれる研究を進展段階（手法無し・手法提案・ツール開発・実証適用）で整理し，研究の成熟度と実務志向の現れ方を分析する．RQ3 では，有効性評価の厳密性を 4 観点で整理し，EGM により「エビデンスが厚い領域」と「不足する領域」を可視化する．RQ4 では，先行研究で指摘された未解決課題が 2019 年以降にどの程度扱われたかを整理し，残存ギャップを特定する．

各 RQ の集計結果と回答は第 4 節で示し，それらを踏まえた考察は第 5 節で議論する．

4　レビュー結果

4.1　調査対象論文の外観

論文の収集と選別を行った結果 37 本の論文が調査対象として選択された．各調査対象論文のタイトル，著者名，出典名，出版年を「付録 A 調査対象論文」に示す．以降の小節で各 RQ への回答を行う．

4.2　RQ1

4.2.1　トピック分類

近年のファンクションポイントに関する研究（2019 – 2025 年，対象論文 37 本）は，その内容から以下の 6 つのトピックに分類できる（表 3）．

表 3: 各トピック及び含まれる論文と本数

トピック	含まれる論文	本数
FP の利点や欠点の評価	[40, 41, 42, 43, 44]	5
計測補助	[12, 13, 25, 26, 27, 45, 46, 47, 48, 49, 50, 51]	12
計測結果の活用	[52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62]	11
計測ルールの変更	[63]	1
計測が難しいとされている対象への適用	[64, 65, 66, 67, 68]	5
その他	[69, 70, 71]	3

以下では各トピックの定義を示す．

1. **FP の利点や欠点の評価**

   FP の計測精度や信頼性を評価する研究．研究成果の評価ではなく，既に存在している FP 法についての評価が行われている研究がこのトピックに含まれる．2019 年以降の論文では 5 本（13.5%）が該当する．

2. **計測補助**

   FP 計測の手間を削減するための自動計測ツールなどの研究．計測自体は可能であるシステムや開発手法に対して補助を行う研究がこのトピックに含まれる．そもそも計測が難しいとされてい

る開発手法などからの計測は後述の「適用」トピックに含まれる．2019 年以降の論文では 12 本（32.4%）を占め，6 トピック中最大の割合となっている．

3. **計測結果の活用**

   主に計測された FP を用いて工数や生産性を導出するための研究，FP の結果を見積もりなどに活用するための研究がこのトピックに含まれる．2019 年以降の論文では 11 本（29.7%）が該当する．

4. **計測ルールの変更**

   FP の計測ルールの校正もしくは新たな計測ルールの提案についての研究．計測要素や重みづけの再定義など，既存の計測ルールを改善するために行われている研究がこのトピックに含まれる．2019 年以降の論文では該当論文 1 本（2.7%）と少数である．

5. **計測が難しいとされている対象への適用**

   FP の適用が難しいとされる開発手法やアプリケーションへの適用についての研究．従来 FP の適用が難しいとされている非ウォーターフォール型開発に対する計測方法の提案などがこのトピックに含まれる．既に計測方法が確立されているウォーターフォール型開発への適用などは上述の「評価」トピックに含まれ，計測の手間の削減などの計測補助が主目的である研究は「計測補助」トピックに含まれる．2019 年以降の論文では 5 本（13.5%）を占める．

6. **その他**

   上記いずれにも分類しにくい FP 研究が該当する．異なる FP 計測法間での規模値変換や FP 研究に関するシステマティックレビュー（SLR）などが含まれる．2019 年以降の論文では 3 本（8.1%）を占める．

要するに，「他の開発規模尺度と比べてどういう点で優れているのか」について研究している論文は「評価」，「どうすれば計測の手間が減らせるか」について研究している論文は「計測補助」，「計測によって得られた FP をどのように工数に変換するか」を研究している論文は「活用」トピックにそれぞれ分類される．また，「どのように重みづけを較正すればより正確な FP が得られるか」を研究している論文は「ルール変更」，「どうすれば FP を得られるようになるか」を研究している論文は「適用」としてそれぞれ分類される．

各トピックにおいて引用数が多い上位 2 本の論文の概要を示す．ただし，「計測ルールの変更」トピックに分類された論文は 1 本のみであるため，当該トピックについてはその 1 本のみを示す．

FP の利点や欠点の評価

- **Suresh Kumar らの研究 [41]：ソフトウェア工数見積もりにおける従来手法（FP 等）の限界と学習モデルの有効性を評価**

ソフトウェア工数見積もりに関する既存研究を整理し，ニューラルネットワークから深層学習までの発展と適用状況を調査した．その上で，FP 法などの従来の見積もり枠組みは，プロジェクト初期の不確実性や，要因間の非線形関係により精度面で限界を持ち得る点を評価した．さらに，学習モデルを用いた推定は，従来手法や統計的手法に比べて高い推定性能を示し得ることを整理した．以上より，FP は技術非依存な尺度として有用である一方，工数見積もりの精度確保という観点では限界があり，学習モデルの導入がその欠点を補完し得ることが示された．

- **Hoc らの研究 [43]：FP 情報を用いた工数見積もりにおける推定モデルの比較を通じて，FP の有用性と限界を評価**

  FP に基づく工数見積もりを対象に，重回帰分析，多層パーセプトロン，深層学習を比較し，推定性能を評価した．調整後 FP，FP カテゴリ，産業分野，相対規模を説明変数として用い，公開データセット（ISBSG）で検証した．その結果，深層学習が最も高い推定性能を示し（標準化精度 72%，予測水準 0.30 で 72%），多層パーセプトロンも重回帰分析や基準手法より良好であることが示された．さらに，調整後 FP は，FP カテゴリに比べて常に精度向上に寄与するとは限らないことが示された．

計測補助

- **Zhang らの研究 [25]：深層学習により要求文書から FP 要素の自動計測**

  産業ドメインの要求文書にアノテーションを付与したデータを用い，系列ラベリングとして FP 要素（種別）の認識を学習させる手法を実装した．具体的には，双方向 LSTM と条件付き確率場を組み合わせたモデルにより，新規要求文書中の FP 種別を自動分類し，計測者は結果の確認に注力する流れを構成した．29 件の実プロジェクトで検証した結果，認識精度は適合率 94.5%，F1 値 80.3% を達成し，さらに計測プロセス全体の効率は平均 38.6% 改善することが示された．
- **Lavazza らの研究 [12]：機械学習により機能規模の推定・FP 計測の補助**

  詳細要求や熟練計測者が十分に揃わない段階でも機能規模（非調整 FP）を見積もるため，高レベル FP（固定重み）や簡易 FP と同等の入力情報から，機械学習による推定モデルを構築・比較した．479 件のプロジェクトデータで，サポートベクタ回帰，ニューラルネットワーク，ランダムフォレスト等を用いた推定を行い，高レベル FP や最小二乗回帰と誤差指標で比較した．その結果，サポートベクタ回帰（5 変数）が最も小さい平均絶対残差（MAR=278）と高い決定係数（$R^2 = 0.985$）を示した一方で，高レベル FP（MAR=303，$R^2 = 0.984$）との差は大きくないことが整理された．別データセット（ISBSG）でも同様の比較を行い，サポートベクタ回帰が最良の推定値を与えるものの，他手法との差の効果量は小さい範囲に留まることが示され，入力負担を増やさずに推定精度を維持・改善するという観点で，機械学習の実務的適用可能性が示された．

計測結果の活用

- **Silhavy らの研究 [54]：FP 計測結果をカテゴリ別に分割して回帰モデルへ投入し，工数見積もりに活用**

  ファンクションポイント分析で得られる計測結果（機能種別の計測値など）を説明変数として用い，カテゴリ変数（例：相対規模，産業分野，業務領域）でデータを分割した上で，各区分ごとに段階的回帰により工数見積もりモデルを作成した．また，従来の IFPUG に基づく一括適用や，クラスタリングに基づくモデル化と比較し，FP 計測結果の「区分別活用」が見積精度に与える影響を評価した．その結果，IFPUG では MAPE が約 45%，PRED（25）が 0.49 であるのに対し，提案手法は PRED を約 4% 向上させ，MAPE を約 11% 低減することが示された．さらに，クラスタリング手法と比べても MAPE が低く（例：提案手法 34%，スペクトルクラスタリング 41%），FP 計測結果を区分別モデルに接続して活用することが有効であると結論を得た．
- **Van Hai らの研究 [61]：FP 計測結果をクラスタ別に適用して見積精度を高める活用法を比較・検証**

  ISBSG データを対象に，カテゴリ変数（開発プラットフォーム，産業分野，言語種別，組織種別，相対規模）および k-means によりデータをクラスタ化し，各クラスタに対して FP 分析に基づく工数見積もりを適用して精度を比較した．併せて，クラスタごとの FP 計測結果から重みづけ（複雑度重み）を学習して推定に用いる手法も適用し，FP 計測結果を「クラスタ別に使い分ける」ことの効果を評価した．その結果，クラスタリングありはクラスタリングなしより常に誤差が小さく，FP 分析では二乗平均平方根誤差（RMSE）で最大 58.06% の平均改善が得られることが示された．また，最良のクラスタリング基準は産業分野であり，FP 分析で 63.68%，学習モデル併用で 72.02% の改善が得られることが示され，FP 計測結果をクラスタ単位で活用することが見積精度向上に有効であるという結論を得た．

計測ルールの変更

- **Effendi らの研究 [63]：標準 IFPUG の代替として「取引機能数」を用いる簡略化ルールを提案**

  標準 IFPUG による計測は要求の詳細分析を要し，早期見積もりでは時間・コスト面で適用が難しい状況がある点を整理した．そこで，FP 計測手順の一部（取引機能の同定）だけを実施し，「取引機能数（#TF）」を規模指標として直接用いることで，標準 FP 計測を置き換える計測ルールを提案した．ISBSG データを用いて，標準 IFPUG FP（Size）と取引機能数（#TF）をそれぞれ説明変数とする工数推定モデルを構築し，推定誤差を比較した．その結果，新規開発では

MAR(Size)=3,244 に対して MAR(#TF)=3,305，拡張（追加中心）では MAR(Size)=1,881 に対して MAR(#TF)=1,877 となり，精度差は僅少であることが示された．また，FP と#TF の相関は Spearman の $\rho = 0.91$ と高く，簡略化ルールでも標準 FP に近い情報を保持できることが示された．さらに，標準 FP 計測を行わず取引機能数を用いる運用では，計測コスト全体で 30% 以上，主要計測作業で少なくとも 50% の削減が可能である点を整理し，標準 FP 計測を簡略化ルールへ置換する実務的妥当性が示された．

計測が難しいとされている対象への適用

- **Rosa らの研究 [64]：政府機関のアジャイル開発に対してファンクションポイントを適用し，早期見積もりでの計測困難性に対応**

  米国政府調達案件におけるアジャイル開発では，契約初期にストーリーポイント等が利用しにくい状況を踏まえ，UFP（未調整ファンクションポイント）および SiFP（簡易ファンクションポイント）を含む複数の規模尺度を，同一条件で工数見積もりへ適用して比較した．プロダクトバックログや要求関連文書を用いて，機能ストーリーを数え上げる手順を整備し，併せて UFP・SiFP を算出してデータ正規化を行った．17 件のアジャイル案件を対象に，規模尺度ごとの単回帰モデルを作成し，適合度（調整済み決定係数）と誤差（MMRE）で比較した．その結果，機能ストーリーは MMRE 29% で最良となり，SiFP（30%）や UFP（32%）も高い予測性能を示し，ファンクションポイント（UFP・SiFP）はストーリーポイント等より有効な説明変数となり得ることが示された．
- **Mushtaq らの研究 [65]：モバイルアプリへの COSMIC Function Point 拡張により，非機能要因を含む対象へ適用**

  モバイルアプリ開発では，端末制約や利用状況に起因する非機能要因が工数へ影響しやすく，従来の機能規模尺度のみでは扱いにくい点に着目した．そこで，COSMIC に基づくモバイル向け規模尺度として Mobile COSMIC Function Points（MCFP）を定義し，技術的複雑度要因（MTCF）と環境的複雑度要因（MECF）を組み合わせて MCFP を算出する手順を提案した．さらに，MCFP を入力として工数を推定する枠組み（CPEEM）を構成し，36 件の実在モバイルアプリを対象に，FPA および COSMIC による推定と比較した．その結果，MMRE は FPA 0.115454，COSMIC 0.064746，CPEEM 0.051385 となり，PRED(25) も CPEEM が 97.22222 と最良であり，非機能要因を含むモバイル開発に対して，ファンクションポイント系尺度の拡張適用が有効であることが示された．

その他

- **Tanabata らの研究 [69]：実装フェーズにおける個人貢献をファンクションポイントで定量化して活用**

  プロジェクト型学習のソフトウェア開発演習を対象に，実装フェーズで各学生が実装した機能を FP で計測し，個人の貢献度を定量化する方法を提案した．GitHub のマイルストーン機能を用いて担当機能と完了状況を追跡し，機能ごとの FP を合算して個人別の貢献量を算出した．その結果，実装機能数が近い場合でも，ILF 等を含む複雑な機能を担当した学生の貢献を FP で区別できることが示された．さらに，相互支援や未完了機能が評価から漏れる課題に対し，支援者に当該機能 FP の 30% を加算する規則，および未完了機能をサブタスク分割して進捗に応じて FP を按分する規則を導入した．以上より，FP を「規模見積もり」以外の目的（個人貢献の定量評価）に転用し，実装成果の活用先を拡張できることが示された．
- **Bluemke らの研究 [70]：テスト工数見積もり研究を広範に整理し，ファンクションポイントの利用形態を体系化**

  ソフトウェアテスト工数見積もりと関連領域を対象に系統的文献レビューを行い，既存手法・ツール・入力情報・制約を広い範囲で分類・整理した．その中で，テスト時間の算出では，システム規模をファンクションポイントで表し，テスト戦略や生産性係数と組み合わせてテスト時間へ変換する考え方が用いられていることを整理した．また，テスト工数見積もりの実用的ツールを見出すことが難しい点を指摘しつつ，Test Point Analysis と自動 Function Point Analysis を組み合わせ，UML 図からテスト時間（工数）を算出するツール実装例を位置づけた．以上より，テスト工数見積もりでは FP が「規模の入力」として活用され得る一方，信頼できる手法・ツールは限定的であることが示された．

4.2.2 各トピックの位置付け変化

トピック分類ごとの件数内訳を，1979 – 2018 年の先行研究 95 本と比較した結果を表 4 に示す．2019 年以降（2019 - 2025 年）と 2018 年以前（1979 - 2018 年）における各トピックの論文数および構成比を整理したものであり，期間で研究関心の重心がどのように変化したかを俯瞰できる．ここでは，単純な本数の増減だけでなく，構成比の変化に着目することで，FP 研究が相対的にどの領域へ移行しているかを明確化する．

この 7 年間の FP 研究動向として以下が読み取れる．計測補助と活用のカテゴリ比率が増加し，主要トピックとしての存在感を増している一方，評価および計測ルール変更の比率は大きく減少している．具体的には，計測補助は全体の約 32% を占める主要トピックとなり，活用も約 30% まで増加した．反面，評価研究は約 16% に低下し，FP 手法自体の評価・検証を主題とする研究が割合として減っていることが分かる．また，計測ルール変更も 1979 – 2018 年の 15% から約 3% へと減少し，新たな FP 手

法提案の頻度が低下したことが読み取れる．適用カテゴリは大きな変化はなく，中程度（約 10% 台）の水準を維持しており，その他も依然としてごく少数に留まる．

表 4: トピック別の論文数（1979 - 2018 と 2019 - 2025 の比較）

トピック	**本数 (1979 - 2018) total=95**	**本数 (2019 - 2025) total=37**	**トレンド**
FP の利点や欠点の評価	28(29.5%)	5(13.5%)	- 16.0%
計測補助	25(26.3%)	12(32.4%)	+ 6.1%
計測結果の活用	14(14.7%)	11(29.7%)	+ 15.0%
計測ルールの変更	14(14.7%)	1(2.7%)	- 12.0%
計測が難しいとされている対象への適用	9(9.5%)	5(13.5%)	+ 4.0%
その他	5(5.3%)	3(8.1%)	+ 2.8%

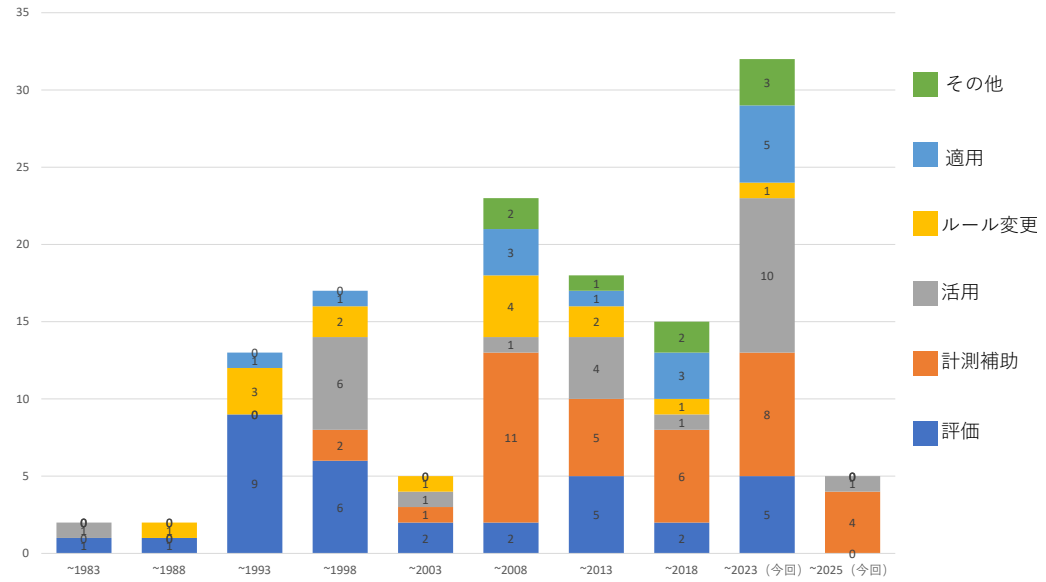
調査対象論文を出版年に基づき概ね 5 年単位で集計し，年代ごとの論文数とトピック構成比の推移を分析する．年代別のトピック構成を図 2a に，割合を図 2b に示す．

なお，直近期間は収集範囲の都合により 2019 - 2023 年と 2024 - 2025 年に分割している．

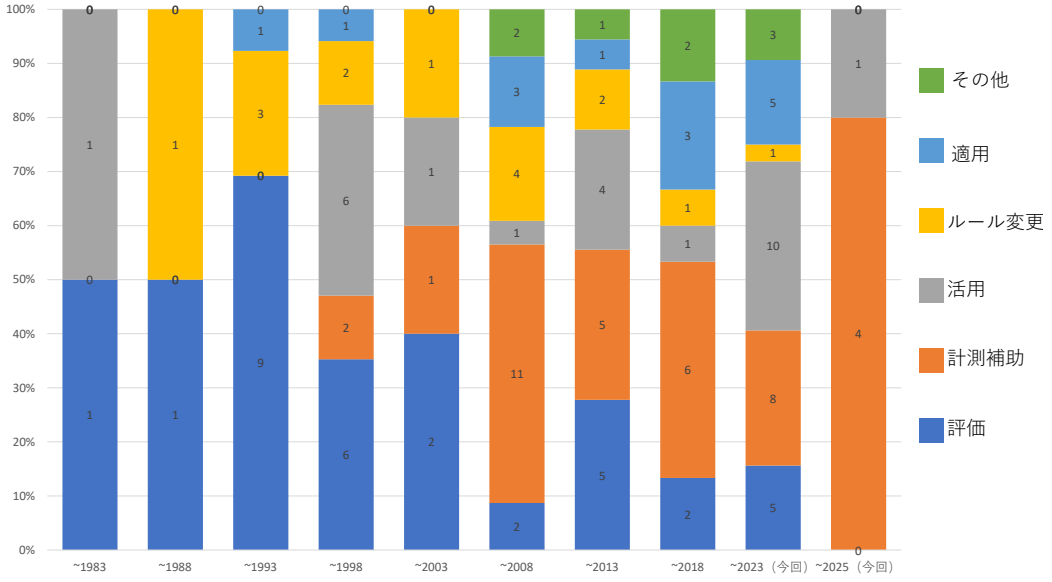
1979 - 1988 年は対象論文数が少なく（各期間 2 本），初期は「評価」および「活用・ルール変更」に限定された議論が中心となる．1989 - 1993 年には論文数が増加し（13 本），「評価」が過半（9/13）を占める．続く 1994 - 1998 年では 17 本へ増加し，「評価」と「活用」が同程度（各 6 本）となり，FP の有効性検証と実務的な利用方法の整理が並行して進む．

1999 - 2003 年は一時的に論文数が減少する（5 本）一方で，2004 - 2008 年には 23 本まで増加し，「計測補助」が主要トピックとなる（11/23）．この傾向は 2009 年以降も継続し，2009 - 2013 年は「評価」と「計測補助」が拮抗（各 5 本）した後，2014 - 2018 年では「計測補助」が最大（6/15）となる．また，2014 - 2018 年では「適用」や「その他」も一定数含まれており（それぞれ 3 本，2 本），FP の適用範囲拡大や周辺技術との組合せを志向した研究が見られる．

2019 - 2023 年は本調査期間内で論文数が最も多く（32 本），「活用」（10 本）と「計測補助」（8 本）を中心にしつつ，「評価」および「適用」も同程度（各 5 本）含む多様な分布となる．一方，2024 - 2025 年は収集対象の期間が短いこともあり論文数が少なく（5 本），「計測補助」への偏り（4/5）が見られる．



(a) 出版年代ごとの各トピックの数量



(b) 出版年代ごとの各トピックの割合

図 2: 出版年代ごとの各トピックの数量と割合

4.2.3　結果の要約

RQ1 では，2019 - 2025 年の FP 関連研究 37 本を 6 トピックに分類し，1979 - 2018 年の先行研究 95 本との比較および年代別推移を整理した．その結果，直近期間の研究関心は「計測補助（12 本）」と「計測結果の活用（11 本）」に集中しており，両者で全体の過半を占める主要トピックであることが確認された．一方，「FP の利点や欠点の評価（5 本）」および「計測ルールの変更（1 本）」は少数であり，FP 手法そのものの検証や新規ルール提案は相対的に周縁化していることが示された．

さらに，先行研究との構成比比較では，「計測補助」と「活用」の比率が増加し，「評価」および「ルール変更」の比率が大きく減少した．すなわち，研究の重心は「FP 手法の妥当性や計測規則そのもの」から，「計測の効率化・自動化」および「計測結果の見積・分析への接続」へ移行していることが明確となった．

年代別推移の分析からは，初期は「評価」中心であったが，2004 年以降は「計測補助」が主要トピックとして定着し，2019 - 2023 年には「活用」と「計測補助」が同時に増加して多様な分布を形成することが確認された．また，直近の 2024 - 2025 年は収集期間が短いことも影響しつつ，「計測補助」への偏りが観測された．以上より，2019 年以降の FP 研究は，自動化・効率化と活用局面に重点を置く方向へ継続的にシフトしていることが示された．

4.3　RQ2

4.3.1　成熟度（進展段階）の定義と判定基準

RQ2 では，各トピックに含まれる研究が，成熟度（進展段階）の観点からどのように構成されているかを明らかにする．本研究における進展段階は，研究成果が「分析・提案・実装・実運用」のどこまで到達しているかを表す枠組みとして定義した．具体的には以下の 4 段階を用いる．

- **A：手法無し**
  新たな手法・手順・尺度の提案を伴わず，既存手法の比較，性能評価，レビュー，統計的分析，分類整理を主目的とする研究．
- **B：手法提案（のみ）**
  新たな手法・モデル・枠組みを提案するが，ツール・システムとしての実装物が提示されていない研究．
- **C：ツール・システム開発（まで）**
  提案手法を実装し，ツール・システムとして動作する形で実験・検証を行っている研究（ただし，実務での定常運用が明示されない場合）．
- **D：実証適用（実際に適用）**

提案ツール・システムが，組織の業務プロセスに組み込まれ，日常的・継続的に利用されていることが本文中で明示される研究.

D の判定については，単に「産業データで評価した」「実プロジェクトでケーススタディを行った」といった記述のみでは該当とせず，日常運用・定常利用・業務フローへの統合などが明示されている場合に限定した.

4.3.2 進展段階の全体分布

まず，対象論文 37 本における進展段階の全体分布を表 5 に示す.

表 5: 進展段階別の論文数と割合（2019 - 2025 年，total=37）

進展段階	本数	割合
A：手法無し	7	18.9%
B：手法提案（のみ）	18	48.6%
C：ツール・システム開発（まで）	11	29.7%
D：実証適用（実際に適用）	1	2.7%

表 5 より，全体として B（手法提案）が約半数を占め，C（実装まで）が約 3 割を占める．一方で，D（実証適用）は 1 本（2.7%）に留まり，実装研究が一定数存在するにもかかわらず，業務への定常導入まで確認できる研究は極めて限定的である．また，A（手法無し）も約 2 割存在し，FP 手法そのものや既存推定枠組みを対象とする分析的研究が一定割合で残っていることが分かる.

4.3.3 トピック内における進展段階の分布

各トピックに含まれる研究を進展段階別に集計した結果を表 6 に示す．本表は，トピックごとに研究が「分析中心（手法無し）」か「提案中心（手法提案）」か「実装中心（ツール開発・実証適用）」かを比較するための基礎集計である.

さらに，トピック内の構成比として把握できるよう，表 7 に割合を示す．この表により，トピック間の成熟度差を定量的に比較できる.

4.3.4 派生指標による成熟度の比較

系統的レビューの結果提示では，単純な件数表に加えて，「実装に到達した割合」を派生指標として示すことで，研究の成熟度構造を明確にすることが多い．そこで，本研究で以下の指標を導入して結果

表 6: トピック内における進展段階の分布（2019 - 2025 年）

トピック	手法無し	手法提案	ツール開発	実証適用	計
FP の利点や欠点の評価	5	0	0	0	5
計測補助	0	5	6	1	12
計測結果の活用	0	9	2	0	11
計測ルールの変更	0	1	0	0	1
計測が難しい対象への適用	1	2	2	0	5
その他	1	1	1	0	3
計	7	18	11	1	37

表 7: トピック内における進展段階の割合

トピック	手法無し	手法提案	ツール開発	実証適用
FP の利点や欠点の評価	100.0%	0.0%	0.0%	0.0%
計測補助	0.0%	41.7%	50.0%	8.3%
計測結果の活用	0.0%	81.8%	18.2%	0.0%
計測ルールの変更	0.0%	100.0%	0.0%	0.0%
計測が難しい対象への適用	20.0%	40.0%	40.0%	0.0%
その他	33.3%	33.3%	33.3%	0.0%

として整理する．

実装到達率：$(C + D)$/total（実装・運用まで到達した研究の割合）．

表 8 に，各トピックの実装到達率を示す．これにより，トピック間で「提案中心」か「実装中心」かを一目で比較できる．

表 8 より，「計測補助」は実装到達率が 58.3% と最も高く，提案研究に留まらずツール・システム実装まで進む研究が相対的に多い．「計測が難しい対象への適用」は 40.0% であり，提案と実装が同程度に現れる構成である．一方，「計測結果の活用」は 18.2% と低く，B（手法提案）が支配的であることが改めて確認できる．「評価」および「ルール変更」は実装到達率が 0 であり，分析的研究（評価）または規則・手順の提案（ルール変更）に集中している．

4.3.5 トピック別の成熟度内訳

以下では，表 6 の分布に基づき，各トピック内でどの段階の研究が中心かを，論文集合の構成として具体的に示す．

表 8: トピック別の実装到達率（C+D）

トピック	C+D（本数）	実装到達率
FP の利点や欠点の評価	0	0.0%
計測補助	7	58.3%
計測結果の活用	2	18.2%
計測ルールの変更	0	0.0%
計測が難しい対象への適用	2	40.0%
その他	1	33.3%

FP の利点や欠点の評価（A. 手法無し に集中）「評価」トピック（5 本）はすべて A に分類され，トピック内構成は A=100% である（表 7）．この集合は，FP 法の適用限界や推定モデルの性能差を明らかにすることを目的とした比較・分析研究から構成される．例えば，文献調査により FP の特性と特定アプリケーション領域の要求を対比する研究 [40] や，工数見積り研究を体系化し FP を含む従来枠組みの限界を整理する研究 [41]，IFPUG と COSMIC の測定差を統計的に分析する研究 [42] などが含まれる．

計測補助（B. 手法提案・C. ツール開発 が中心，D. 実証適用 は 1 本）「計測補助」（12 本）は，B=41.7%，C=50.0%，D=8.3% であり（表 7），提案と実装が中心となる成熟度を示す．B に分類される研究群 [12, 13, 27, 47, 48] は，推定モデルを提示し，精度指標や誤差指標により有効性を示す形式が中心である．一方，C に分類される研究群 [25, 26, 45, 46, 50, 51] は，要求文書からの要素抽出や計測支援の処理パイプラインを実装し，プロトタイプとして動作する形で評価を行う．これらは産業ケースや実データを用いた検証を含む場合があるが，業務プロセスへの定常統合が明示されないため C に留めている．D に分類されるのは [49] のみであり，ツールが FP 計測チームの日常業務に統合されたことが本文で明示される点で，他のケーススタディ型研究と区別される．このように，計測補助は「実装まで到達する研究が多い」一方で，「運用まで到達した研究は限定的」であるという二層構造を結果として示している．

計測結果の活用（B. 手法提案 が支配的）「活用」（11 本）は B=81.8% が支配的であり（表 7），FP 計測結果を工数見積り・計画・モデル化に接続する研究の多くが，手法の提案と比較実験により構成されることを示す．具体的には，既存近似計測法の有効性比較 [53]，外れ値処理や前処理設計を通じた推

定精度改善 [58]，極端に簡略化した尺度（#TF 等）を用いる推定枠組み [62] などが B に分類される．C に分類される研究は 2 本 [52, 55] であり，具体的に実装し，実行可能な形で評価する点で B の多数派と区別される．ただし，本トピックでは D（定常適用）に該当する研究は確認されなかった．

計測ルールの変更（B. 手法提案 のみ）「ルール変更」は 1 本 [63] のみであり，B に分類される．この結果は，本調査対象期間において，計測規則の再定義・校正に関する研究が少数であること，およびその内容が主として規則や調整因子の提案に留まっていることを示す．

計測が難しい対象への適用（B. 手法提案・C. ツール開発 が並立）「適用」(5 本) は A=20%, B=40%, C=40% であり，提案と実装が同程度に含まれる構成である（表 7）．A に分類される [67] は，ファンクションポイントと工数の関係を大規模に検証する実証研究であり，新規計測法の提案を目的としないため A に位置付く．B に分類される [64, 66] は，適用対象の特性（アジャイル調達，モバイル等）を踏まえた尺度・手順の提示が中心であり，ツール実装の提示は主目的ではない．C に分類される [65, 68] は，提案枠組みをシステムとして実装し，実験・評価を行う点で B と区別される．ただし，本トピックでも D が明示された研究は見られない．

その他（A. 手法無し・B. 手法提案・C. ツール開発 が分散）「その他」(3 本) は A・B・C が各 1 本ずつであり，段階が分散する．文献レビュー [70] は A，評価文脈での方法提案 [69] は B，実装を伴うシステム枠組み [71] は C に分類される．このトピックは対象数が少ないため，一般化は避けつつ，少数ながら多様な成熟度段階の研究が含まれることを結果として示す．

4.3.6 図による可視化

表の結果を直観的に比較するため，トピック別の進展段階割合を積み上げ棒グラフとして可視化することが有効である．図 3 はその例であり，「活用」が B に偏ること，「計測補助」が C を中心に幅広い段階を含むこと，「評価」が A に集中することを視覚的に確認できる．

4.3.7 結果の要約

RQ2 では，2019 - 2025 年の FP 関連研究 37 本を，成熟度（進展段階）A-D（手法無し・手法提案・ツール開発・実証適用）で整理した．その結果，全体では B（手法提案）が約半数，C（ツール開発）が約 3 割を占め，提案研究が中心である一方，D（実証適用）は 1 本に留まった．すなわち，実装に到達した研究が一定数存在するにもかかわらず，業務プロセスへの定常統合まで本文上で確認できる研究は極めて限定的である．

トピック別に見ると，成熟度構成には明確な差が確認された．「評価」は A に集中し，既存枠組みの

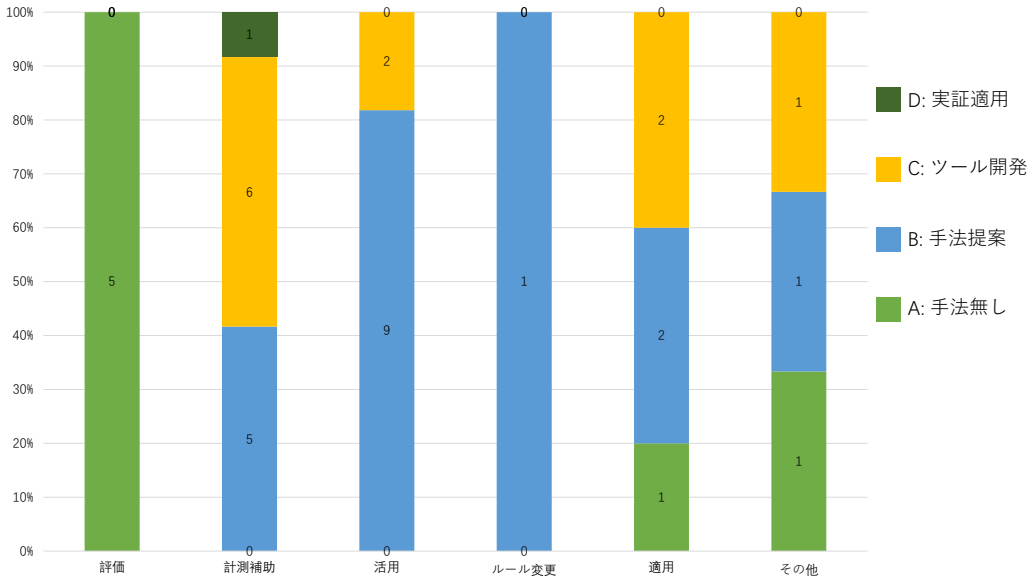


図 3: トピック別の進展段階割合

比較・分析が主である．「活用」は B が支配的であり，計測結果を工数見積り等へ接続する枠組み提案が中心となる一方，実装まで到達した研究は少数である．これに対して「計測補助」では B と C が中心で，実装到達率（C+D）が最も高い．ただし，このトピックでも定常運用（D）に該当する研究は 1 本のみであり，「実装研究の増加」と「運用定着の希少性」が併存する二層構造が示された．

以上より，2019 - 2025 年の FP 研究は，トピックにより成熟度の到達点が異なり，特に「運用まで到達した証拠」が研究全体として不足していることが明らかとなった．

4.4 RQ3

4.4.1 評価対象と評価観点

RQ3 では，提案手法の有効性評価がどの程度厳密に行われているかを把握するため，RQ2 で進展段階 A（手法無し）に分類された研究を除外し，少なくとも手法提案以上（B・C・D）に該当する研究を評価対象とした．対象論文 37 本のうち A は 7 本であるため，RQ3 の評価対象は 30 本となる．

有効性評価の厳密性は，ソフトウェア工学分野の妥当性議論に照らし，以下の 4 観点を満たすかどうかを Y・N（Yes・No）で記録した．なお，判定は論文本文・付録・オンライン補足資料（公開 URL，リポジトリ等）に基づく．記述が曖昧で第三者が確認できない場合は，過大評価を避けるため N とする．

- **(1) データセットの多様性・規模**

  評価データについて，(i) データの出所（企業データ・公開データ等）または収集方法が明記され，(ii) 規模数（プロジェクト数，要求文数，事例数など）が明記され，(iii) 複数プロジェクト

（または複数データセット）を含む，のすべてを満たす場合を Y とする．いずれかが欠ける場合（例：出所のみで規模不明，単一事例のみ等）は N とする．

- **(2) 比較対象の妥当性**

  提案手法が，少なくとも 1 つの比較対象（既存法または明示的ベースライン）と同一データ・同一評価指標・同一評価手順で比較されており，かつ比較対象の選定理由が説明されている場合を Y とする．ここで「妥当な比較対象」とは，実務・研究で広く用いられる代表的手法，または先行研究により標準的ベースラインとして位置づくものを含み，「提案手法に有利な比較対象」を恣意的に選ぶ（弱すぎるベースラインのみ，主要ベースラインの欠落，条件を揃えない比較など）ことを避けた比較を指す．比較が無い，条件不一致，選定理由の欠如，または上記のような恣意性が疑われる場合は N とする．

- **(3) 統計的検証の有無**

  手法間の差について，(i) 有意性検定（例：Wilcoxon 検定等）を実施している，または (ii) 信頼区間や効果量を提示している，のいずれかを満たす場合を Y とする．評価指標（平均誤差等）のみで，不確実性・差の根拠が示されない場合は N とする．

- **(4) 再現性**

  第三者が追試可能となるように，(i) データ，(ii) コード（実装・スクリプト），(iii) 手順（前処理，設定，パラメータ等の再実行に必要な詳細）のうち，少なくとも 1 つが公開され，入手可能である場合を Y とする．公開が一切無い場合や，入手不能（URL 不通・非公開等）の場合は N とする．

4.4.2　評価段階（評価無し・下・中・上）の定義

4 観点を満たすかどうかの Y・N 記録に基づき，評価対象 30 本を「評価段階」へ集約した．具体的には，有効性評価そのものを実施していないものを「評価無し」とし，それ以外は Y の数に応じて次の 3 段階に分類した．すなわち，Y が 0 - 1 個の研究を「評価下」，Y が 2 - 3 個の研究を「評価中」，Y が 4 個すべての研究を「評価上」とした．

4.4.3　評価段階の全体分布

表 9 に，評価対象 30 本の評価段階別内訳を示す．

表 9 より，評価対象の大半は「評価下」または「評価中」に集中している．すなわち，多くの研究で何らかの実験・検証は行われている一方で，4 観点すべてを満たす厳密な評価（評価上）は少数（2 本）に留まる．また，「評価無し」も 2 本存在し，提案・実装が提示されても有効性を定量的に示す枠組みが欠落するケースが確認された．

表 9: 評価段階別の論文数と割合（RQ3 対象：total=30）

評価段階	本数	割合
1. 評価無し	2	6.7%
2. 評価下	14	46.7%
3. 評価中	12	40.0%
4. 評価上	2	6.7%

4.4.4 トピック別評価段階（Evidence Gap Map）の結果

研究トピック（RQ1 の 6 分類）と評価段階（1 - 4）を組み合わせ，Evidence Gap Map（エビデンスの不足・偏りを可視化するマップ）として整理した結果を表 10 に示す．なお，RQ3 は進展段階 A を除外するため，「評価」トピックは全数が対象外となり，該当セルは 0 となる．

表 10: Evidence Gap Map：トピック別評価段階（論文数）

トピック	評価無し	評価下	評価中	評価上	計
FP の利点や欠点の評価	0	0	0	0	0
計測補助	1	5	5	1	12
計測結果の活用	0	5	5	1	11
計測ルールの変更	0	1	0	0	1
計測が難しい対象への適用	1	1	2	0	4
その他	0	2	0	0	2
計	2	14	12	2	30

表 10 から，トピックごとに評価段階の分布が異なることが読み取れる．「計測補助」と「計測結果の活用」はいずれも件数が多く，それぞれ「評価下」と「評価中」が同程度（各 5 本）含まれる一方，「評価上」に該当する研究は各トピックで 1 本のみである．また，「計測が難しい対象への適用」では「評価中」が 2 本ある一方，「評価無し」が 1 本含まれ，評価実施のばらつきが相対的に大きい．「その他」は少数（2 本）ながら両方が「評価下」に留まる．「計測ルールの変更」は 1 本のみであり，「評価下」に位置付く．

4.4.5　4 観点（データ・比較・統計・再現性）の達成状況

次に，評価段階の根拠となる 4 観点そのものの達成状況を集計する．「評価無し」を除く 28 本（評価下・中・上）について，4 観点の Y・N 記録を集計した結果を表 11 に示す．

表 11: 4 観点別の達成状況（評価を実施した 28 本）

評価観点	Y の本数	割合
データセットの多様性・規模	15	53.6%
比較対象の妥当性	15	53.6%
統計的検証の有無	14	50.0%
再現性（公開）	3	10.7%

表 11 より，データ・比較・統計は半数前後で Y が確認される一方，再現性だけが著しく低く，Y は 3 本に留まる．すなわち，多くの研究が実験結果は提示するものの，第三者が追試可能な形での公開（データ・コード・手順の公開）は限定的である．

4.4.6　評価段階別の「欠けやすい要素」

系統的レビューでは，単に段階別件数を示すだけでなく，段階が異なるときに「どの要素が欠けやすいか」を併せて示すことで，厳密性のボトルネックを明確化することが多い．そこで，本研究でも評価段階（評価下・中・上）ごとに 4 観点の達成状況を整理した（表 12）．ここでは「評価下＝ Y が 0 - 1 個（14 本）」「評価中＝ Y が 2 - 3 個（12 本）」「評価上＝ Y が 4 個（2 本）」として集計した．

表 12: 評価段階別の 4 観点達成状況（評価を実施した 28 本）

評価段階	本数	データ Y	比較 Y	統計 Y	再現 Y
評価下（0 - 1 個）	14	5	1	2	0
評価中（2 - 3 個）	12	8	12	10	1
評価上（4 個）	2	2	2	2	2

表 12 より，「評価下」に分類された研究では，比較対象の妥当性（1/14）および統計的検証（2/14）が特に欠けやすく，再現性は 0/14 である．一方，「評価中」では比較対象（12/12）と統計（10/12）が整備される傾向が見られるが，再現性は依然として 1/12 に留まる．「評価上」は 2 本のみだが，4 観点

すべてを満たしている．

これらの分布は，評価段階を引き上げる際に「どこを優先して整備すべきか」の目安も与える．まず「評価無し」から「評価下」への移行では，有効性評価を実施し，4 観点のうち少なくとも 1 つを満たして Y を得ることが必要である．表 12 では，「評価下」においてデータ観点の Y が最も多い（5/14）ことから，最初の一歩としては，評価データの出所・収集方法と規模（件数）を明記し，可能であれば複数事例を含めるなど，データセットの多様性・規模に関する根拠を整備することが，比較的取り組みやすい改善といえる．さらに余力があれば，同一データ上でのベースライン比較や，不確実性を示す統計的根拠（検定・効果量・区間推定のいずれか）を付与することで，厳密性の底上げに繋がる．

次に「評価下」から「評価中」への移行では，Y の不足が生じやすい観点を優先して埋めることが重要となる．表 12 が示すように，「評価下」では比較 Y が 1/14 と極端に少なく，統計 Y も 2/14 に留まるのに対し，「評価中」では比較 Y が 12/12，統計 Y が 10/12 と大きく改善している．すなわち，評価下の研究が評価中へ到達するうえで最も効果が大きいのは，(i) 妥当なベースラインを設定した公平な比較の導入，および (ii) 評価指標に留まらない統計的根拠の提示である．

一方で，再現性（公開）は「評価中」でも 1/12 に留まり，段階が上がっても欠けやすい要素である．以上より，再現性の不足は「評価下」だけでなく「評価中」でも残りやすい主要ギャップであり，評価厳密性を「評価上」へ引き上げる際の最後のボトルネックとして表れる．

4.4.7　トピック別の評価段階の特徴

以下では，表 10 の分布に基づき，各トピックの評価段階がどのように構成されているかを結果として具体化する．ここでは「なぜそうなったか」の要因分析は行わず，どのような評価設計・証拠提示が多いかを観測結果として整理する．

計測補助（評価下・中が中心，評価無しと評価上が各 1 本）「計測補助」（12 本）は，評価下 5 本・評価中 5 本が中心であり，評価無しと評価上が各 1 本ずつ存在する．評価中に分類された研究では，複数プロジェクト（または複数データセット）による評価，明示的なベースライン比較，統計検定（Wilcoxon 等）や効果量の提示など，少なくとも 2 観点以上が満たされる構成が多い [12, 45, 47]．一方，評価下に分類された研究では，比較が同一枠組み内の変種に限定される，あるいは強い既存法と条件を揃えた対照が不足する，統計的根拠が評価指標に留まるなど，厳密性を押し下げる要素が複数確認される [25, 26, 27, 50]．また，評価無しに分類された研究として，提案ツールの機能紹介に留まり有効性評価を実施しない例が含まれる [51]．評価上は 1 本であり，データ・比較・統計・再現性が同時に満たされた研究として位置付く [13]．

計測結果の活用（評価下・中が同数，評価上が 1 本）「活用」（11 本）は評価下 5 本・評価中 5 本と拮抗し，評価上が 1 本含まれる．評価中に分類された研究では，多データセット（あるいは大規模データ）上での比較，統計的検証の導入，比較設計が研究主張に整合する形で組まれる傾向が見られる [55, 60, 61, 62]．一方，評価下に分類された研究では，単一データソースへの依存や比較対象の狭さ，統計検証の欠如などが重なりやすい [54, 56, 58, 59]．評価上に分類される研究は 1 本であり，比較の設計と統計的根拠が揃った形で提示されている [52]．

計測が難しい対象への適用（評価無しを含み，ばらつきが大きい）「適用」は RQ1 では 5 本だが，そのうち進展段階 A を除外するため，RQ3 対象は 4 本となる．内訳は評価無し 1 本，評価下 1 本，評価中 2 本である．評価中に分類された研究では，複数組織データや複数尺度との比較，統計的検証が取り入れられる例が見られる [64]．一方，評価無しとしては，改良案の提示に留まり実データによる有効性検証を伴わない研究が含まれる [66]．また，評価下の研究では，比較は行うが統計的根拠や外的妥当性が不足するなど，厳密性が限定される [65]．再現性については，本トピックでは一部でレプリケーションパッケージの提供が確認される例がある [68] ものの，トピック全体としては少数である．

計測ルールの変更・その他（少数だが評価下に集中）「ルール変更」は 1 本のみであり評価下に分類された [63]．データが単一事例に限定され，比較・統計・公開の観点が整備されない形である．「その他」も RQ3 対象は 2 本であり，いずれも評価下に分類される [69, 71]．これらは提案内容の提示はあるが，評価設計（比較，統計，公開）の整備が限定的であることが共通している．

4.4.8 結果の要約

RQ3 では，手法提案以上（B・C・D）に該当する 30 本を対象に，有効性評価の厳密性について，4 観点（データセットの多様性・規模，比較対象の妥当性，統計的検証，再現性）を満たすかどうかを Y・N（Yes・No）で判定し，その結果に基づき評価段階（評価無し・下・中・上）へ集約した．その結果，評価対象の大半は「評価下」および「評価中」に集中し，4 観点をすべて満たす「評価上」は少数に留まった．また，提案・実装が提示されても有効性評価自体が欠落する「評価無し」も存在し，評価設計の不在が一部で確認された．

EGM の結果から，研究が集中する「計測補助」「活用」においても，評価段階は下位〜中位に偏っており，厳密な評価（評価上）は各トピックで限定的であることが示された．さらに 4 観点別集計では，データ・比較・統計は半数前後で整備される一方，再現性（データ・コード・手順の公開）は著しく低く，評価厳密性の主要なボトルネックとして最も顕著に表れた．

評価段階別に見ると，「評価下」では比較と統計が特に欠けやすく，再現性はほぼ確認されない．一

方,「評価中」では比較と統計が整備される傾向が見られるが，再現性は依然として低く，上位段階へ移行する際の最後の障壁として残りやすいことが確認された．以上より，2019 - 2025 年の FP 研究では有効性評価は広く実施されているものの，厳密性には段階差が大きく，とりわけ再現性確保の不足が継続的課題であることが明らかとなった．

4.5 RQ4

4.5.1 先行研究で指摘された未解決課題の整理

RQ4 では，2018 年時点で指摘された未解決課題が，2019 - 2025 年の研究においてどの程度緩和・解消されたかを整理する．ここでいう「未解決課題」とは，研究トピックの多寡そのものではなく，研究成果を実務へ移転する際に障壁となる論点（導入判断や運用定着を阻害する要因）を指す．

先行研究では，実務側の主要なニーズとして，(1) 汎用的自動計測ツールの不在，(2) 保守案件への適用困難，(3) 生産性評価モデルの未確立，(4) 発注側への普及不足，(5) FP で計測できない要素（非機能等）への対応不足，(6) 有効な教育（普及）手法の不在，が重要な未解決課題として挙げられていた．さらに，研究成果の適用可能性を判断するために必要なプロジェクト・コンテキスト情報（開発言語，業種，規模等）の記載不足が広範に存在し，実務導入の判断材料が欠落している点も，横断的な課題として位置づけられていた．

本レビューでは，これらの未解決課題を 2019 - 2025 年の文献集合に対応付け，課題ごとに「解消」「部分的解消」「未解消」の 3 段階で状況を整理する．「解消」は，当該課題に関する研究が複数存在し，かつ提案が特定事例に閉じず，一般化の見通し（適用範囲の明示，複数事例での検証，運用定着の報告等）まで踏み込んでいる状態を指す．「部分的解消」は，提案・実装・検証に関する蓄積が確認できる一方で，対象範囲の限定や運用・普及の未到達などにより，実務適用に向けたギャップが残存する状態を指す．「未解消」は，対応研究が乏しい，または研究が存在しても課題の核心に対する改善が確認できない状態を指す．

4.5.2 解消状況の全体像

表 13 に，主要な未解決課題と 2019 - 2025 年における解消状況をまとめる．本表は，RQ1（研究関心の集中領域），RQ2（成熟度），RQ3（評価の厳密性）の結果を踏まえ，未解決課題が「どの領域の研究によって」「どの段階まで」押し上げられたかを俯瞰するための統合表である．

4.5.3 主要 3 課題の解消状況

以下では，本レビューで重要度が高いと判断した 3 課題（コンテキスト記載，教育（普及），非機能等への対応）について，2019 - 2025 年の文献集合における対応状況をより詳細に整理する．

表 13: 未解決課題の解消状況（2019 - 2025 年の文献集合に基づく整理）

未解決課題	解消状況	2019 - 2025 年における観測（結果の要点）
コンテキストの記載不足（開発言語・規模・業種等）	未解消	データ出所の説明はある一方で，適用判断に必要なコンテキスト項目が十分に揃わない例が多い．企業内データや NDA 下データに依存する研究も多く，外的妥当性の説明が弱い傾向が残る．
有効な教育（普及）手法の不在	未解消	自動化・活用・精度改善に研究関心が集中し，教育手法を体系化して比較・検証する研究は見当たらない．
FP で計測できない要素（非機能等）への対応	未解消	当該課題を主題として扱い，FP の枠組み側の拡張や再設計に踏み込む研究はほとんど確認できない．非機能要因への言及は散見されるが，実務で参照可能な手順・指針として整理された知見は乏しい．
汎用的自動計測ツールの不在	部分的解消	計測補助領域でツール開発研究が増加し，一部で実務統合も報告されるが，汎用性・標準準拠・広範な運用実績の観点ではギャップが残る．
保守案件への適用困難	未解消	保守を中心課題として FP 手順へ反映する枠組みを提示する研究は見当たらない．そのため，保守への適用困難という障壁は十分に緩和されたとは言い難い．
生産性評価モデルの未確立	未解消	工数推定の研究は増加しているが，生産性を安定して比較・移転可能にする評価モデルとしての合意形成は弱い．
発注側企業への普及不足	未解消	発注側のメリット提示，合意形成，契約実務との接続を主題化する研究は見当たらない．

(1) コンテキスト（開発言語・規模・業種等）の記載不足（未解消） コンテキストの記載不足は，研究成果の移転可能性を判断できないという点で，実務適用の根本的なボトルネックとなる．2019 - 2025 年の研究では，評価に用いたデータの出所（企業名やデータ種別等）には言及される場合が多いが，開発言語・業種・アーキテクチャ・規模といった「適用判断に必要な最低限の項目」が揃って報告される例は限定的である．とくに企業内データを用いる研究では，秘密保持契約（Non-Disclosure Agreement; NDA）等の制約によりデータ公開や詳細記述が難しく，その結果として「どの条件で成立する知見か」が明確化されないまま結論が提示されるケースが残存する．

この傾向は RQ3（有効性評価の厳密性）の結果とも整合する．すなわち，データセットの多様性や規模の条件を満たす研究は限定的であり，外的妥当性を補強する情報が十分でない研究が依然として多い．また，再現性（データ・コード・手順の公開）も少数にとどまるため，第三者検証を通じたコンテキスト依存性の確認が進みにくい．以上より，コンテキスト報告の不足は 2019 年以降も明確に残存していると整理できる．解消に向けては，少なくとも報告テンプレートの導入と，適用範囲を明示する記述作法の標準化が必要である（提案は考察で扱う）．

(2) 有効な教育（普及）手法の不在（未解消） 2019 - 2025 年の研究動向（RQ1）では，「計測補助」および「計測結果の活用」が中心であり，教育・普及を主目的とした研究は主要トピックとして十分に形成されていない．FP 計測の学習プロセスを体系化し，教育介入の有効性を比較検証する研究も確認できない．このため，先行研究が指摘した「教育研究の蓄積が乏しい」というギャップは，本レビュー範囲では大きく改善していないと整理される．

ただし，本調査期間では計測補助ツールに関する研究が増加しており，これらを教育支援へ転用できる可能性（理由提示，判断支援，ログに基づく弱点分析等）が相対的に高まっている点は補足できる．一方で，現時点では教育手法そのものを主題化し，方法論として比較・確立する研究が十分に蓄積しているとは言い難い．

(3) FP で計測できない要素（非機能等）への対応（未解消） FP は機能規模の計測を主眼とするため，非機能要件（性能，信頼性，環境制約等）や開発環境に依存する要因を，FP の計測結果に一貫して取り込むことが難しいという課題が指摘されてきた．2019 - 2025 年の研究では，非機能要因を説明変数として推定モデルに追加する，あるいは特定領域で追加尺度を導入する試みが散見される．

しかし，これらは「FP で計測できない要素」を FP の計測枠組みとして扱えるようにするものではなく，追加特徴量や別尺度として個別に付与するに留まる場合が多い．その結果，どの非機能要因をどの条件でどのように扱うべきかという実務導入上の障壁に対して，一般に参照可能な形で改善が確認できる段階には至っていない．したがって，本課題は未解消と位置づけられる．

4.5.4　その他の課題

先行研究が挙げたその他の課題についても，本レビュー範囲では以下の傾向が得られる．

汎用的自動計測ツールの不在（部分的解消）　2019 - 2025 年では計測補助トピックが最大規模であり，ツール・システム開発（RQ2 の C）も相対的に多い．さらに，一部では実務への統合（RQ2 の D）も報告され，研究としての前進は確認できる．しかし，多くは特定ドメイン・特定組織のデータや工程に依存しており，標準準拠かつ汎用ツールとして広く導入される状態に到達しているとは言い難い．従って，本課題は部分的解消に留まる．

保守案件への適用困難（未解消）　本レビュー範囲では，保守を主要対象として位置付け，FP の計測手順や解釈を保守実務に合わせて具体化する研究は見当たらない．改修・拡張に関連する推定研究や変更規模の尺度を扱う研究は見られるものの，保守固有の論点（影響分析，変更波及，既存資産制約，技術的負債等）を FP の枠組みとして一貫して扱うことを主目的としていない場合が多い．以上より，保守適用に関する障壁は未解消と整理される．

生産性評価モデルの未確立・発注側企業への普及不足（未解消）　工数推定研究の増加は確認できるが，生産性評価を「異なる組織・条件間でも比較可能なモデル」として安定化させる研究や，発注側の合意形成・契約実務との接続を主題化する研究は見当たらない．従って，これらの課題はいずれも未解消と整理される．

4.5.5　結果の要約

RQ4 では，2018 年時点で指摘された未解決課題を 2019 - 2025 年の文献集合（37 本）に対応付け，「解消・部分的解消・未解消」の 3 段階で整理した．その結果，実務適用の前提となるコンテキスト情報（開発言語・規模・業種等）の記載不足と，教育（普及）手法の不在は改善が限定的であり，未解消と判断される．とくにコンテキストについては，データ出所への言及はある一方で適用判断に必要な最低限項目が揃わない例が多く，RQ3 で確認された再現性の限定とも整合して，第三者検証を通じた移転可能性の確認が進みにくい構造が残存している．

一方，汎用的自動計測ツールの不在については，関連研究の増加，提案・実装，実務統合が観測され，部分的解消と整理できる．ただし，多くの研究は特定ドメインや特定組織の前提に依存しており，汎用性，標準化，再現可能性，広範な運用実績の観点でギャップが残るため，解消と断定できる段階には至っていない．

これに対して，FP で扱いにくい要素（非機能等）への対応は，当該課題を主題として扱う研究がほ

とんど確認できず，FP の枠組みとしての整理・統合も進んでいないことから未解消と判断される．また，保守案件への適用困難についても，保守を中心課題として扱う研究が不足であり，実務で参照可能な適用指針としての蓄積が十分でないため未解消と整理される．さらに，生産性評価モデルの確立および発注側企業への普及は，工数推定研究の増加にもかかわらず，合意形成や契約実務との接続を主題化する研究が少なく，未解消と整理される．

以上より，2019 - 2025 年の研究は自動化や適用拡張の側面では前進が認められる一方，実務適用を左右するコンテキスト報告，教育（普及），再現性に関する不足が主要なギャップとして残存していることが示された．

5　考察

本節では，2019 - 2025 年の FP 関連研究 37 本を対象とした SLR 結果を統合し，研究動向の形成要因，現在の研究ポートフォリオが抱える構造的な課題，ならびに改善に向けた具体的方策を議論する．ここでの考察は，複数 RQ の結果を束ねて「なぜその状況が生じているか」「何がボトルネックか」「どう改善できるか」を軸に整理する．

なお，本節で参照する根拠は，主として RQ1（トピック分布と研究関心の偏り），RQ2（進展段階による成熟度分布），RQ3（有効性評価の厳密性と再現性），RQ4（未解消課題の残存状況）である．各小節では，議論の出発点となる RQ 結果（複数に跨る場合は併記）を明示したうえで，その背景要因と含意を整理し，改善の方向性へ接続する．

5.1　研究関心の再配置が示す FP の役割変化

本小節は主として RQ1（トピック分布）に基づき，必要に応じて RQ4（未解消課題の残存）で整理した「計測負担・運用上の障壁」の観測結果を踏まえて考察する．本調査期間において研究が「計測補助」と「計測結果の活用」に集中したことは，FP が単なる規模尺度として成熟した結果，研究焦点が「尺度そのものの是非」から「現代の開発現場で使える形への再統合」へ移ったことを示唆する．この再配置は偶然ではなく，開発現場の前提条件が変化したことにより，従来の論点が「優劣の議論」だけでは解きにくくなったためと解釈できる．近年は，短いリリース周期での計画更新，意思決定の高速化，データに基づく管理の重視が進み，見積もり・比較・説明責任を素早く回すこと自体が要求される．その一方で，FP は手作業中心の計測プロセスが残りやすく，専門判断と工数の負担がボトルネックになりやすい（RQ4 で残存課題として整理）．したがって，尺度の妥当性を抽象的に論じるよりも，「限られた時間と情報で，どこまで一貫した計測を実現し，運用へ組み込めるか」が研究課題として前面に出やすくなる．このギャップが，計測の自動化・支援へ研究を促進した主因であると位置付けられる．

同時に，活用領域の増加は，FP が「入力特徴量」として再評価されていることと整合する（RQ1 で「活用」研究が一定割合を占めた）．この背景には，実務データの蓄積と学習モデルの普及により，「単一尺度で一回だけ推定する」形から，「複数の属性と組み合わせて継続的に推定・最適化・説明へ接続する」形へ移行した事情がある．言い換えると，現場が求めるのは FP の理論的完成度そのものというよりも，意思決定に耐える説明可能性を保ちながら，取得コストを抑えて継続運用できる情報である．そのため，FP は単独の尺度としてよりも，プロジェクト属性や派生尺度と併用しやすい「説明変数」として使われやすい．特に，簡易尺度（#TF など）との比較や代替可能性の検証が複数見られる点は，FP が「理論的に優れた尺度」かどうかではなく，「実務の制約下で，どの情報がどの程度のコストで得られ，どの程度の性能と説明力を提供するか」という観点で再設計されつつあることを示している

[13, 47, 62]．以上より，「活用」へ重心が移ったのは，FP が不要になったからではなく，意思決定の場で使うために，周辺情報と統合される形へ役割が変化したためと整理できる．

一方で，評価やルール変更が相対的に少ないこと（RQ1）は，FP の基本枠組みが成熟し，ルール自体の大改変が研究として成立しにくくなった可能性を示す．ただし，これは「問題が解消した」ことを意味しない．むしろ，現代の開発文脈（アジャイル，生成 AI，マイクロサービス等）では，従来の前提が崩れやすく，適用限界の再点検やルールの局所的な補正が必要になる局面が増えている．評価・ルール変更が少数である現状は，研究関心の偏りによって「変化する前提への検証」が追いついていない可能性も含むため，将来的な再活性化の余地を残す領域といえる．

5.2　成熟度ギャップ：提案・実装から定常運用への移行

本小節は主として RQ2（進展段階の分布）に基づき，補助的に RQ3（再現性・外的妥当性の制約）および RQ4（運用移転を阻む残存課題）で観測された要因と接続して考察する．成熟度（進展段階）の分布を見ると，研究の多くが「手法提案（B)」と「ツール開発（C)」に集中し，「実証適用（D)」が極めて少ない．この傾向は第 4 章 RQ2 の進展段階分析（5）でも確認でき，トピックを問わず B・C で成果が止まりやすい構図が示された．この構造は，研究コミュニティにおける成果の出し方（論文として成立しやすい貢献）と，組織内での定常運用に必要な条件（統合コスト，責任分界，継続保守）が一致していないことに起因する可能性が高い．ここから本節では，「なぜ D まで到達しにくいのか」を，RQ2 の観測事実を起点に整理する．

第一に，FP 計測は見積り・契約・工数管理に直結し，誤りが生じた場合の影響が大きい（RQ4 で残存課題として整理した「運用上のリスク」)．研究としては精度指標や効率改善を示せても，運用要件（例外処理，監査可能性，責任分界，継続保守）まで満たす作り込みには追加コストが大きい．結果として，論文ではツール開発と限定的評価で止まりやすい．

第二に，実務統合にはデータ連携が不可欠であり，要求管理・設計文書・リポジトリといった複数システムとの接続が必要になる．しかし，こうした統合は組織固有の環境に強く依存し，研究成果を一般化しにくく，また論文としての再現性も確保しにくい（RQ3 で再現性の低さが顕在化)．本調査でも，産業データや NDA 制約の言及が多く，公開可能な形での再利用に壁がある．そのため，「実装はあるが再現性が弱い」研究が増え，結果として D の蓄積が進みにくい．

第三に，実証適用の判定において本研究が厳格な基準（定常運用の明示）を採用したことも重要である（RQ2 の分類基準)．多くの研究が「実プロジェクトで評価」と記述しても，それが「導入評価」なのか「日常業務での常態利用」なのかは別である．本調査で D に分類される研究が少数であったことは，運用レベルの根拠提示が論文上不十分であるか，あるいは実際に運用まで到達した研究が少ないかのいずれかを意味する．前者の場合，運用の有無を記述する作法（採用前後の比較，運用期間，利用頻

度，運用上の例外対応など）が研究報告として定着していない課題が残る（RQ4 の「移転・定着」観点とも整合）．

改善策としては，研究成果を D へ押し上げるための「移行パス」を明確化する必要がある．例えば，(i) 手法提案（B），(ii) 再利用可能なプロトタイプ（C），(iii) 限定部署での試験導入，(iv) 定常運用（D）という段階を想定し，各段階で満たすべき要件（精度，説明可能性，運用負荷）をチェックリスト化することが有効である．これにより，研究者側は「論文化に必要な最小要件」と「運用に必要な追加要件」を切り分けられ，組織側も導入判断に必要な情報を得やすくなる．特に計測補助のような人間の判断が残る領域では，完全自動化を目指すより，「確認と修正を前提に支援する」として設計し，誤り時の影響を局所化することが導入障壁を下げる．この方向性は，実務統合を明示した研究の位置付けとも整合する．

5.3　エビデンスの質におけるボトルネック：外的妥当性と再現性

本小節は主として RQ3（有効性評価の厳密性）に基づき，とりわけ「再現性（公開）」と「外的妥当性（一般化可能性）」に関する観測結果を起点に考察する．RQ3 の結果は，評価が「存在する」ことと「厳密である」ことの間に大きな隔たりがあることを示した．データ・比較・統計の 3 観点は半数前後で満たされる一方，再現性（公開）が極端に低い．

まず，本節で扱う外的妥当性・再現性の問題は，FP 研究に固有の課題というより，ソフトウェア工学の実証研究一般に広く見られる構造的制約に根差す．具体的には，企業データの秘匿性，商用データベースの利用制限，実験環境・前処理の依存性といった要因が，データや手順の共有を難しくし，研究間比較と追試を阻害する．その上で FP 研究では，要求文書，計測結果，生産性やコストに直結する属性が機密になりやすく，さらに派生尺度やデータ形態の多様性も大きいことから，SE 一般の制約が相対的に強く現れやすい．

また，これらの問題が 2019 年以降に「悪化した」と断定できるかは慎重に扱う必要がある．本レビューの範囲では，再現性の低さが顕著であることは確認できるが，過去期間との厳密な時系列比較（同一基準での縦断評価）を行っていないため，悪化の有無そのものは結論づけられない．一方で，2019 年以降は AI・機械学習を含むデータ駆動型の研究が増え，モデル・特徴量・前処理・学習設定など共有すべき要素が増大した結果，公開が不十分な場合に再現性欠如がより目立ちやすいという側面はある．したがって本研究では，「悪化した」とは言わず，近年の研究潮流の中で再現性がボトルネックとして顕在化しやすくなっている，という位置付けで整理する．

外的妥当性（一般化可能性）については，単一組織・単一ドメインに依存するデータが多いことが主要因となる．企業内データを用いる研究は，現実的で価値が高い一方，開発言語・規模・業種・プロセスといった条件が変わると性能が変動する可能性がある．この点を補うためには，複数データセットでの検証，あるいは公開データによる再検証が必要であるが，データ公開が難しいために，研究間の横比

較が進まない．結果として，研究は増えても「信頼できる結果」が形成されにくい．

比較の妥当性については，「研究目的に整合した対照」を設計できている研究と，「公正性を欠く対照」や「同一枠組み内の変種比較」に留まる研究が混在していた（RQ3）．前者は，問題設定に合うベースラインを明示し，同一データ・同一手順で公平に比較しようとする．一方後者では，強い既存手法との比較が欠落し，改善幅が過大に見積もられるリスクがある．これは，FP 領域が多様な派生尺度・データ形態を持ち，分野横断で共有される「標準的ベースライン」や評価プロトコルの合意が形成されにくいこととも関連する．

統計的検証については，検定や効果量の導入が進んでいる研究がある一方，評価指標のみで「改善した」と結論づける研究も残る（RQ3）．評価指標の多様性も，結論の比較可能性を下げる要因となる．

最も大きなボトルネックは再現性である（RQ3）．再現性が低いと，第三者による追試・再分析ができず，系統的レビューとしても研究間比較の根拠が弱くなる．この課題は「データが公開できないから仕方がない」で終わらない．例えば，(i) 匿名化・集約化した特徴量レベルの公開，(ii) 前処理・モデル学習・評価のスクリプト公開，(iii) データ入手手順（ISBSG 等）の明示と処理後データの生成手順公開，(iv) 擬似データによるパイプライン検証，といった代替手段がある．特に，コード公開や評価手順の固定化は，データ公開が困難でも再現性を引き上げる現実的な方策になり得る．

5.4　コンテキスト報告の標準化による「研究の累積性」への影響

本小節は主として RQ4（未解消課題の残存状況）に基づき，RQ3（外的妥当性・再現性）との関係を明示しながら考察する．RQ4 で「未解消」と整理されたコンテキスト記載不足は，単なる記述の問題ではなく，エビデンスの外的妥当性と再現性を同時に損ねる構造要因である．多くの研究がデータ出所を述べても，開発言語，業種，規模，工程，チーム構成などが欠落すれば，「どの条件で有効か」を判断できない．さらに，コンテキストが不足すると，後続研究が同条件で追試することも難しくなり，結果として研究成果が点として散在し，累積しない．

この問題は，FP 研究が扱う対象が多様であるほど深刻化する．例えば，同じ要件文書でも，金融・行政・組込みでは用語や要求粒度が異なり，NLP ベースの抽出器の性能は変わり得る [45, 49]．また，ISBSG のような多様なデータを使う場合でも，フィルタ条件（品質ランク，開発種別，新規・改修）により分布が変わり，結論が変わり得る．したがって，コンテキスト報告は「望ましい」ではなく，研究の比較可能性を担保するための必須要件と位置付ける必要がある．

改善策として，本研究が示唆するのは，最低限の「コンテキスト報告テンプレート」を必須項目として固定することである．具体的には，(1) 対象の種類（新規・改修，ウォーターフォール・アジャイル等），(2) ドメイン（業種・システム種別），(3) 規模分布（FP，要求文数，タスク数など主要な規模指標），(4) 技術スタック（言語・プラットフォーム），(5) データ入手と前処理（匿名化の有無，除外条

件)，(6) 評価設計（評価指標）を最低限セットとして明示する．このテンプレートは，比較の前提条件を揃えることでレビューの統合可能性を高め，実務側が適用判断を行うための材料にもなる．

一方で，「テンプレートを作るべき」という主張は正論である反面，現実に実装できるのか，という問いが生じる．これまで標準化が進みにくかった理由として，第一に，FP 研究の多くが企業データや契約制約を伴い，業種や規模，工程などの情報が機微情報として扱われやすい点がある．第二に，研究目的が自動計測，活用，評価などに分散しており，どの項目を必須とするかの合意形成が難しかった点も大きい．第三に，論文誌・会議ごとのフォーマットやページ制約が異なり，統一的な記載枠組みが採用されにくかった，という運用上の障壁もある．したがって，標準化は「誰かが良い案を出す」だけでは進まず，採用を促す主体と実行手段を併せて設計する必要がある．

標準化を担う主体としては，(a) FP 計測法のコミュニティ（IFPUG 等の標準化団体)，(b) ソフトウェア工学の学術コミュニティ（査読付き会議・論文誌)，(c) データ提供者を含む実務コミュニティ（ISBSG 等）の三者が現実的な候補になり得る．例えば，論文誌・会議が「コンテキスト報告チェックリスト」を投稿要件または推奨項目として明示すれば，研究者側の記述行動は比較的速く変わり得る．一方，標準化団体がテンプレート（必須項目と任意項目）を公開し，実務データの秘匿性に配慮した記載例（範囲表現，カテゴリ化，匿名化の粒度）を示せば，企業データでも採用可能性が上がる．ISBSG のようなデータ基盤が，データ利用ガイドとして推奨テンプレートを提示することも，研究と実務の接続点として有効である．

実現可能性の観点では，「完全な統一」を狙うより，(1) 最小必須セット，(2) 分野別の追加項目，(3) 秘匿性を確保した記載ルール，の三層構造にすることが現実的である．最小必須セットは，本節で述べた 6 項目のように比較可能性へ直結する要素に絞り，具体値が出せない場合はレンジやカテゴリ（例：規模は区分，業種は分類）で代替する．そのうえで，NLP による抽出研究なら要件文書の粒度やアノテーション手順，見積り活用研究なら組織の意思決定プロセスや運用頻度など，トピックに応じた追加項目を任意で付す．以上により，ページ制約や機密制約を踏まえつつも，最低限の比較可能性と追試可能性を確保でき，結果として FP 研究の「累積性」を底上げする基盤となる．

5.5　教育・普及の空白：研究対象としての位置付け不足

本小節は主として RQ4（教育・普及が未解消として残る）に基づき，RQ1（当該トピックが相対的に少数である）および RQ2（運用定着が進みにくい成熟度分布）との関係として考察する．教育（普及）手法の未整備は，FP が標準であるにもかかわらず「一貫性・信頼性が揺らぐ」問題の根にある．FP は定義の解釈や境界判断を含むため，学習者の理解の差が計測結果のばらつきにつながる．にもかかわらず，2019 - 2025 年の研究関心は自動化・活用・精度改善に集中し，教育介入を体系的に設計・比較・評価する研究は依然として少ない（RQ1，RQ4)．これは，教育研究が（i）短期で成果が出にくい，（ii）

研究評価が難しい，（iii）産業データほど注目されにくい，という構造的要因を持つためと考えられる．

ただし，教育・普及は「別領域の問題」ではなく，計測補助研究の成果を定着させるための鍵でもある．自動化ツールが導入されても，計測者が結果を理解し，誤りを検知・修正できなければ，組織は運用リスクを抱える．したがって，計測補助ツールは「計測代替」だけでなく，教育補助として再設計できる余地が大きい．さらに，教育は成果が見えにくい一方で，既存の学習支援・説明可能性・運用ログ活用といった蓄積を応用できるため，研究テーマとして具体化し得る．ここでは，(1)-(4) の方向性について，それぞれ「何を足せば研究として成立するか」と「実現可能性の見通し」を併せて述べる．

具体的には，(1) 判断理由の提示（なぜ ILF・EIF と判定したか，参照箇所はどこか）は，説明可能性（根拠提示）やレビュー支援の既存アプローチを FP 判断に持ち込む形で実装可能である．要件文中の根拠スパン（参照箇所）を抽出し，判定ラベルと併せて提示する設計にすれば，学習者は「どの記述が境界判断に効いたか」を追跡できる．実現面では，ツールが内部で用いた特徴や参照箇所を出力できるようにし，熟練者の判断根拠（レビューコメント等）と突き合わせて整合性を検証することで，教育効果の前提となる妥当性を担保しやすい．

(2) 迷いやすい判断点のガイド（境界条件，例外パターン）は，ルールブックの単純な再掲ではなく，誤りが多い論点に焦点化した「学習順序」と「分岐条件」を設計対象にできる．例えば，典型的な曖昧表現や例外パターンを類型化し，該当する要件記述が出たときに注意喚起と確認質問を出す形は，既存のガイド付き入力やチェックリスト型支援の延長として実現できる．研究としては，どの論点を優先して提示すべきか（学習効果と負荷のトレードオフ）を実験設計で比較でき，短期的にも評価可能な単位（特定の論点セット）へ分割しやすい．

(3) 修正ログの蓄積と弱点分析（どの種別で誤りが多いか，誰がどこで迷うか）は，運用ログや学習分析の考え方をそのまま適用できる．計測結果の修正履歴（差分）と理由（コメント）を最小限の形式で収集できれば，誤りの集中箇所（論点・要件タイプ・判定カテゴリ）や学習曲線を定量化できる．実現面でも，入力と出力を保存するだけで始められ，個人情報や機密の扱いが難しい場合は，匿名化・抽象化（テンプレ化した要件例）で段階的にデータ整備できる．そのうえで，ログから得た「つまずき分布」を (2) のガイド設計へ還流させることで，教育介入を反復的に改善する研究サイクルが作りやすい．

(4) ケースベース学習（類似要件の参照）は，検索・推薦に基づく学習支援として位置付けられる．既存の要件例とその FP 判断（加えて根拠スパンや注意点）をケースとして蓄積し，新規要件に対して類似ケースを提示すれば，「似た例から学ぶ」学習が可能になる．実現可能性は，まず少数の代表ケースから開始し，(3) のログで増分的にケースを拡充する構成にすれば高い．また，類似度の誤りが問題になる場合も，学習者が参照ケースを評価・修正する仕組み（フィードバック）を併設すれば，運用しながら品質を上げられる．

以上の (1)-(4) は，個別に完結するというより，相互に接続した「教育補助としての計測補助」を構成

できる点に意義がある．例えば，(1) の根拠提示で理解を支え，(2) で論点を誘導し，(3) で弱点を可視化して介入を最適化し，(4) で例示学習を回す，という統合が考えられる．この方向は，D（実務統合）の希少性（RQ2）とも関係する．教育補助としての価値を明確化できれば，組織は「精度が完全でなくても，説明可能性と学習効果がある」ことを理由に導入しやすくなる．すなわち，教育は独立トピックとしてだけでなく，運用移転（B/C→D）を促進する横断的要素として位置付けるべきである．

5.6 「FP で測れない要素」への対応研究の不足と統合原理の欠如

本小節は主として RQ1（当該観点を主題とする研究の少なさ）および RQ4（未解消課題の残存）に基づき，その解釈上の限界も含めて考察する．非機能要素や FP で計測しにくい要因への対応については，本調査範囲では，当該課題を主題として扱う研究がほとんど見当たらない．ただし，本結果の解釈には注意が必要である．本レビューは検索語に「function point」を用いており，いわゆる「FP で測れない要素」への対応は，品質特性，リスク，技術負債，運用負荷などの語で議論され，必ずしも「function point」を明示しない研究として蓄積されている可能性がある．そのため，本調査での「少なさ」は，研究の実在数だけでなく，検索設計上の検出限界を含むことを明確にしておく必要がある．

そのうえで，FP の活用分野をどう捉えるかを整理すると，近年の見積り・意思決定の場面では，FP が入力情報として支配的な影響を持つ局面が確かに存在する．例えば，初期見積り，契約・発注，生産性分析，ベンチマーク，ポートフォリオ管理では，FP が共通尺度として機能し，議論の出発点になりやすい．一方，実務で必要とされる説明力は FP 単独では満たしにくく，FP 以外の要素（非機能，要求不確実性，運用・保守特性等）も，同時に重要になる．したがって，本論点は「FP の代替」ではなく，「FP を軸にしつつ，追加要因をどのように位置付け，接続するか」という統合問題として捉えるのが適切である．

モバイルやアジャイルなど従来前提と異なる対象に言及し，追加要因を取り込んで推定へ接続する試みはごく少数に留まる．しかし，これらは散発的な対応に留まり，実務で参照可能な手順・指針として整理された知見の蓄積は乏しいため，本課題は未解消と整理される（RQ4）．その結果として，議論するための共通土台自体が十分に形成されていない．

例えば，非機能要因を導入しても，要因の定義や測定方法が研究ごとに異なれば，再現性・比較可能性が失われる．また，非機能の寄与はドメイン依存が大きく，コンテキスト報告が不足すると一般化が困難になる（RQ4）．結局，非機能対応は「尺度拡張」だけでは不十分であり，(1) 要因カタログの標準化，(2) 要因の測り方の合意，(3) FP と要因の統合モデル（加算・補正・階層モデル等）の比較，(4) どの条件で何が有効かの整理，まで進めて初めて「解消」に近づく．この意味で，非機能対応は RQ4 の課題であると同時に，RQ3 と RQ4 が示した「比較可能性・再現性・条件記述」の基盤整備を前提とする領域でもある．

5.7　今後の改善に向けた統合的ロードマップ

本小節は RQ1-RQ4 の結果を統合し，とくに RQ2（成熟度ギャップ）と RQ3（再現性・評価厳密性の不足），RQ4（コンテキスト・教育の未解消）を「依存関係のある横断課題」として整理した上で，優先順位付きの改善案を提示する．以上の議論を踏まえると，2019 - 2025 年の FP 研究は量的には活発化しているが，(1) 運用定着の不足（成熟度ギャップ：RQ2），(2) エビデンスの厳密性の不足（再現性・外的妥当性：RQ3），(3) コンテキスト報告の不足（RQ4），(4) 教育・普及の空白（RQ4），という横断的課題によって，研究成果が累積しにくい構造を持つ．そこで本研究は，これらの課題を個別に足し算するのではなく，「研究の比較可能性を先に固め，次に再現性と外的妥当性を段階的に底上げし，その上で運用定着と適用範囲拡張へつなぐ」という順序で統合的に進めるロードマップを提案する．

本ロードマップは，(i)-(iii) を基盤整備として最優先に置き，それを前提に (iv) で運用定着（成熟度ギャップ：RQ2）を縮小し，(v) のような「FP で測れない要素」への拡張（RQ1・RQ4 で不足・未解消）は中長期で並走させる．図 4 に，優先順位，項目間の依存関係，および短期・中期・長期の実現イメージを示す．

以下では，図 4 の各項目を補足する．

(i) ベンチマーク化と評価プロトコルの固定化　比較対象の妥当性を高めるためには，タスクごとに最低限含めるべきベースライン（強い既存法，単純ベースライン，アブレーション）を合意し，評価手順をプロトコル化する必要がある（主に RQ3 の観測に対応）．特に，提案点が「追加要素の効果」なら，同一条件でのアブレーション比較を必須とし，偏りを避けるために重複検証や固定ルールを採用することが望ましい．これは (iii) の再現性確保や，(v) の統合モデル比較の前提にもなる．

(ii) コンテキスト報告テンプレートの必須化　研究の累積性を高める最短経路は，コンテキスト報告を「任意」から「必須」へ引き上げることである（RQ4）．テンプレートを導入すれば，外的妥当性の議論が明確になり，研究間比較も容易になる．また，実務側が導入判断を行う際の材料にもなるため，研究成果の移転可能性を同時に高められる．さらに (iii) で外部検証へ進む際にも，検証条件の記述を標準化できる．

(iii) 再現性の段階的確保　データ公開が難しい領域でも，コード，設定，前処理手順，評価スクリプトは公開可能な範囲が残ることが多い（RQ3）．まずは「実行可能な再現性」を確保し，次に匿名化・集約化した特徴量や合成データでパイプラインを検証し，最終的に公開データや共有契約の枠組みで外部検証を進める，という段階的戦略が現実的である．これにより，RQ3 で顕著だった再現性ギャップを埋められる．なお，この段階化は (i) のプロトコル固定と (ii) のコンテキスト標準化があるほど進めやすい．

(iv) 教育・普及を「研究成果の定着条件」として扱う　教育は独立研究としてだけでなく，計測補助ツールの導入・運用を支える条件として扱うべきである（RQ4，および成熟度ギャップを示した RQ2 と接続）．説明可能性と学習支援（理由提示，ガイド，ログ分析）を組み込むことで，ツールが「計測者の置き換え」ではなく「計測者の能力増強」として受け入れられやすくなる．結果として，成熟度ギャップの縮小にも寄与する．また，(iii) の再現可能な実装が共有されれば，教育・普及の教材化も進めやすい．

(v) 非機能・測れない要素の統合　非機能や測れない要素の対応は，本調査範囲では当該課題を主題として扱う研究自体が少なく（RQ1），現状の知見は実務指針として参照できる形に整理されていない（RQ4）．要因定義の共有と測定方法の合意，統合モデルの比較可能性，そして再現性・コンテキスト報告を伴う外部検証が揃って初めて，実務にとって利用可能な指針となる．したがって，非機能や測れない要素の対応は独立に先行させるよりも，(i)-(iii) で「比較できる・再現できる・条件が分かる」基盤を固めた上で，(iii) の外部検証と並行して中長期で積み上げる必要がある．

5.8　総括

本調査の結果は，FP 研究が近年，対象領域を拡大しながら急速に進展している一方で，運用定着（RQ2），エビデンスの厳密性（RQ3），コンテキスト報告（RQ4），教育（RQ4）といった横断要素が追いついていないことを示した．特に，研究成果を実務へ移転し，研究知見として蓄積するためには，(1) 評価プロトコルとベンチマークの整備，(2) 再現性確保の段階的戦略，(3) コンテキスト報告の必須化，(4) 教育・普及を含む運用設計を統合的に進める必要がある．これらは個別トピックの追加研究だけでは解消しにくく，研究コミュニティとしての作法・基盤整備を伴う課題である．その意味で，本研究が示した RQ1-RQ4 の結果は，FP 研究の「何が増えたか」だけでなく，「どの条件が整えば実務と研究のギャップを縮められるか」を，RQ ごとの観測に基づいて明確化した点に意義がある．

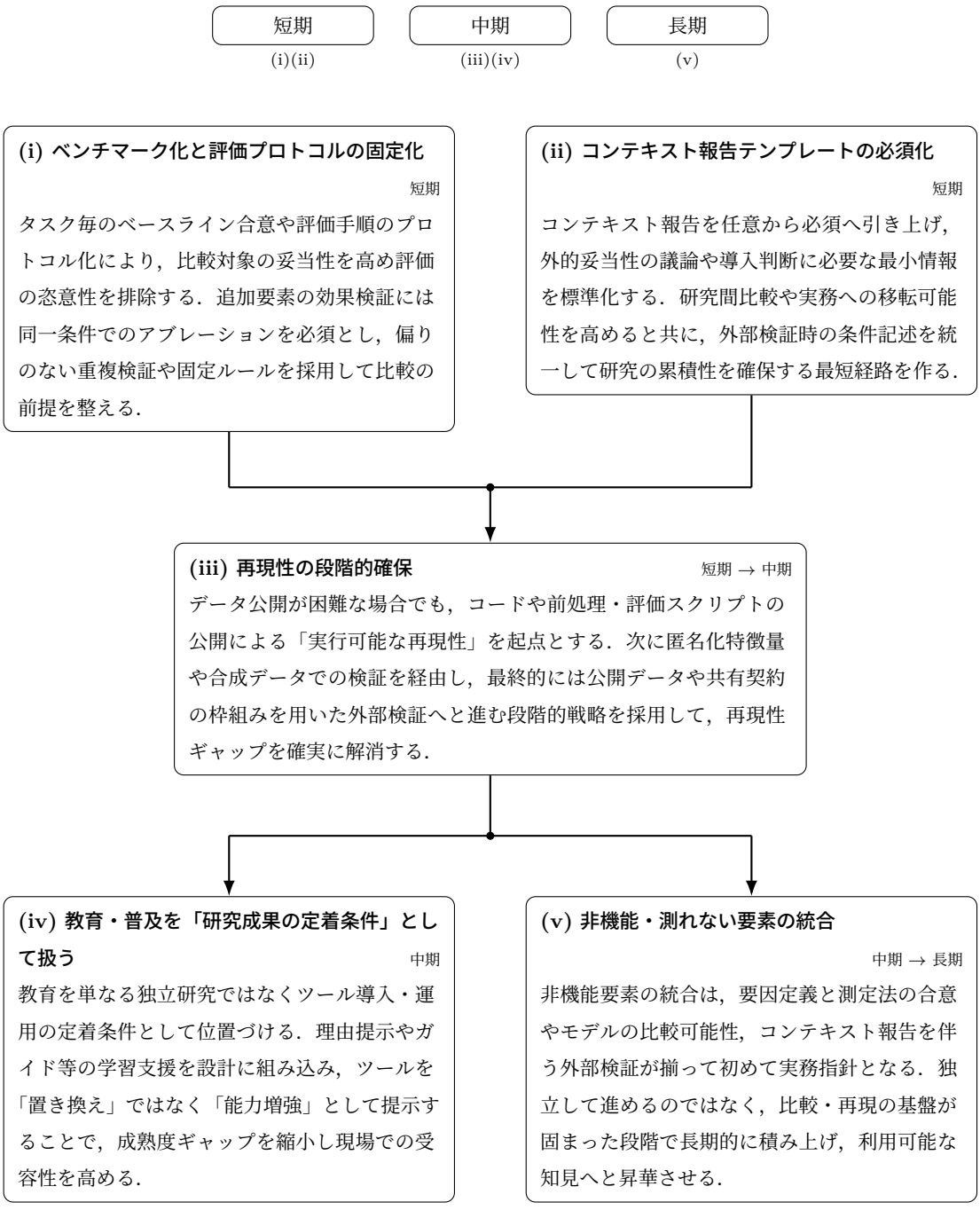


図 4: 統合的ロードマップ：優先順位・依存関係・時系列の実現イメージ

6　おわりに

本研究では，2019 - 2025 年に公表されたファンクションポイント法（FPA）関連研究を対象に，系統的文献レビューを実施した．先行研究（〜2018 年）以降，技術・実務環境が変化し，FP は手作業で算定する成果物であるだけでなく，見積りや意思決定の入力情報としても扱われるようになっている．こうした変化を踏まえ，近年の研究動向と，実務適用可能性を左右する根拠の整備状況を，一貫した観点で系統的文献レビューを実施した．

レビュー結果として，2019 - 2025 年の FP 関連研究は「計測補助」および「計測結果の活用」に集中し，研究関心が「尺度そのものの是非」や「ルール変更」から「現場で使える形への再統合」へ移行していることが確認された（RQ1）．一方で，成熟度の観点では手法提案・ツール開発まで到達する研究が多い反面，業務プロセスに組み込まれた日常的・継続的利用が明示される実証適用は極めて限定的であり，提案・実装から定常運用への移行にギャップが存在することが示された（RQ2）．また，有効性評価は広く実施されているものの，データ・比較・統計の整備にばらつきがあり，とりわけ再現性（データ・コード・手順の公開）が顕著に不足していた（RQ3）．さらに，先行研究で指摘された未解決課題のうち，コンテキスト報告の不足と教育（普及）手法の不在は依然として未解消であり，FP で計測しにくい要素（非機能等）への対応も体系的な蓄積が乏しいことが確認された（RQ4）．

本研究の貢献は，(i) 近年文献を 6 トピックで再整理し，研究関心の移行を明確化したこと，(ii) 進展段階（手法無し・手法提案・ツール開発・実証適用）により成熟度構造を可視化したこと，(iii) 有効性評価の厳密性を 4 観点（データ・比較・統計・再現性）で Y・N 記録し，評価段階と EGM（エビデンスの不足・偏りを可視化するマップ）として整理したこと，にある．これにより，研究が増加している領域と，根拠が厚い領域が必ずしも一致しない点，および再現性が横断的なボトルネックとなっている点を，レビュー結果として具体化した．

今後は，研究成果を実務へ移転し，研究の累積性を確保するために，(1) ベンチマーク化と評価プロトコルの固定化，(2) コンテキスト報告テンプレートの必須化，(3) データ公開が困難でも実行可能な再現性（コード・手順公開）を起点とした段階的戦略，を優先して整備する必要がある．その上で，(4) 教育・普及を研究成果の定着条件として位置づけ，実証適用へ接続する移行パスを設計することが重要である．また，(5) 非機能要因や保守特性など FP で捉えにくい要素は，標準化と比較可能性の基盤が整った段階で，中長期に統合的に積み上げるべき課題である．

FP 法は今後も実務で重要な尺度であり続ける一方，現代の開発環境では取得コストと説明可能性を両立した運用が求められる．本研究が示した整理とロードマップが，FP 研究の全体像理解と，研究と実務のギャップ縮小に向けた議論の土台として活用されることを期待する．

謝辞

本論文の執筆にあたり，多くの方々からご指導とご支援を賜りました．ここに記して，心より感謝申し上げます．

まず，指導教員である楠本真二教授に深く御礼申し上げます．修士課程の 2 年間を通じて，研究テーマの設定から，研究計画の立案，設計と実施，結果の解釈，考察の組み立てに至るまで，研究の進め方を基礎から丁寧にご指導いただきました．また，発表準備や練習の場では，限られた時間で要点を伝えるための構成，図表の見せ方，質疑応答での論点整理の仕方まで具体的に助言をいただき，発表に向けた実践的な訓練を積むことができました．研究上の悩みを相談した際にも，状況を整理した上で次に取るべき行動を示していただき，迷いなく研究を継続できました．

次に，柗本真佑准教授に感謝申し上げます．研究室における発表練習で，結論の飛躍や前提の抜けを鋭く指摘していただき，主張と根拠の対応関係を厳密に整えることができました．特に，結果の見せ方に関して，表現が曖昧になりやすい箇所を具体例とともに示しながら改善点を提示していただいたことは，本論文の可読性と説得力を高めるうえで大きな助けとなりました．また，研究室運営に関わる多くの業務を担ってくださり，学生が研究に集中できる環境が維持されていたことにも感謝いたします．

事務補佐員の橋本美砂子氏にも厚く御礼申し上げます．日常の事務手続きや各種連絡，提出物に関する対応など，研究以外の業務を丁寧かつ迅速に支えていただきました．細かな確認事項にも的確に対応していただいたことで，学生として安心して研究活動を進めることができました．

さらに，楠本研究室の皆様に感謝申し上げます．研究室内での議論や発表練習の機会を通じて，多様な視点からのコメントをいただき，本研究の問題設定と結論の妥当性を見直す機会を得ました．特に，修士課程の 2 年間にわたりチューターとしてご指導くださった藪下友氏，岡本琉生氏には，研究の進捗管理や論文執筆の進め方に関して，具体的かつ継続的な助言をいただきました．日本語の学術表現についても，言い回しの不自然さや論理関係が曖昧になる箇所を細かく指摘していただき，読み手に伝わる文章へと改善することができました．また，同期である玉置文人氏，忠谷晃佑氏，呂 蔚暄氏には，研究室生活の中で日常的に支えていただくとともに，困ったときに気軽に相談できる環境をつくっていただきました．議論や雑談を通じて気持ちを切り替えられたことも，研究を継続するうえで大きな支えとなりました．

最後に，家族ならびに友人に心より感謝いたします．研究が思うように進まず不安を感じたときや，締切が重なり負荷が高い時期にも，変わらず励ましの言葉をかけて支えてくれました．周囲の支えがあったからこそ，最後まで本論文をまとめることができました．

以上，ご指導・ご支援を賜りましたすべての皆様に，改めて深く感謝申し上げます．

参考文献

[1] Jørgensen, M. and Shepperd, M.: A Systematic Review of Software Development Cost Estimation Studies, *Transactions on Software Engineering*, Vol. 33, No. 1, pp. 33–53 (2007).

[2] Albrecht, A. J. and Gaffney, J., John E.: Software Function, Source Lines of Code, and Development Effort Prediction: A Software Science Validation, *Transactions on Software Engineering*, Vol. SE-9, No. 6, pp. 639–648 (1983).

[3] Van Hai, V., Nhung, L. T. K. and Hoc, H. T.: A Review of Software Effort Estimation by Using Functional Points Analysis, in *Proceedings of the International Conference on Advances in Systems and Software*, pp. 471–484 (2019).

[4] Freitas Junior, de M., Fantinato, M. and Sun, V.: Improvements to the Function Point Analysis Methods: A Systematic Literature Review, *Transactions on Engineering Management*, Vol. 62, No. 4, pp. 495–506 (2015).

[5] Kemerer, C. F. and Porter, B. S.: Improving the Reliability of Function Point Measurement: An Empirical Study, *Transactions on Software Engineering*, Vol. 18, No. 11, pp. 1011–1024 (1991).

[6] Lavazza, L.: Automated Function Points: Critical Evaluation and Discussion, in *Proceedings of the Workshop on Emerging Trends in Software Metrics*, pp. 15–21 (2015).

[7] Meli, R. and Santillo, L.: Function Point Estimation Methods: A Comparative Overview, in *Proceedings of the FESMA Conference*, pp. 1–10 (1999).

[8] Finnie, G. R., Wittig, G. E. and Desharnais, J.-M.: A Comparison of Software Effort Estimation Techniques: Using Function Points with Neural Networks, Case-Based Reasoning and Regression Models, *Journal on Systems and Software*, Vol. 39, No. 3, pp. 281–289 (1997).

[9] 山田涼太, 楠本真二, 肥後芳樹, 柗本真佑, 倉重誠：系統的文献レビューを用いたファンクションポイント研究の調査, 電子情報通信学会論文誌 D, Vol. J103-D, No. 3, pp. 144–158 (2020).

[10] Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J. and Linkman, S.: Systematic Literature Reviews in Software Engineering - A Systematic Literature Review, *Journal on Information and Software Technology*, Vol. 51, No. 1, pp. 7–15 (2009).

[11] Lavazza, L., Locoro, A. and Meli, R.: Software Development and Maintenance Effort Estimation Using Function Points and Simpler Functional Measures, *Journal on Software*, Vol. 3, No. 4, p. 22 (2024).

[12] Lavazza, L., Locoro, A., Liu, G. and Meli, R.: Estimating Software Functional Size via Machine

Learning, *Transactions on Software Engineering and Methodology*, Vol. 32, No. 5, pp. 1–27 (2023).

[13] Lavazza, L., Locoro, A. and Meli, R.: Using Machine Learning and Simplified Functional Measures to Estimate Software Development Effort, *Journal on IEEE Access*, Vol. 12, pp. 142505–142523 (2024).

[14] Diev, S.: Use Cases Modeling and Software Estimation: Applying Use Case Points, *Journal on ACM SIGSOFT Software Engineering Notes*, Vol. 31, No. 6, pp. 1–4 (2006).

[15] ISO/IEC International Standard - Software and Systems Engineering – Software Measurement – IFPUG Functional Size Measurement Method, *ISO/IEC 20926:2009(E)*, pp. 1–24 (2009).

[16] ISO/IEC International Standard - Software Engineering – Mk II Function Point Analysis – Counting Practices Manual, *ISO/IEC 20968:2002(E)*, pp. 1–93 (2002).

[17] ISO/IEC International Standard - Software Engineering – NESMA Functional Size Measurement Method – Definitions and Counting Guidelines for the Application of Function Point Analysis, *ISO/IEC 24570:2018(E)*, pp. 1–70 (2018).

[18] ISO/IEC International Standard - Software Engineering – COSMIC: A Functional Size Measurement Method, *ISO/IEC 19761:2011(E)*, pp. 1–14 (2011).

[19] (IFPUG), I. F. P. U. G.: *Function Point Counting Practices Manual*, International Function Point Users Group (IFPUG) (1986).

[20] Symons, C. R.: *Software Sizing and Estimating: Mk II FPA (Function Point Analysis)*, Wiley (1991).

[21] Netherlands Software Metrics Association (NESMA), : *Definitions and Counting Guidelines for the Application of Function Point Analysis*, Netherlands Software Metrics Association (NESMA) (1990).

[22] Abran, A., Desharnais, J.-M., Oligny, S., St-Pierre, D. and Symons, C. R.: *COSMIC-FFP Measurement Manual*, Software Engineering Management Research Laboratory, Université du Québec à Montréal (UQAM) (1999).

[23] Kemerer, C. F.: Reliability of function points measurement: A Field Experiment, *Journal on Communications of the ACM*, Vol. 36, No. 2, pp. 85–97 (1993).

[24] Czarnacka-Chrobot, B.: Standardization of Software Functional Size Measurement Methods, in *Proceedings of the Advanced Information Technologies for Management*, pp. 41–50 (2009).

[25] Zhang, K., Wang, X., Ren, J. and Liu, C.: Efficiency Improvement of Function Point-Based Software Size Estimation with Deep Learning Model, *Journal on IEEE Access*, Vol. 9, pp.

107124–107136 (2021).

[26] Zhao, Z., Jiang, H., Zhao, R. and He, B.: Emergence of a Novel Domain Expert a Generative AI-based Framework for Software Function Point Analysis, in *Proceedings of the International Conference on Automated Software Engineering*, pp. 2245–2250 (2024).

[27] Sun, S. and Tao, Y.: Research on a Software System Size Evaluation Model Combining BiLSTM-CRF with Domain Adaptation Learning, in *Proceedings of the International Conference Proceeding Series*, pp. 492–497 (2024).

[28] Dybå, T. and Dingsøyr, T.: Strength of Evidence in Systematic Reviews in Software Engineering, in *Proceedings of the International Symposium on Empirical Software Engineering and Measurement*, pp. 178–187 (2008).

[29] Hevner, A. R., March, S. T., Park, J. and Ram, S.: Design Science in Information Systems Research, *Journal on MIS Quarterly*, Vol. 28, No. 1, pp. 75–105 (2004).

[30] Peffers, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S.: A Design Science Research Methodology for Information Systems Research, *Journal on Management Information Systems*, Vol. 24, No. 3, pp. 45–77 (2007).

[31] Gorschek, T., Garre, P., Larsson, S. and Wohlin, C.: A Model for Technology Transfer in Practice, *Journal on IEEE Software*, Vol. 23, No. 6, pp. 88–95 (2006).

[32] Wieringa, R. J., Maiden, N., Mead, N. and Rolland, C.: Requirements Engineering Paper Classification and Evaluation Criteria: A Proposal and a Discussion, *Journal on Requirements Engineering*, Vol. 11, No. 1, pp. 102–107 (2006).

[33] Shaw, M.: Writing Good Software Engineering Research Paper, in *Proceedings of International Conference on Software Engineering*, pp. 726–737 (2003).

[34] Kitchenham, B. A. and Mendes, E.: Why Comparative Effort Prediction Studies May Be Invalid, in *Proceedings of the 5th International Conference on Predictor Models in Software Engineering*, pp. 1–5 (2009).

[35] Shepperd, M. and MacDonell, S. G.: Evaluating Prediction Systems in Software Project Estimation, *Journal on Information and Software Technology*, Vol. 54, No. 8, pp. 820–827 (2012).

[36] Whigham, P. A., Owen, C. A. and MacDonell, S. G.: A Baseline Model for Software Effort Estimation, *Transactions on Software Engineering and Methodology*, Vol. 24, No. 3, pp. 1–11 (2015).

[37] Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets, *Journal on Machine Learning Research*, Vol. 7, pp. 1–30 (2006).

[38] Association for Computing Machinery, : Artifact Review and Badging, ACM Publications Policy (2016), Accessed: 2026-01-22.

[39] Hirsch, J. E.: An Index to Quantify an Individual's Scientific Research Output, in *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102, pp. 16569–16572 (2005).

[40] Hillman, M. F. and Subriadi, A. P.: 40 Years Journey of Function Point Analysis: Against Real-Time and Multimedia Applications, *Journal on Procedia Computer Science*, Vol. 161, pp. 266–274 (2019).

[41] Pemmada, S. K., Behera, H. S., K, A. K., Nayak, J. and Naik, B.: Advancement from Neural Networks to Deep Learning in Software Effort Estimation: Perspective of Two Decades, *Journal on Computer Science Review*, Vol. 38, p. 100288 (2020).

[42] Quesada-López, C., Madrigal-Sánchez, D. and Jenkins, M.: An Empirical Analysis of IFPUG FPA and COSMIC FFP Measurement Methods, in *Proceedings of the Information Technology and Systems*, pp. 265–274 (2020).

[43] Hoc, H. T., Silhavy, R., Prokopova, Z. and Silhavy, P.: Comparing Multiple Linear Regression, Deep Learning and Multiple Perceptron for Functional Points Estimation, *Journal on IEEE Access*, Vol. 10, pp. 112187–112198 (2022).

[44] Sakhrawi, Z., Sellami, A. and Bouassida, N.: Investigating the Impact of Functional Size Measurement on Predicting Software Enhancement Effort Using Correlation-Based Feature Selection Algorithm and SVR Method, *Journal on Lecture Notes in Computer Science*, Vol. 12541, pp. 229–244 (2020).

[45] Li, M., Shi, L., Wang, Y., Wang, J., Wang, Q., Hu, J., Peng, X., Liao, W. and Pi, G.: Automated Data Function Extraction from Textual Requirements by Leveraging Semi-Supervised CRF and Language Model, *Journal on Information and Software Technology*, Vol. 143, p. 106770 (2022).

[46] Quesada-López, C., Martínez, A., Jenkins, M., Salas, L. C. and Gómez, J. C.: Automated Functional Size Measurement: A Multiple Case Study in the Industry, in *Proceedings of the Product-Focused Software Process Improvement*, pp. 263–279 (2019).

[47] Liu, G. and Lavazza, L.: Early and Quick Function Points Analysis: Evaluations and Proposals, *Journal on Systems and Software*, Vol. 174, p. 110888 (2021).

[48] Rankovic, N., Rankovic, D., Ivanovic, M. and Kaljevic, J.: Interpretable Software Estimation with Graph Neural Networks and Orthogonal Array Tunning Method, *Journal on Information*

Processing & Management, Vol. 61, No. 5, p. 103778 (2024).

[49] Shi, L., Li, M., Xing, M., Wang, Y., Wang, Q., Peng, X., Liao, W., Pi, G. and Wang, H.: Learning to Extract Transaction Function from Requirements: An Industrial Case on Financial Software, in *Proceedings of the European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1444–1454 (2020).

[50] Han, D., Gu, X., Zheng, C. and Li, G.: Research on Structured Extraction Method for Function Points Based on Event Extraction, *Journal on Electronics*, Vol. 11, No. 19, p. 3117 (2022).

[51] Bluemke, I. and Malanowska, A.: Tool for Assessment of Testing Effort, in *Proceedings of the Engineering in Dependability of Computer Systems and Networks*, pp. 69–79 (2020).

[52] Barros, M. d. O. and Gonçalves, V. P.: A Function Point Formulation for the Software Release Planning Problem, in *Proceedings of the International Symposium on Empirical Software Engineering and Measurement*, pp. 1–11 (2019).

[53] Di Martino, S., Ferrucci, F., Gravino, C. and Sarro, F.: Assessing the Effectiveness of Approximate Functional Sizing Approaches for Effort Estimation, *Journal on Information and Software Technology*, Vol. 123, p. 106308 (2020).

[54] Silhavy, P., Silhavy, R. and Prokopova, Z.: Categorical Variable Segmentation Model for Software Development Effort Estimation, *Journal on IEEE Access*, Vol. 7, pp. 9618–9626 (2019).

[55] Hoc, H. T., Silhavy, R., Prokopova, Z. and Silhavy, P.: Comparing Stacking Ensemble and Deep Learning for Software Project Effort Estimation, *Journal on IEEE Access*, Vol. 11, pp. 60590–60604 (2023).

[56] Prokopova, Z., Silhavy, P. and Silhavy, R.: Influence Analysis of Selected Factors in the Function Point Work Effort Estimation, in *Proceedings of the Intelligent Systems in Cybernetics and Automation Control Theory*, pp. 112–124 (2019).

[57] Ersoy, E., Bagriyanik, S. and Sozer, H.: On the Accuracy of Effort Estimations Based on COSMIC Functional Size Measurement: A Case Study, in *Proceedings of the Empirical Software Engineering and Measurement*, pp. 528–537 (2024).

[58] Silhavy, P., Silhavy, R. and Prokopova, Z.: Outliners Detection Method for Software Effort Estimation Models, in *Proceedings of the Software Engineering Methods in Intelligent Algorithms*, pp. 444–455 (2019).

[59] Silhavy, P., Silhavy, R. and Prokopova, Z.: Spectral Clustering Effect in Software Development Effort Estimation, *Journal on Symmetry*, Vol. 13, No. 11, p. 2119 (2021).

[60] Silhavy, P., Silhavy, R. and Prokopova, Z.: Stepwise Regression Clustering Method in Function Points Estimation, in *Proceedings of the Computational and Statistical Methods in Intelligent Systems*, pp. 333–340 (2019).

[61] Van Hai, V., Nhung, H. L. T. K., Prokopova, Z., Silhavy, R. and Silhavy, P.: Toward Improving the Efficiency of Software Development Effort Estimation via Clustering Analysis, *Journal on IEEE Access*, Vol. 10, pp. 83249–83264 (2022).

[62] Lavazza, L., Liu, G. and Meli, R.: Using Extremely Simplified Functional Size Measures for Effort Estimation: An Empirical Study, in *Proceedings of the Empirical Software Engineering and Measurement*, pp. 1–9 (2020).

[63] Effendi, A., Setiawan, R. and Rasjid, Z. E.: Adjustment Factor for Use Case Point Software Effort Estimation (Study Case: Student Desk Portal), *Journal on Procedia Computer Science*, Vol. 157, pp. 691–698 (2019).

[64] Rosa, W. and Jardine, S.: Data-Driven Agile Software Cost Estimation Models for DHS and DoD, *Journal on Systems and Software*, Vol. 203, p. 111739 (2023).

[65] Mushtaq, Z. and Wahid, A.: Inclusion of Functional and Non-Functional Parameters for the Prediction of Overall Efforts of Mobile Applications, *Journal on Computer Standards & Interfaces*, Vol. 71, p. 103404 (2020).

[66] Kaur, A. and Kaur, K.: Investigation on Test Effort Estimation of Mobile Applications: Systematic Literature Review and Survey, *Journal on Information and Software Technology*, Vol. 110, pp. 56–77 (2019).

[67] Tawosi, V., Moussa, R. and Sarro, F.: On the Relationship between Story Points and Development Effort in Agile Open-Source Software, in *Proceedings of the International Symposium on Empirical Software Engineering and Measurement*, pp. 183–194 (2022).

[68] Alhamed, M. and Storer, T.: Playing Planning Poker in Crowds: Human Computation of Software Effort Estimates, in *Proceedings of the International Conference on Software Engineering*, pp. 1–12 (2021).

[69] Tanabata, K., Hazeyama, A., Yamada, Y. and Furukawa, K.: Proposal of an Evaluation Method of Individual Contributions Using the Function Point in the Implementation Phase in Project-Based Learning of Software Development, *Journal on Procedia Computer Science*, Vol. 192, pp. 1524–1531 (2021).

[70] Bluemke, I. and Malanowska, A.: Software Testing Effort Estimation and Related Problems: A Systematic Literature Review, *Journal on ACM Computing Surveys*, Vol. 54, No. 3, pp.

1–38 (2022).

[71] Maia, A. W. and Farias, P. P. M.: Transactions as a Service, *Journal on Advances in Intelligent Systems and Computing*, Vol. 984, pp. 415–423 (2019).

付録 A 調査対象論文

[I] Hillman, M. F. and Subriadi, A. P.: 40 Years Journey of Function Point Analysis: Against Real-Time and Multimedia Applications, *Journal on Procedia Computer Science*, Vol. 161, pp. 266–274 (2019)

[II] Pemmada, S. K., Behera, H. S., K, A. K., Nayak, J. and Naik, B.: Advancement from Neural Networks to Deep Learning in Software Effort Estimation: Perspective of Two Decades, *Journal on Computer Science Review*, Vol. 38, p. 100288 (2020)

[III] Quesada-López, C., Madrigal-Sánchez, D. and Jenkins, M.: An Empirical Analysis of IFPUG FPA and COSMIC FFP Measurement Methods, in *Proceedings of the Information Technology and Systems*, pp. 265–274 (2020)

[IV] Hoc, H. T., Silhavy, R., Prokopova, Z. and Silhavy, P.: Comparing Multiple Linear Regression, Deep Learning and Multiple Perceptron for Functional Points Estimation, *Journal on IEEE Access*, Vol. 10, pp. 112187–112198 (2022)

[V] Sakhrawi, Z., Sellami, A. and Bouassida, N.: Investigating the Impact of Functional Size Measurement on Predicting Software Enhancement Effort Using Correlation-Based Feature Selection Algorithm and SVR Method, *Journal on Lecture Notes in Computer Science*, Vol. 12541, pp. 229–244 (2020)

[VI] Lavazza, L., Locoro, A., Liu, G. and Meli, R.: Estimating Software Functional Size via Machine Learning, *Transactions on Software Engineering and Methodology*, Vol. 32, No. 5, pp. 1–27 (2023)

[VII] Lavazza, L., Locoro, A. and Meli, R.: Using Machine Learning and Simplified Functional Measures to Estimate Software Development Effort, *Journal on IEEE Access*, Vol. 12, pp. 142505–142523 (2024)

[VIII] Zhang, K., Wang, X., Ren, J. and Liu, C.: Efficiency Improvement of Function Point-Based Software Size Estimation with Deep Learning Model, *Journal on IEEE Access*, Vol. 9, pp. 107124–107136 (2021)

[IX] Zhao, Z., Jiang, H., Zhao, R. and He, B.: Emergence of a Novel Domain Expert a Generative AI-based Framework for Software Function Point Analysis, in *Proceedings of the International Conference on Automated Software Engineering*, pp. 2245–2250 (2024)

[X] Sun, S. and Tao, Y.: Research on a Software System Size Evaluation Model Combining

BiLSTM-CRF with Domain Adaptation Learning, in *Proceedings of the International Conference Proceeding Series*, pp. 492–497 (2024)

[XI] Li, M., Shi, L., Wang, Y., Wang, J., Wang, Q., Hu, J., Peng, X., Liao, W. and Pi, G.: Automated Data Function Extraction from Textual Requirements by Leveraging Semi-Supervised CRF and Language Model, *Journal on Information and Software Technology*, Vol. 143, p. 106770 (2022)

[XII] Quesada-López, C., Martínez, A., Jenkins, M., Salas, L. C. and Gómez, J. C.: Automated Functional Size Measurement: A Multiple Case Study in the Industry, in *Proceedings of the Product-Focused Software Process Improvement*, pp. 263–279 (2019)

[XIII] Liu, G. and Lavazza, L.: Early and Quick Function Points Analysis: Evaluations and Proposals, *Journal on Systems and Software*, Vol. 174, p. 110888 (2021)

[XIV] Rankovic, N., Rankovic, D., Ivanovic, M. and Kaljevic, J.: Interpretable Software Estimation with Graph Neural Networks and Orthogonal Array Tunning Method, *Journal on Information Processing & Management*, Vol. 61, No. 5, p. 103778 (2024)

[XV] Shi, L., Li, M., Xing, M., Wang, Y., Wang, Q., Peng, X., Liao, W., Pi, G. and Wang, H.: Learning to Extract Transaction Function from Requirements: An Industrial Case on Financial Software, in *Proceedings of the European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1444–1454 (2020)

[XVI] Han, D., Gu, X., Zheng, C. and Li, G.: Research on Structured Extraction Method for Function Points Based on Event Extraction, *Journal on Electronics*, Vol. 11, No. 19, p. 3117 (2022)

[XVII] Bluemke, I. and Malanowska, A.: Tool for Assessment of Testing Effort, in *Proceedings of the Engineering in Dependability of Computer Systems and Networks*, pp. 69–79 (2020)

[XVIII] Barros, M. d. O. and Gonçalves, V. P.: A Function Point Formulation for the Software Release Planning Problem, in *Proceedings of the International Symposium on Empirical Software Engineering and Measurement*, pp. 1–11 (2019)

[XIX] Di Martino, S., Ferrucci, F., Gravino, C. and Sarro, F.: Assessing the Effectiveness of Approximate Functional Sizing Approaches for Effort Estimation, *Journal on Information and Software Technology*, Vol. 123, p. 106308 (2020)

[XX] Silhavy, P., Silhavy, R. and Prokopova, Z.: Categorical Variable Segmentation Model for Software Development Effort Estimation, *Journal on IEEE Access*, Vol. 7, pp. 9618–9626

(2019)

[XXI] Hoc, H. T., Silhavy, R., Prokopova, Z. and Silhavy, P.: Comparing Stacking Ensemble and Deep Learning for Software Project Effort Estimation, *Journal on IEEE Access*, Vol. 11, pp. 60590–60604 (2023)

[XXII] Prokopova, Z., Silhavy, P. and Silhavy, R.: Influence Analysis of Selected Factors in the Function Point Work Effort Estimation, in *Proceedings of the Intelligent Systems in Cybernetics and Automation Control Theory*, pp. 112–124 (2019)

[XXIII] Ersoy, E., Bagriyanik, S. and Sozer, H.: On the Accuracy of Effort Estimations Based on COSMIC Functional Size Measurement: A Case Study, in *Proceedings of the Empirical Software Engineering and Measurement*, pp. 528–537 (2024)

[XXIV] Silhavy, P., Silhavy, R. and Prokopova, Z.: Outliners Detection Method for Software Effort Estimation Models, in *Proceedings of the Software Engineering Methods in Intelligent Algorithms*, pp. 444–455 (2019)

[XXV] Silhavy, P., Silhavy, R. and Prokopova, Z.: Spectral Clustering Effect in Software Development Effort Estimation, *Journal on Symmetry*, Vol. 13, No. 11, p. 2119 (2021)

[XXVI] Silhavy, P., Silhavy, R. and Prokopova, Z.: Stepwise Regression Clustering Method in Function Points Estimation, in *Proceedings of the Computational and Statistical Methods in Intelligent Systems*, pp. 333–340 (2019)

[XXVII] Van Hai, V., Nhung, H. L. T. K., Prokopova, Z., Silhavy, R. and Silhavy, P.: Toward Improving the Efficiency of Software Development Effort Estimation via Clustering Analysis, *Journal on IEEE Access*, Vol. 10, pp. 83249–83264 (2022)

[XXVIII] Lavazza, L., Liu, G. and Meli, R.: Using Extremely Simplified Functional Size Measures for Effort Estimation: An Empirical Study, in *Proceedings of the Empirical Software Engineering and Measurement*, pp. 1–9 (2020)

[XXIX] Effendi, A., Setiawan, R. and Rasjid, Z. E.: Adjustment Factor for Use Case Point Software Effort Estimation (Study Case: Student Desk Portal), *Journal on Procedia Computer Science*, Vol. 157, pp. 691–698 (2019)

[XXX] Rosa, W. and Jardine, S.: Data-Driven Agile Software Cost Estimation Models for DHS and DoD, *Journal on Systems and Software*, Vol. 203, p. 111739 (2023)

[XXXI] Mushtaq, Z. and Wahid, A.: Inclusion of Functional and Non-Functional Parameters for the Prediction of Overall Efforts of Mobile Applications, *Journal on Computer Standards &*

Interfaces, Vol. 71, p. 103404 (2020)

[XXXII] Kaur, A. and Kaur, K.: Investigation on Test Effort Estimation of Mobile Applications: Systematic Literature Review and Survey, *Journal on Information and Software Technology*, Vol. 110, pp. 56–77 (2019)

[XXXIII] Tawosi, V., Moussa, R. and Sarro, F.: On the Relationship between Story Points and Development Effort in Agile Open-Source Software, in *Proceedings of the International Symposium on Empirical Software Engineering and Measurement*, pp. 183–194 (2022)

[XXXIV] Alhamed, M. and Storer, T.: Playing Planning Poker in Crowds: Human Computation of Software Effort Estimates, in *Proceedings of the International Conference on Software Engineering*, pp. 1–12 (2021)

[XXXV] Tanabata, K., Hazeyama, A., Yamada, Y. and Furukawa, K.: Proposal of an Evaluation Method of Individual Contributions Using the Function Point in the Implementation Phase in Project-Based Learning of Software Development, *Journal on Procedia Computer Science*, Vol. 192, pp. 1524–1531 (2021)

[XXXVI] Bluemke, I. and Malanowska, A.: Software Testing Effort Estimation and Related Problems: A Systematic Literature Review, *Journal on ACM Computing Surveys*, Vol. 54, No. 3, pp. 1–38 (2022)

[XXXVII] Maia, A. W. and Farias, P. P. M.: Transactions as a Service, *Journal on Advances in Intelligent Systems and Computing*, Vol. 984, pp. 415–423 (2019)