PLS-SEM for Software Engineering Research: An Introduction and Survey

DANIEL RUSSO, Department of Computer Science, Aalborg University
KLAAS-JAN STOL, Lero—The Irish Software Research Centre *and* University College Cork

Software Engineering (SE) researchers are increasingly paying attention to organizational and human factors. Rather than focusing only on variables that can be directly measured, such as lines of code, SE research studies now also consider unobservable variables, such as organizational culture and trust. To measure such latent variables, SE scholars have adopted Partial Least Squares Structural Equation Modeling (PLS-SEM), which is one member of the larger SEM family of statistical analysis techniques. As the SE field is facing the introduction of new methods such as PLS-SEM, a key issue is that not much is known about how to evaluate such studies. To help SE researchers learn about PLS-SEM, we draw on the latest methodological literature on PLS-SEM to synthesize an introduction. Further, we conducted a survey of PLS-SEM studies in the SE literature and evaluated those based on recommended guidelines.

CCS Concepts: • **Software and its engineering** → *Software creation and management*; *Software organization and properties*; • **General and reference** → **Empirical studies**; **Evaluation**.

Additional Key Words and Phrases: Partial Least Squares, Structural Equation Modeling, Research Methodology, Critical review

ACM Reference Format:
Daniel Russo and Klaas-Jan Stol. 2021. PLS-SEM for Software Engineering Research: An Introduction and Survey. *ACM Comput. Surv.* 54, 4, Article 1 (December 2021), 37 pages. https://doi.org/10.1145/3447580

1 INTRODUCTION

> "The way the paper is presented makes it very difficult to follow for computer science people. The shift away from mainstream software engineering to psychological and sociological issues is a real problem. In the space of only a few pages, the authors introduce the use of: PLS, Cronbach, AVE, HTMT, VIF, and Stone-Geisser. I have never heard of these. I felt overwhelmed with jargon and metrics that to be honest seemed like an exercise in reader confusion and statistical overkill."
>
> *—a reviewer at a top software engineering conference*

The nature of software engineering (SE) research has evolved considerably over the past decades. The last decade has witnessed increased attention for human factors, leveraging frameworks and methods from the social sciences [37], including partial least squares structural equation modeling (PLS-SEM). While we believe the SE field is enriched by studying a wide array of topics and leveraging appropriate methods from other fields, the introduction and use of new methods into the SE field can also face resistance, as the opening quote illustrates. If the SE field adopts methods

Authors' addresses: Daniel Russo, daniel.russo@cs.aau.dk, Department of Computer Science, Aalborg University, Selma Lagerlöfs Vej 300, 9220, Aalborg, Denmark; Klaas-Jan Stol, Lero—The Irish Software Research Centre *and* University College Cork, School of Computer Science and Information Technology, Cork, Ireland, klaas-jan.stol@lero.ie.

0360-0300/2021/12-ART1
https://doi.org/10.1145/3447580

from other fields, then SE researchers need resources to equip themselves to understand these methods.

Many studies in SE now address topics such as "organizational culture" [61], "project success" [8], and "trust" [149]. These concepts are *latent* variables or non-observable variables that cannot be directly measured with a single metric. Instead, they are indirectly measured using a set of *indicators* or *manifest variables*. PLS-SEM is one way to analyze such non-observable variables and the relationships between them. Indeed, we observed that PLS-SEM studies had been published in key SE journals, including the Journal of Systems and Software and Information and Software Technology, as well as conferences, including the International Conference on Software Engineering.

Hwang et al. suggest many reasons to use PLS-SEM [94]. First, PLS-SEM is appropriate when the latent variables of interest are composites, rather than "common factors" [135, 162] (we briefly address the difference in Sec. 2). Second, PLS-SEM is also useful when the research goal is to predict key constructs [169], or to identify key constructs ("driver" constructs) [73]. Third, when the researcher has complex models that comprise many different constructs, indicators and relationships [32, 73, 75]. Fourth, when the research model is evaluated with secondary or archival data [65, 133]. Further, PLS-SEM provides accurate estimates with small sample sizes [75] and can therefore be used when the researcher has access to smaller sample sizes, which can be useful if population sizes are small (e.g., an investigation in a single organization). We note PLS-SEM is often incorrectly justified by claiming that it is appropriate when using small sample sizes [104, 132]. While PLS-SEM models can converge to a solution with smaller samples than what is typically required in a covariance-based structural equation model (CB-SEM), smaller samples will reduce the statistical power.

A traditional notion of PLS-SEM is that it is ideally suited for exploratory research, that is, to conduct analyses if the researcher has little insight as to how the various constructs might relate or if theory is lacking [74]. While this is a valid reason to use PLS-SEM, there is increasing agreement that researchers can also conduct confirmatory and explanatory research [21, 82].

While not widely used, we observe that PLS-SEM is gaining interest within the software engineering research community. One potential reason for the interest in PLS-SEM is the increasing attention for human factors in the field and the need to rely on psychometric measures to study such factors [107, 110]. A critical reflection on the use of PLS-SEM in software engineering research is now needed for several reasons.

First, the PLS-SEM approach to analysis has seen considerable advances in recent years. New techniques have been proposed and evaluated and sometimes replaced older evaluation techniques. Thus, this paper seeks to offer a snapshot of many of these recent methodological advances. To that end, we summarize the PLS-SEM method and synthesize guidance from other disciplines to provide a coherent and systematic set of guidelines for SE researchers. In so doing, rather than presenting extensive technical explanations of the various available tests, we seek to offer a high-level overview, a 'map,' with numerous pointers to the methodological literature on PLS-SEM. While such overviews exist in other fields, these are indeed not focused on software engineering research or are now outdated (e.g., [128]), given the advances made in PLS-SEM in recent years.

Second, from a methodological point of view, this review assesses the extent to which the method is used and reported correctly. We identified 29 articles that present 30 research models that were evaluated using PLS-SEM. Our review revealed that very few of the analyzed models follow a complete model assessment.

Third, by reviewing PLS-SEM studies within the SE literature, we focus on the substantive topics that have been studied using this method. We showcase and draw attention to this method, which is very suitable to study a range of topics within the field. A review of articles reporting PLS-SEM in a software engineering context can serve as a source of inspiration for other researchers to study

topics for which they felt an appropriate method was hitherto lacking. Notwithstanding, while highly promising, PLS-SEM is not a panacea or a "silver bullet" [114], and critiques of its misuse have been voiced and addressed [83, 118, 145, 147, 148].

The remainder of this article is structured as follows. Given that PLS-SEM is not widely known within the SE community, we include an overview in Section 2. Section 3 draws on literature from other disciplines to synthesize evaluation guidelines based on the current state of the art of PLS-SEM. In Section 4, we review the studies published within the software engineering field. We summarize the topics where PLS-SEM has been applied to software engineering in the past in Sec. 5. The results are further discussed in Sec. 6, which provides suggestions for future work.

2 OVERVIEW OF PLS-SEM

2.1 Introduction to Latent Variables

Ullman [185, p. 35] defined SEM as:

> "a collection of statistical techniques that allow a set of relations between one or more independent variables (IVs), either continuous or discrete, and one or more dependent variables (DVs), either continuous or discrete, to be examined."

Of particular interest is SEM's ability to define latent variables, which cannot be measured directly. Examples include constructs such as trust, organizational culture, and project success. To measure such a construct, a researcher identifies a series of observed variables that represent the construct. A set of variables that together represent a latent variable is referred to as a measurement instrument. For example, one measurement instrument for the construct 'trust' contains the following items [33], each of which is an observed variable—typically a question on a survey instrument:

- People in my organization will always keep the agreements they make with one another.
- People in my organization behave in a consistent manner.
- People in my organization are truthful in dealing with one another.

Instruments such are these are common in the social sciences; human respondents are asked to rate each of the statements on a Likert-scale, typically with five or seven points. Different instruments may have been developed for the same construct, and which one to use is a matter of how well the instrument fits the specific research context. Values of latent (unobserved) variables can thus be "inferred" through combinations of several observed variables. PLS-SEM is one approach to calculating models with latent variables, though this has been a topic of some controversy and discussion due to the way constructs are represented in a mathematical sense. We return to this point in more detail in Sec. 2.3.

A recent mapping study observed that LVA is still a niche approach in empirical software engineering research [40]. In that review, De Oliveira Neto et al. found that among LVA, the most common techniques are factor analysis with 150 papers, principal component analysis (133) and, structural equation modeling with 47 studies. These results suggest that SEM has not yet gained widespread adoption within the software engineering community.

2.2 Elements of a PLS-SEM Model

A PLS-SEM research model consists of a structural model (or inner model) and a measurement model (the outer model). Figure 1 shows an abstract example. The structural model is the most relevant model for theory development and evaluation of hypotheses. In Figure 1 constructs are represented by ovals, and hypotheses are represented by single-headed arrows.

The measurement model must be measured by observed variables called items, indicators, or manifest variables, which are represented as rectangles in Fig. 1. Each item is conceptually unidimensional and, in survey research among human respondents, often measured using a Likert

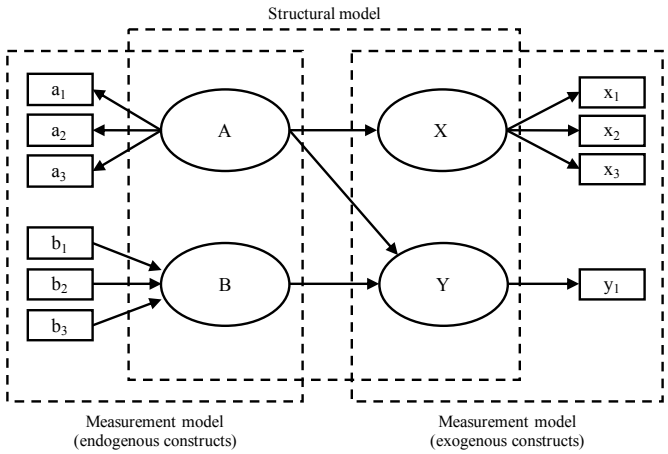

Fig. 1. Abstract PLS-SEM model example. Constructs A, X, and Y are reflectively measured; B is formatively measured.

scale. However, other items could measure age or tenure—these are typical examples of control variables. The measurement model refers to the relationship between the items and their non-observable variables; that is, how is the non-observable latent variable operationalized through a set of observable manifest variables. Non-observable variables can be measured in different ways. While our description below draws on traditional terminology that distinguishes 'reflective' from 'formative' constructs, the mathematical representations is different from those in CB-SEM, suggesting these terms are inappropriate, which has been a source of considerable debate [145, 147, 148]. We return to this issue in Sec. 2.3.

First, constructs can be said to be 'reflective.' In this representation, any changes in the unobservable reflective construct latent variable are "reflected" in the values of its indicators; that is, a change in the construct "causes" a change in its items. The standard notation for this are arrows from the construct to the indicators (see Fig. 1). In reflective measurement, items can be considered a representative sample of all the possible items available within the construct's conceptual domain; thus, they are interchangeable [72]. Dropping an item from the measurement model does not change the meaning of the construct. For this reason, having a "weak" item that does not correlate with the other items may not represent a good indicator to measure the construct.

Second, constructs can be said to be 'formative.' Formative measurements are used when items *form* or define a construct. In this case, the phenomenon of interest is formed by underlying measures [97]. To represent formative constructs in a graphical model, the arrows originate in the items and point to the construct. In contrast to reflective constructs, indicators of formative constructs are not interchangeable; every indicator captures a specific aspect of the latent variable. Thus, indicators do not have to be strongly correlated since it would suggest that they are interchangeable and not unique.[1] The consequence is that the construct's conceptual coverage should be as broad as possible to capture the meaning of the construct [43]. In other words, if a relevant formative item is missing, it may bias the interpretation of its latent construct since one piece of understanding is lacking.[2] As an intuitive example, consider the construct 'drunkenness.' Assuming one would not have a

[1]Section 3.6 clarifies how to address this issue, checking for collinearity.

[2]For this reason, we illustrate in Section 3.6 the way to test items' significance and relevance. Similarly, we have to check for content and convergent validity to assess the measurement error. Also, scholars should be aware that the number of items

breathalyzer (which would give a direct measurement of alcohol in a person's blood), one could "measure" drunkenness in terms of the number of beer pints. However, without considering other indicators, such as the number of glasses of wine and shots of hard liquor, one could incorrectly conclude that a person who drank zero pints of beer is not drunk. In a software engineering context, consider the example of software quality. If software quality would be measured through the number of reported defects only, then one would miss other indicators, such as maintainability, usability, or any other non-functional attribute. Which indicators to include will depend on the specific context of the study.

A clear-cut definition of when a measurement model should be reflective or formative is not possible. Guidelines for using reflective-formative type models have been proposed by Becker et al. [10]. The research community has debated this issue extensively [16, 17, 70], but consensus has as of yet not been reached about how to specify models correctly. This will depend on the nature of constructs, which are not intrinsically formative or reflective, but rather depend on their conceptualization. That is, whether a construct should be modeled reflectively or formatively, depends on the way a researcher defines the nature of the construct or on the literature they draw on. Outlines to define the appropriate measurement specification have been proposed by Hair et al. [72, p. 47]. This allows scholars to test the appropriateness of the reflective or formative specification [75].

2.3 PLS as a Member of the SEM Family

As mentioned, SEM represents a family of methods. In this article, we focus specifically on partial least squares SEM (PLS-SEM), which some scholars have suggested to be a leading SEM technique [74, 141], or even a "silver bullet" [74]. It is a causal-predictive approach to SEM, in the sense that it focuses on the prediction of statistical models, whose hypotheses suggest causal explanations [75, 132, 162]. PLS-SEM is often considered as an alternative to covariance-based SEM (CB-SEM). Similarly to PLS-SEM, CB-SEM has also been used in SE research [131, 150], though it has not seen widespread adoption.

However, PLS-SEM and CB-SEM rely on different statistical representations and computations, and assumptions [135]. Rigdon et al. present a thoughtful and accessible comparison of PLS-SEM and CB-SEM and provide a set of recommendations for researchers to consider when designing their studies [135].

Some scholars have critiqued PLS, some even claiming that it does not constitute structural equation modeling [65, 145, 147, 148]. A key issue in this disagreement relates to the representation of constructs as latent variables, which can be mathematically modeled as *common factors* (as is done in CB-SEM), or as *composites* or *emergent variables* [14], as is the case in PLS-SEM [76, 135].

In PLS-SEM, reflective and formative measurement is often referred to as 'Mode A' and 'Mode B,' respectively. However, equating Mode A with reflective (and Mode B with formative) measurement is incorrect, despite the widespread reference to this terminology in much of the methodological literature [1, 145]. Rönkkö et al. traced the origins of this association, suggesting the link between Mode B and formative measurement is accidental [146, p. 659]. The terms Mode A and Mode B, then, refer to the different algorithms used to generate the 'latent' variable scores [1, 145].[3] Linking Mode A to reflective measurement (and Mode B to formative measurement) has become a convention, rather than a reflection of reality [14, 136, 137, 146]. Dijkstra pointed out that PLS-SEM constructs "proxies" for the latent variables [45, p.32]. The latent variables' values are calculated as

also have a significant impact on measurement uncertainty, which threatens the validity of the research [134], meaning that they should be used parsimoniously.

[3] A third mode, Mode C, is a combination of Mode A and B; see Dijkstra for further details [45].

composites, which are linear combinations of their indicators, which is different from the common factor representation of latent variables in CB-SEM. A key difference is that composites do not completely incorporate measurement error and are, therefore, only approximations of common factors that are used in CB-SEM [104, p. 675].

Benitez et al. reserve the term 'latent variable' for behavioral concepts (such as personality traits) and use the term 'emergent variable' to refer to a construct that represents an artifact [14]. The assumption that the indicators that make up a composite are free of measurement error aligns well with the artifactual nature; that is, a construct representing a human-made object, as opposed to a behavioral construct that *"exists in nature"* [14]. Benitez et al.'s example of an emergent construct illustrates the human-made nature [14, p. 4]:

> "Based on theory, bread is made from wheat, water, salt, and yeast. Although the correlations between the amounts of wheat, water, salt, and yeast in a sample of loaves of bread are likely to be high, one would not conclude that bread is something that should be measured, i.e., that bread causes (or is caused by) wheat, water, salt, and yeast. Rather, wheat, water, salt, and yeast are the simple entities (ingredients) combined to form the emergent variable representing the artifact we call bread."

While further technical details of this discussion are beyond the scope of this introductory paper, these are discussed in detail in other sources and we recommend aspiring PLS-SEM users to become cognizant of these issues [14, 76, 83, 90, 104, 135]. In this paper we follow Henseler et al.'s rebuttal [83] of Rönkkö and Evermann's critiques [145] in accepting that PLS is, in fact, a member of the SEM family. Similarly, we echo Rigdon's assertion that both common factors (as calculated in CB-SEM) and composities (as calculated in PLS-SEM) are approximations of the constructs under study [136].

2.4 Recent Developments of PLS-SEM

For many years, PLS-SEM has been considered as a "lesser" version of CB-SEM. Rigdon provides an account of the reasons for this [136]. Among these reasons is a long-held belief that factor-based representations of theoretical concepts were somehow more accurate than the "approximations" offered by PLS-SEM [136, p. 344]. While the details underpinning the mathematical representation of factor-based and composite-based modeling is beyond the scope of this article, Rigdon has argued to treat both as alternative proxies that seek to represent concepts of interest [136, p. 347].

To address the shortcomings of the way PLS calculates latent variables, Dijkstra developed consistent PLS (shortened to PLSc) [44–46, 48]. Wold (who lay the foundations of PLS-SEM) himself noted that path coefficients and indicator loadings were not consistent, but *consistent at large* [192]. This may lead to Type I errors (false positive) since an effect may be considered significant, although it may not be observed in the population [102, 130, 184]. Also, for this reason, a confidence interval of 95% and a power level of 80% is usually recommended [67, 115]. PLSc was developed to correct the estimates of reflectively measured constructs through a reliability coefficient [45, 46, 48]. Practically, it consistently estimates the indicators loadings, inner construct correlations, as also path coefficients. A Monte Carlo simulation study showed that the bias of PLSc is comparable to CB-SEM for medium to large sample sizes [44]. As a result, it minimizes Type I and Type II errors.

In recent studies, Henseler and Schubert [90], and Schuberth et al. [165] discuss confirmatory composite analysis (CCA), which is a set of procedures for specifying and assessing composite models, as an alternative to confirmatory factor analysis (CFA) that is generally used in a CB-SEM setting. In other words, it is a hybrid version of CB-SEM applied to composite models, where typical CB-SEM steps are followed (i.e., model specification, model identification, model estimation, and model assessment), where the estimated model is assessed both globally and locally. Although

CCA is an independent analysis approach, some scholars tried to embed it as an evaluation step of PLS-SEM [71]. This attempt has been criticized as there is no evidence of its efficacy [164].

Another common problem is the limitation in the prediction accuracy caused by sampling constraints, mainly depending on the participants' selection [24, 154]. Sampling shortcomings might lead to biased results, undermining generalizability claims. Therefore, Becker and Ismail developed the WPLS-SEM (wPLS) algorithm to *ex post* adjust the sampling weights in the model estimation [9]. For example, it is possible to weigh demographic characteristics. So, if gender or education level distributions of a target population are known *ex ante*, and the sampling procedure could not determine an exact representative population, this can be adjusted with wPLS.

3 DEVELOPMENT AND EVALUATION OF PLS-SEM MODELS

A PLS-SEM model needs to be carefully evaluated to provide valid and reliable outcomes. We propose a systematic approach to pursue and review PLS-SEM studies based on the field's most recent advances. Fig. 2 provides an overview of the main steps involved.

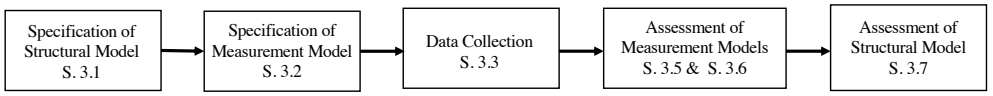

Fig. 2. Flowchart of the most relevant steps of a PLS-SEM analysis and review

3.1 Specification of the Structural Model

As a first step, a researcher identifies the constructs that are relevant to the phenomenon of interest. Informed and justified by prior theory, constructs can be proposed to be related; that is, a researcher draws on prior empirical observations or theories that seem appropriately fitting to identify a set of hypotheses that link a number of constructs together. The entire set of hypotheses represents the structural model. In addition to these 'normal' hypotheses, which propose a relationship between an exogenous variable *A* and an endogenous variable *B*, other types of relationships between variables can be proposed: mediating and moderating relationships. Figure 3 presents conceptual models that include a mediator (left) and a moderator (right). Mediation and moderation analyses can be evaluated within the overall PLS-SEM framework as well [157].

A *mediator* is a third variable *M* which is positioned in between an exogenous variable *A* and an endogenous variable *B*. The relation between *A* and *B* is hypothesized to be mediated by *M*. (To be clear, such a mediating hypothesis implies two normal hypotheses, linking *A* and *M*, and *M* and *B*.) In order to assess whether a mediating relationship exists, a direct relation between *A* and *B* must be modeled as well.

Researchers can also propose a variable to have an influence on a proposed relationship between two variables *A* and *B*. This is called a *moderator*: depending on the value of the moderator *W*, the relationship between *A* and *B* is either weakened or strengthened. Moderation analysis is an advanced topic that is beyond the scope of this introductory article. Moderation analysis is not widely used but can offer useful and actionable insights when moderating variables are within managers' control. For example, if perceived organizational support (POS)—that is, the level of support that a developer experiences in their work—is shown to moderate a relationship that affects developer productivity, then managers would be wise to explore ways to increase developers' POS. Hair et al. have discussed moderation analysis in more detail [72, 77].

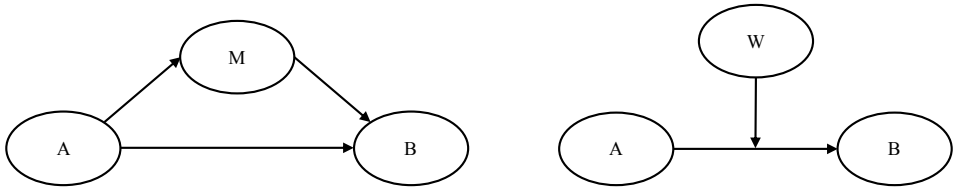

Fig. 3. Left: M mediates the relationship between A and B. Right: W moderates the relationship between A and B.

3.2 Specification of the Measurement Model

In step two, the researcher should operationalize each of the constructs of interest; that is, identify suitable measurement instruments for each construct with the help of an auxiliary theory [53], or, if no suitable instrument is available, develop one themselves. This is an essential step because it is concerned with deciding how a non-observable variable is measured. This is known as content validity, which Straub et al. defined as *"the degree to which items in an instrument reflect the content universe to which the instrument will be generalized"* [178, p. 424]. This step involves a degree of subjectivity, and there are no clear rules to guarantee content validity [172]. The researcher should carefully consider measurement instruments and consider whether they are appropriate.

Chin [28] points out that using a high number of items does not imply a better understanding (i.e., estimates) of the construct, but it reduces standard errors. Moreover, from an operational perspective, using a number of items allows the iterative process of model evaluation and optimization [34] by discarding those which are not adequate to measure the target construct (i.e., low loading of an item onto a construct). We discuss this issue in more detail in Sec. 3.5.

Single-item constructs may be used (see construct Y in Fig. 1 with a single item y_1). Because each item is a question on a survey, reducing the overall number of items by using single-item constructs could therefore lead to a better response rate [64]. However, Sarstedt et al. [155] argue it is best not to use single-item constructs. Diamantopoulos et al. [41] provide a thoughtful discussion and detailed guidelines on the use of single-item constructs.

3.3 Data Collection and Examination

In step three, the researcher collects and examines the data. Frequently, scholars collect primary data through a structured questionnaire, but secondary (or archival) data can also be used. In either case, issues such as missing data, unlikely response patterns (i.e., inconsistent answers), outliers, and data distribution should be considered. Many issues may arise due to a poor survey design, and it is therefore recommended to pay careful attention to this. Dillman et al. present useful guidelines in this respect [49]. Different strategies can be used to enhance construct validity, such as a pre-test to receive feedback on the survey and a pilot-test to see how well the items perform [186].

After specifying a model, data collection can start. The data sample must be representative of the target population that it purports to represent, which means (a) that respondents must be knowledgeable about the subject matter and (b) that a sufficient number of respondents participate. While it might be straightforward to define what 'knowledgeable' means — a survey may include competence screening, and papers tend to report on demographics including education level, job position, and gender — determining an appropriate sample size is far from trivial. Goodhue [67] pointed out that small sample sizes have been incorrectly associated with PLS-SEM use. Although PLS-SEM performs quite well with small sample sizes [74], this interpretation has sometimes been overstretched as a justification for having a small sample, including in software engineering

research [5]. This may lead to significant problems because underpowered studies (i.e., with too small sample sizes) are not well able to capture the difference between the chosen population parameter (such as mean equals 0) and the null hypothesis value (i.e., effects), leading to a false negative, or, a Type II error [6].

Several guidelines to determine the sample size have been mentioned in the PLS literature [2, 29, 106]. The most trivial one is the so-called "ten times rule" introduced by Chin and Newsted [29], suggesting that the minimum sample is represented by the most significant number of structural paths directed at one construct in the structural model, multiplied by ten, which can be very small (for example, 20 observations). Probably also due to a misappropriation of this "rule" [141], ten years later, Marcoulides, Chin, and Saunders criticized that suggestion, stating that this "rule" was not supposed to be a guideline and noted that "for a more accurate assessment, one needs to specify the effect size[4] for each regression analysis and look up the power tables" [115, 327]. Power tables for regression analysis, developed by Cohen [36] are a reasonable trade-off solution to determine sample sizes [72] (Section 4.2 discusses their use further). However, Aguirre-Urreta and Rönkkö [2] pointed out this approach also has some limitations and is not well suited for PLS-SEM. The most advanced simulations for power analysis[5] are Monte Carlo simulations [193]. However, this solution is not without issues. Performing a Monte Carlo simulation is not a trivial task and may not be feasible for many scholars. Further, its use in PLS-SEM literature is still debated [2].

Regarding the examination of data, researchers should first consider whether outliers are random errors and consequently be deleted or whether they represent a distinct and unique subgroup of the sample. In this case, several approaches have been proposed to uncover unobserved heterogeneity, such as Finite Mixture Partial Least Squares (FIMIX-PLS), which can be used to identify so-called latent (unobserved) respondent subgroups [161].

Despite the fact that PLS-SEM does not make any assumptions of data distribution normality, scholars have nevertheless recommended conducting a data distribution analysis because an excessive non-normality may harm parameters' significance assessment [72]. A distribution with a mean larger than 1 or lower than −1, can be considered as substantially skewed [72]. Such cases might suggest inaccurate data and might bias the model. Therefore, cases of extreme kurtosis and skewness should be identified and analyzed and, if present, eliminated, for example, by conducting data transformations. Such decisions should not be taken lightly but carefully considered on a case by case basis. For example, some constructs, such as perceived quality, tend to be heavily skewed. Thus, authors should discuss whenever the skewness of the measured construct depends on the construct's nature or because of a flawed measurement process.

3.4 PLS-SEM Model Evaluation

Once the model is developed, it can be evaluated using specialized software packages or libraries. Several libraries are available for PLS-SEM analysis in the open source package R [190], such as `sempls` [122], `pls` [120] and `plspm` [151]. Alternatively, several commercial packages are available, such as ADANCO [84], SmartPLS [142], PLS-Graph [31], XLSTAT [55], and WarpPLS [105].

A standard practice in evaluating CB-SEM models is the evaluation of the overall model fit. While the earlier literature on PLS-SEM suggested that a 'global' model fit was not appropriate [92, p. 202], some confusion exists, and this remains to be a topic of some discussion [14, 72, 86]. Some model fit measures have been proposed for PLS-SEM models. Among the earliest was the Goodness-of-Fit (GoF) [182]; however, studies have shown that the GoF is not suitable to evaluate fit as it cannot distinguish misspecified models from well-specified ones [89]. Several scholars have

[4]The effect size measures if the measured result is real and how large it is [51].
[5]It is an analysis to define the probability of finding a statistically significant effect.

argued that due to PLS-SEM's predictive nature, it is inappropriate to consider model fit at all since it does not focus on the confirmation of the research model to the same degree as CB-SEM [56, 168]. Nevertheless, several other measures have been suggested, but these have not been widely adopted and are still considered to be in an early phase of development [73, 77]. The Standardized Root Mean Square Residual (SRMR) is a common fit measure for CB-SEM [91]. Henseler et al. suggest that the SRMR is also appropriate to detect misspecification of PLS-SEM models [83]. Others include the squared Euclidean distance (d_{ULS}) and the geodesic distance (d_G) [47], RMS_{theta} [113], and the Normed Fit Index (NFI) [15]. While some software packages report these fit measures, researchers should consider their appropriate use carefully. For example, some fit measures are only appropriate when using the consistent PLS-SEM (PLSc) estimator discussed earlier. Benitez et al. pointed out that more research is needed to establish sound thresholds for these fit measures [14].

3.5 Assessing PLS-SEM results of the Reflective Measurement Model

Depending on whether a model includes reflective or formative measures, different criteria apply. This section discusses the evaluation of reflective measurement models. Sec. 3.6 discusses criteria for formatively measurement models.

Reflective measurement models are by far the most common. They are tested for their (i) internal consistency reliability, (ii) convergent validity, and (iii) discriminant validity. Table 1 summarizes the relevant assessment criteria, along with their desirable values.

3.5.1 Internal consistency reliability. Internal consistency reliability is measured through different measures. Among those, Cronbach's alpha [38] is one of the most commonly used, providing an estimate of the reliability based on the intercorrelations of the observed item's variables. Another common test is composite reliability [191], which is similar to Cronbach's alpha, indicating the level of reliability from 0 to 1, but is more scale-independent and less conservative. Recently, Dijkstra and Henseler introduced the consistent reliability coefficient to measure internal consistency (rho_A), addressing shortcomings[6] of Cronbach's alpha [44]. For all criteria, values between 0.60 and 0.70 are acceptable in exploratory research; that is, a field that is in a nascent state ideal. Other than that, values between 0.70–0.90 reflect satisfactory to good results [75, 123]. Values above 0.95 may suggest that items are measuring the same phenomenon, decreasing construct validity typically; this suggests that items are semantically redundant [50, 75]. This could suggest a potential common method bias [178].[7] Hair et al. suggest using bootstrap confidence intervals to assess whether the lower bound of a 95% confidence interval exceeds the minimum threshold of 0.70 [75].

3.5.2 Convergent validity. Convergent validity is the degree to which a measure correlates positively with alternative measures of the same construct. We use reflective specification when items are interchangeable since they (theoretically) all represent the construct equally (as opposed to formative constructs, when dropping an indicator may alter the conceptual meaning of that construct). Convergent validity is measured by the Average variance Extracted (AVE), which is the grand mean value of the squared loadings of the items associated with the construct. The AVE's desired values are above 0.50 since it suggests that the construct represents more than 50% of the variance of its items [60]. It is commonly agreed that items should share at least 50% (or 0.5) of their variance. The indicator reliability is then measured as the square root of this shared variance, which is 0.708—hence, the loading of an item onto its construct (the outer loading) should be at least 0.708; this is commonly simplified to a threshold of 0.70 [28, 72]. However, newly defined

[6] In case of small sample sizes, if some specific conditions are not met, Cronbach's alpha will underestimate the internal consistency reliability [171].

[7] Common method bias is a well-known phenomenon of applied statistics which happens when the response variation is derived from the measurement instrument and not by informants' knowledge [20, 129].

Table 1. Assessment of reflective measurement models

Criteria	Description	Test	Desirable values	Reference
Internal consistency	Type of reliability to evaluate result's consistency across items. The aim is to discover if the correlation between items are high enough, suggesting similarities between the items of the same latent variable.	Cronbach's alpha	0.70 (satisfactory) to 0.90 (good); 0.60-0.70 acceptable for exploratory research	[38]
		Composite reliability	0.70 (satisfactory) to 0.90 (good); 0.60-0.70 acceptable for exploratory research	[191]
		rho_A	0.70 (satisfactory) to 0.90 (good); 0.60-0.70 acceptable for exploratory research	[44]
		Confidence intervals	Minimum threshold above lower bound of the 95% bootstrap confidence interval	[75]
Convergent validity	Indication of which items correlate positively with different items of the same latent variable.	AVE	>0.50	[60]
		Indicator reliability	>0.70	[28]
Discriminant validity	Indicates if a latent variable is measuring a distinct construct and the degree of which items exemplify the target construct.	HTMT	<0.90 if constructs are conceptually similar, otherwise <0.85	[86]
		Bootstrapping	Bootstrapped confidence interval of HTMT should not contain 1.0	[72]

constructs, scales, or research models may have loadings lower than 0.70 [92]. Items with a loading of 0.40 – 0.70 should be considered for removal [72]. If dropping the item that loads poorly increases the AVE significantly (or from an unacceptable level to an acceptable level, i.e., >0.50), it should be discarded [72].

3.5.3 Discriminant validity. Discriminant validity represents the degree of uniqueness of one construct in relation to another. Discriminant validity is critical to ensure that different constructs capture different concepts; this can be assessed with the Heterotrait-Monotrait ratio of correlations (HTMT) [86]. This ratio is based on the comparison of the Heterotrait-Heteromethod correlations and the Monotrait-Heteromethod correlations, which identifies the lack of discriminant validity at a high sensitivity rate [72]. The cut-off value is 0.90 if the constructs are conceptually similar (such as different developers skills, e.g., decomposition and abstraction); a more conservative cut-off value is 0.85 [86]. It is also suggested to conduct an inferential test through a bootstrapping procedure to ensure that the HTMT ratios are statistically significantly different from 1.0 [63, 72]. This is because cut-off values are very conservative and have a high likelihood of falsely rejecting discriminant validity (i.e., Type II error) [63].

For the sake of completeness, we list two other outdated methods. Both methods are still frequently reported but have been shown to be insufficient to assess discriminant validity [86]. The first one is the test of crossloadings of items onto other constructs. If items "crossload" then their loading is higher for another construct than that it purports to measure, which is problematic [28]. Items should load high on the construct that they purport to measure but low on all other constructs. A second outdated test is the Fornell-Larcker criterion [60] which requires that all correlations between constructs be less than the lowest of the square root values of the AVEs.

3.6 Assessing PLS-SEM results of the Formative Measurement Model

Formative constructs differ from reflective constructs in that changes in the observed indicators of a formative construct are thought to 'cause' changes in that construct. Therefore, the mathematical modeling of formative constructs is different, thus requiring other assessment techniques (see Table 2). In contrast to items of a reflective construct, items are not interchangeable, as each one is supposed to be relevant to explain the construct. Referring back to the construct drunkenness: if one drop (or forgets) the indicator 'glasses of wine,' only counting pints of beer and shots of hard liquor, one could mistakenly assess a person as sober, despite having drunk a bottle of wine.

Tests for formative constructs include content validity, convergent validity, collinearity, and indicator weights' significant reliability. Testing internal consistency reliability is conceptually incorrect since measurement error is not captured in the mathematical representation of formative constructs [42].

3.6.1 Content validity. Content validity is important as stressed in Section 3.3. Particularly when dealing with formative measurement models, the highest care should be devoted to the content specification in which the scholar specifies the domain's content of the measured items since they are not replaceable and interchangeable. Consider once again the measurement of drunkenness; leaving out either pints of beer or glasses of wine will capture a different concept; nor can one interchange pints of beer with glasses of wine. From an operative point of view, a review of items by experts and an extensive literature review contributes to increasing content validity[8] [43].

3.6.2 Convergent validity. Hair et al. define convergent validity as *"the extent to which a measure correlates positively with other measures (indicators) of the same construct"* [72]. A redundancy analysis should be pursued in these cases, originally described by Chin [28]. This procedure involves measuring a construct twice (thus, redundantly): once as an (exogenous) formatively measured latent variable that predicts an (endogenous) reflectively measured latent variable representing the same construct [72]. The strength of the path between the two latent variables is an indicator of convergent validity. A value above 0.70 is recommended (in our example of Fig. 1 this is represented between the path B and Y) [25, 72]. This would contribute to a high R^2 value of the endogenous constructs.

3.6.3 Collinearity. High correlations suggest interchangeability of the items that make up a construct. While this is desirable for reflectively measured constructs, as indicators are treated as interchangeable, it is undesirable for formative constructs where each item represents a unique aspect of a construct. High correlations imply collinearity [72], which reduces the ability to estimate weights and may lead to biased estimation result [77]. A common criterion to assess collinearity is the Variance Inflation Factor (VIF) [121]. A VIF value larger than 5 suggests a collinearity problem [74]. This would mean that the item's variance is mostly explained by the other formative items of the same construct, suggesting semantic interchangeability. Some guidelines suggest a cut-off value for VIF of 10 [186], while others advise using a more conservative cut-off value of 3 [124]. There

[8]In this regard, practical tools such as www.inn.theorizeit.org developed by Larsen & Bong [109] might be of help.

Table 2. Assessment of the Formative Measurement Model

	Meaning	Test	Desirable values	Reference
Convergent validity	Extent to which an item relates to other items of the same construct	Redundancy analysis	Path between formatively measured LV and reflectively measured LV > 0.70	[25, 28, 72]
Collinearity	Degree of correlation between items of a construct	Variance Inflation Factor (VIF)	<5.00 (soft cut-off value); <3.00 (hard cut-off value)	[13]
Significance and relevance	Testing of the statistical significance and semantically relevance of formative items' outer weights	Bootstrapping, outer loadings	Retain items with loadings > 0.50 even if nonsignificant. If loading <0.50: if nonsignificant, delete indicator, if significant, consider deleting.	[72]

are also other assessment techniques, which are not readily supported by the available software packages but which researchers might want to consider when evaluating borderline cases. One of these is the condition index (CI) which tests critical collinearity levels in formative measurement models, and tolerance (i.e., the amount of variance of one formative item that is not represented by the other items) [68]. In cases with high collinearity, the researcher must carefully reflect on the item's theoretical contribution to the construct before dismissing it. For example, when using validated measurement instruments (from other studies), it might be a good idea to retain them since they explain a specific latent variable's unique facet.

3.6.4 Significance and Relevance of Formative Items. The final step is the assessment of the significance and relevance of the formative items. This means to test the statistical significance and relevance of formative items' outer weights to establish each item's relative contribution to the construct. To test significance, we again rely on a bootstrapping procedure [39, 54]. Especially in the case of a large number of formative items, most of them may be non-significant [72]. However, this does not mean that non-significant items must be discarded per se; instead, the theoretical contribution to its construct should first be considered. A loading above 0.50 is desirable [72]. An item with a loading below this threshold and not statistically significant may be removed if it lacks strong theoretical support [72]. However, considering that the elimination of formative specified items has almost no effect on the parameter estimates when re-estimating the measurement model, formative items should not be removed just for statistical reasons but for theoretical reasons concerning the research model.

3.7 Assessing PLS-SEM results of the Structural Model

Once the measurement (or outer) model has been assessed, the next step is to assess the structural or inner model. This involves assessing the predictive power and the significance of the relationships between the constructs of the structural model. Based on this evaluation, the researcher can accept or reject the proposed hypotheses. Table 3 summarizes the various tests to conduct.

3.7.1 Collinearity. The evaluation of the structural model involves evaluating a series of regression equations that represent the relationships between constructs [75]. The estimated regression results that represent the relationships between constructs could potentially be biased if constructs exhibit too much collinearity—that is, two (or more) constructs may, in essence, represent similar concepts. Variance Inflation Factor (VIF) values are used to assess collinearity. Hair et al. suggest that VIF

Table 3. Assessment of the Structural Model

Term	Meaning	Test	Desirable values	Reference
Collinearity	Degree of correlation between constructs	VIF	<5.00 (hard cut-off value), <3.00 (soft cut-off value)	[75, 121]
Path significance	Assesses how strongly two constructs relate to each other	Path coefficient	Depending on f^2 and p	[77]
		p-value	Standard cut-off values are 0.05 and 0.01; sometimes a cut-off of 0.10 is used to highlight weak support	[58]
Path relevance	Extent to which two constructs relate to each other	IPMA	No threshold values	[138]
Coefficient of determination	Measure of the predictive accuracy of the structural model	R^2	Context and discipline dependent	[75]
Effect size	Measure of the impact of the exogenous construct on the endogenous one	f^2	>0.02	[35]
Predictive relevance	Measure of a model's predictive power	Q^2	>0.00	[66, 176]
Relative measure of predictive relevance	Measure of the predictive relevance of paths relations to the reflectively measured endogenous constructs	q^2	>0.02	[35]
Heterogeneity	Unobserved heterogeneity: groups are not known up front	FIMIX-PLS, PLS-POS, PLS-TPM, REBUS-PLS, PLS-GAS, PATHMOX	Discover subpopulation groups	[11, 62, 77, 137, 140, 163]
	Observed heterogeneity: groups are known up front	Multi-Group Analysis (MGA)	Group segmentation	[158]
Prediction-oriented results assessment	Measure of a model's out-of-sample predictive power	PLSpredict (Q^2_{predict})	>0.00	[169, 170]
Robustness checks	Complementary methods for assessing the robustness of PLS-SEM results	CTA-PLS Unobserved heterogeneity	Test dependent	[70, 160]

values should be close to 3 or lower [75]. If collinearity remains an issue, then the researcher could consider using a second-order construct [75], which is a construct that has as its "indicators" (or items) two or more other constructs. Further discussion of higher-order constructs is beyond the scope of this article and refer interested readers to Sarstedt et al.'s discussion on the topic [156].

3.7.2 Significance of the structural model relationships. Path coefficients represent the hypothesized relationships among the constructs whose standardized values range between −1 and 1 (for strongly negative or positive relationships, respectively), whereas values close to 0 suggest a weak relationship. The significance of path coefficients depend on their p-values and the effect sizes, both of which should be reported and discussed [77]. Because PLS-SEM does not assume the data are normally distributed and thus does not rely on normal theory, standard parametric tests cannot be relied upon. In order to assess whether a relationship is statistically significant, a bootstrap procedure can be used. The bootstrap procedure generates a sample distribution that approximates the normal distribution, which can then be used to establish critical t-values and subsequently the p-values. Drawing on the clinical statistical literature, we recommend also discussing the clinical or practical significance [108], which is one of the three questions which concerns behavioral scientists (along with statistical significance and effect size) [101]. Practical significance can be considered as the magnitude of the observed effect and whether it is big enough to modify research conclusions. For example, if the difference in lines of code (LOC) in two groups of programs is statistically significant, but the difference in absolute terms is only 10 LOC, does this represent a practically significant difference given the variability inherent in any program? Does this result lead to a better understanding of, say, defect prediction in software? Probably not. In other words, a statistically significant relationship may not be practically significant. There is no consensus about the measurement of practical significance, so judgments as to the practical significance relies on experts' considerations [108].

3.7.3 Relevance of the structural model relationships. The next step is to consider the relevance of the relationships. Some path coefficients may be significant, but have a very small effect size. Therefore, taking into account the relative importance of path coefficients is essential to draw appropriate conclusions. Former techniques to assess include total effects analysis, which sums a construct's direct and indirect effects' importance of the relationship between exogenous and endogenous constructs [59]. Building on prior literature [116, 173], scholars have recently introduced the importance-performance map analysis (IPMA) which provides a deeper understanding of the relevance, but also of the performance of an exogenous construct in explaining the endogenous (target) construct [138]. This allows discovering which constructs might improve a particular target construct, also making a group comparison (e.g., based on gender, age, experience). For example, through an IPMA we could discover that older people struggle to deal with relevant technology (such as healthcare) much more than younger people [181]. As a result of this insight, we might conclude that special attention is paid to user interfaces or that special training is provided. Another example: research on software quality in open source projects may find that failing tests locally before submitting a pull request is more important than writing a good pull request title. If this were the case, project leaders and maintainers should mainly focus on improving test requirements by committers rather than asking them to write good pull request descriptions. In other words, IPMA is a criterion that helps a researcher to provide appropriate recommendations. However, IPMA makes some assumptions. For example, items should be measured only through metric or quasi-metric scales [159]. Moreover, items should have the same scale direction, i.e., they should address every construct either positively or negatively. Once computed, constructs appear on a Cartesian plane, where the x-axis represents the relative relevance (or importance) of each construct, and the y-axis their performance.

3.7.4 Model's explanatory power. The coefficient of determination explained variance, or R^2 value is an essential measure in PLS-SEM since it measures the model's explanatory power. It measures the proportion of variance explained by each endogenous construct. For example, a construct A with an R^2 of 0.30 means that 30% of the variation in A is explained by the constructs that point to A. The resulting value lies between 0–1, suggesting the level of explanatory power. It is not possible to provide threshold values because they depend on the subject matter and the model's complexity. For instance, a R^2 value of 0.20 may be considered as substantial in some disciplines and very weak in others [74]. Since the value of R^2 is related to the size of the model (i.e., complex models have, on average, higher R^2 values than smaller models), it is good practice to also consider the Adjusted R^2 criterion, which adjusts the R^2 value based on the model size [96].[9] In software engineering, the community has not reached a consensus about thresholds for R^2 values.

To assess whether a specific exogenous latent variable has a substantial impact on the endogenous ones, we may use the f^2 effect size [35]. It measures if the exogenous construct has a substantial impact on the endogenous one. Cohen [35] proposed the following indicative thresholds: a value below 0.02 represents no effect, 0.02–0.15 represents a small effect size, 0.15-0.35 a medium-sized effect, and above 0.35 a large effect size.

3.7.5 Predictive relevance. To assess the predictive relevance of a given endogenous latent variable, researchers can calculate Stone-Geisser's Q^2; values should be greater than 0 [66, 176]. Predictive accuracy can be small (higher than 0), medium (larger than 0.25), or large (greater than 0.50) [75]. We speak of predictive relevance, since Q^2 predicts the data points of the items of reflective measurement endogenous models (it does not apply to formative endogenous constructs) [72]. To compute it, we use blindfolding to obtain cross-validated redundancy measures for each endogenous latent variable [30, 88, 183]. Since blindfolding works as an iterative process, which repeats until each data point has been omitted, to re-estimate the model, the data point elimination depends on the omission distance (D), which the researcher chooses for the model computation. Ideally, it should be between 5 and 10 [72]. In order not to delete the same set of observations in each round from the data matrix, the number of observations for the model estimation divided by the omission distance should not be not an integer.

Several methodological researchers recommend against the use of Q^2 however [166, 170]. Shmueli et al. [169, 170] proposed an alternative approach to assess a model's predictive relevance (see Sec. 3.7.6). As a relative measure of predictive relevance, we can use q^2 effect size. It measures the predictive relevance of paths relations to the reflectively measured endogenous constructs. Since its computation is similar to the f^2, its threshold values are the same: 0.02, 0.15, and 0.35 suggest small, medium, and large predictive relevance for a certain endogenous latent construct [35].

3.7.6 Prediction-oriented results assessment. The PLSpredict algorithm developed by Shmueli et al. [169] provides an assessment regarding a model's predictive power by mimicking the evaluation of the predictive power for out-of-sample data[10] [169, 170]. Using this procedure, the data set is divided into k equal-sized subsets; the model is run ('trained') with k-1 subsets. The model's predictive power can then be evaluated to assess whether the model can predict the k^{th} subset, which Shmueli et al. refer to as the 'holdout' [169]. Further details on this procedure are offered by Shmueli et al. [169, 170] and Hair et al. [75]. PLSpredict aims to assess whether the model outperforms the most naïve linear regression benchmark (referred as Q^2_{predict}). Therefore, a Q^2_{predict} value higher than 0

[9]For the sake of completeness, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are also useful alternatives to the Adjusted R^2 [79]; however, they might not be readily supported in all software packages.
[10]Previously discussed techniques (e.g., R^2 or Q^2) compute their estimates on the entire sample. For that reason, this predictive power does not apply for out-of-sample data.

indicate that the PLS path model's prediction error is smaller than the prediction error given by the (most) naïve benchmark.

To assess the predictive power of a model, several statistics can be calculated, such as the mean absolute error (MAE), mean absolute percentage error (MAPE), and the root mean squared error (RMSE) [166, 170]. All of them are based on the differences between the predicted values and the actual (observed) values, but these statistics differ in some aspects, making some of them more useful in certain circumstances than others; thus, we advise researchers using PLS-SEM to become cognizant of these details that have been discussed by several methodological researchers [166, 169, 170].

3.7.7 Observed heterogeneity. Theoretical models propose relationships between constructs. To test these models, researchers collect a sample from a population, which is analyzed using PLS-SEM software. The assumption is that the population is homogenous, but this is not always true. For example, novice developers and senior developers are likely to differ in their training needs. Open source software developers may vary in their motivation to continue contributing [7]. Treating a sample as if it were homogenous can lead to distorted results [137, 140, 163]. Researchers can include a priori groupings and control variables in their study [11]. This is called observed heterogeneity: the variation across subgroups is captured in variables such as age, gender, and tenure, often as control variables. PLS-SEM Multi Group Analysis (MGA) [158] can assess whether there are significantly different behaviors between groups. Before running a MGA, some analyses need to be performed. First, testing through a multi-step procedure for measurement invariance of composite models (MICOM) allows assessing whether or not the MGA is a meaningful test [87]. Permutation testing is recommended to verify whether pre-defined data groups (for example, based on gender) have significant differences in their path coefficients and other parameter estimates [32].

3.7.8 Robustness checks. Finally, scholars should also consider conducting robustness checks [153, 160]. The first is the CTA-PLS test and considers the correct choice of a formative versus reflective measurement model [70]. In a reflective measurement model, the CTA-PLS should be 0. However, if the result is significantly different from zero, a formative measurement model is preferred instead. Ultimately, however, we emphasize that theoretical grounding should prevail on the results of such empirical test [77]. Another robustness test is to assess unobserved heterogeneity. This is the case when subgroups exist in the sample that are not known to the researcher [11]. In such a case, a subset of the sample exhibits specific shared characteristics without a researcher being aware of this. When data subgroups differ considerably, the model estimates will be biased. If it is possible to identify these subpopulations, then this unobserved heterogeneity can become *observed* heterogeneity: that is, researchers can identify and label these subpopulations for use in future studies [11]. Several methods have been proposed to achieve this [11, 152, 158], such as FIMIX-PLS, PLS-POS, PLS-TPM, REBUS-PLS, PLS-GAS, and PATHMOX. Each of these approaches has benefits, and limitations [11, 152]. A detailed discussion of these approaches is beyond the scope of this introductory article; Sarstedt's offers a detailed review, and discussion [152, 162]. Scholars should identify possible explanatory variables that characterize a population's uncovered segments (e.g., gender). Also, additional analyses, such as moderation [12], or multigroup analysis [158] can provide further insights into the studied phenomenon. Sarstedt et al. also suggest to test for nonlinear effects and endogeneity [153, 160]. The assessment of nonlinear effects can be tested with the approach suggested by Svensson et al. to verify whether nonlinear effects are significant and conclude that the linear effects model is robust [180]. Endogeneity can be assessed using Hult et al.'s procedure [93].

4 REVIEW OF PLS-SEM RESEARCH IN SOFTWARE ENGINEERING

We now present an analysis of the use and reporting of PLS-SEM in the software engineering community. This review seeks to identify how well the method has been used and reported, as well as draw attention to areas for improvement, which may be of use to other SE scholars who are considering PLS-SEM. We also analyze what questions SE researchers have answered using PLS-SEM. This may help to understand aspiring PLS-SEM users how this technique can answer research questions in a SE context.

4.1 Search Protocol

To conduct the systematic literature review, we followed established guidelines by Kitchenham and Charters for the software engineering domain [103]. We followed the following steps:

(1) Eligibility criteria: Studies using PLS-SEM.
(2) Information sources: Identification of the most relevant software engineering venues according to Google Scholar.
(3) Search: Within each repository of the target venues, we made the following query: "partial least squares" OR "PLS." We did not set any limitations to the year of publication.
(4) Study selection: We found a total of 36 papers, which we screened according to the following exclusion criteria:
   (a) Papers that did not use PLS-SEM as a method to evaluate a theoretical model were excluded.
   (b) No inclusion criteria were defined as we only surveyed the target venues.
(5) Data collection process: Each paper was coded adapting an established PLS-SEM coding standard [78] to the most recent assessment criteria.

To identify relevant publication venues, we consulted Google Scholar, which lists the h5-index for conferences and journals. We excluded specialized venues such as Patterns of Programming Languages (PoPL) and included those with a very high reputation (e.g., ACM Transactions on Software Engineering and Methodology). We excluded magazines (e.g., IEEE Software) and non-peer-reviewed articles (e.g., those published on public sites such as arXiv.org). All data were extracted in February 2019. Our goal was not to conduct an exhaustive literature review but rather to present a snapshot of the use of PLS-SEM in general-purpose software engineering venues.

The final set of 36 papers included in our SLR is presented in Table 4; these papers included the strings "partial least squares" OR "PLS." We checked whenever the papers referred to PLS path modeling and not PLS regression, which is a different technique. Of these, 29 used PLS-SEM. The other seven papers refer to the method within their literature review. One paper [4] presented a comparative analysis of two models. Since the authors ran two separate models with similar but different characteristics, we reviewed them as two distinct models. We also excluded one paper [125] since it presents a mixed PLS-SEM and Confirmatory Factor Analysis (CFA) computation which followed a mixed evaluation procedure; thus, it would have biased our evaluation scheme.

Figure 4 lists the identified papers, ordered by their year of publication year. We observe that PLS-SEM's adoption is a relatively recent development by the software engineering community, with the first paper in our sample published in 2005. The number of articles published over the years appears to be reasonably stable.

4.2 Data Characteristics

Our first consideration concerns data. While PLS-SEM does not assume a normal distribution of the data and thus uses non-parametric statistics, reporting kurtosis and skewed data is still useful because they might be an indicator of weak data measurements [40, 72]. Unfortunately, while some

Table 4. Number of papers and models identified in the systematic literature review

Venue	Name	No. Found Papers	No. Relevant Papers	No. Models
Conferences	ICSE	3	1	1
	FSE	0	0	0
	ASE	0	0	0
	MSR	0	0	0
	ICSME	0	0	0
	ICST	0	0	0
Journals	TSE	0	0	0
	TOSEM	0	0	0
	JSS	15	14	14
	IST	16	13	14
	EMSE	1	1	1
	SCP	0	0	0
	SoSyM	1	0	0
Total		36	29	30

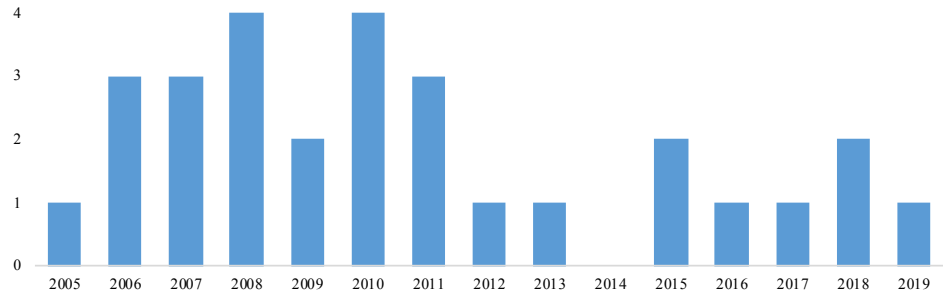

Fig. 4. Number of papers using PLS-SEM published per year

authors provide tables with descriptive statistics about their data set, none of the reviewed papers mentions whether or not the data were distributed normally.

Similarly, almost none of the papers made an a priori power analysis of the sample size; only two papers included a power analysis [8, 19]. This is not uncommon in software engineering research. A review of 103 controlled experiments published in 12 major SE venues between 1993–2002 concluded that the statistical power of such experiments was substantially inadequate [52].

We recommend performing a power analysis to identify the minimum sample size before data collection. One tailored approach for PLS-SEM is proposed by Aguirre-Urreta and Rönkkö [2], using a Monte Carlo simulations used for power analysis supported by the R package matrixpls [144]. However, other tools do not require the use of R, designed for regression analysis which provides an acceptable approximation, such as G*Power [57], which is informed by Cohen's work on power analysis [35]. To conduct a power analysis to establish an appropriate sample size, the suitable test family is the F-test with multiple linear regression, using an a priori test to compute the required sample size. The researcher is asked to provide the following:

- Effect size f^2

- Error probability α
- Power (1-α)
- Number of predictors (i.e., arrows pointing at a latent variable in the structural model)

Section 3 discusses how we defined the recommended values for the planned study. The "ten-times rule" [72], which suggests that for each predictor there should be at least ten observations, is usually quite inaccurate [141]; more sophisticated techniques should be used to assess the statistical power and effect sizes a priori [36]. As briefly pointed out in Section 3.3, there is a growing debate about appropriate methods to determine the minimum sample size due to its pivotal importance for a correct model computation [106]. As a general recommendation, we suggest authors to perform a power analysis to gain an a priori understanding of the minimum sample size needed to detect effects of a specified size. The G*Power tool offers a user-friendly interface for scholars not familiar with the R statistical package to conduct such analyses. However, tools such as G*Power provide only the technically minimum sample size. Data should always be representative of the target population in the first place.

4.3 Model Characteristics

Table 5 (which we adapted from Hair et al. [78]) reports the outcome of the model characteristics evaluation. Our main observation is that 80% of authors did not specify the measurement specification. In other words, it is unclear if the constructs are measured in a formative or reflective fashion. Just this shortcoming alone impedes a thorough review process. On average, the structural models in the reviewed set of studies in software engineering have six constructs and seven path relationships, numbers comparable to other disciplines (e.g., [78]).

The mean number of items for the entire model is 25, with as few as 4 and as many as 80 items. Other communities use more items to specify their constructs, such as in Hospitality Management, where PLS-SEM models have, on average, 35 indicators [139]. The reason may be that the software engineering community has mostly drawn on other fields to import their measurement instruments without developing their own ones. Since not all instruments can be adopted from other disciplines, software engineering specific ones are somewhat limited. On a positive note, the use of single-item constructs is not common (13%), supporting good predictive validity, meaning that more items are more likely to effectively predict the target construct with respect to a single-item [41]. Model complexity tends to be relatively high. Higher-order constructs allow modeling variables on a more abstract dimension by adding an additional layer of abstraction. Comprehensive guidelines for using higher-order constructs in PLS-SEM have been advanced by Sarstedt et al. [156]. Eighty percent of the studies use a higher-order structure that contains further layers of abstraction to define a construct. Seventy percent of the models use interaction effects, and 67% include mediators. Finally, all reviewed papers presented a graphical representation of their models.

4.4 Measurement Model Evaluation

The lack of model specification in 80% of the cases can make the evaluation of the measurement model quite challenging. In other words, authors rarely state whether they have used formative or reflective (or both) constructs. Thus, we sought both formative and reflective assessment criteria in all surveyed papers, although not all criteria might apply for every model. Table 6, adapted from Hair et al. [78], summarizes the results. We particularly stress that authors should specify whether the model has reflective or formative specified constructs to facilitate a thorough evaluation. As shown in Table 5, only one paper uses formative measurement models, whereas 4 articles use reflective ones; only one paper specifies both formative and reflective models.

Table 5. Model characteristics

Criterion		Results (n = 30)	Proportion (%)
Number of non-observable variables	Mean	6	
	Median	6	-
	Range	(2; 9)	
Number of structural model path relations	Mean	7	
	Median	6	-
	Range	(1; 16)	
Mode of measurement models	Only reflective	4	13%
	Only formative	1	3%
	Reflective & formative	1	3%
	Not specified	24	80%
Number of items per reflective construct (min)	Mean	4	
	Median	3	-
	Range	(2; 13)	
Number of items per reflective construct (MAX)	Mean	6	
	Median	5	-
	Range	(3; 13)	
Number of items per formative construct (min)	Mean	4	
	Median	4	-
	Range	(1; 13)	
Number of items per formative construct (MAX)	Mean	5	
	Median	4	-
	Range	(1; 14)	
Total number of items in the model	Mean	25	
	Median	23	-
	Range	(4; 80)	
Number of models with single-item constructs		4	13%
Mediating effects		20	67%
Interaction effects		21	70%
Higher-order model		24	80%
Model graphical representation		30	100%

For reflective measurement models, most researchers report the item loading (82%). All reviewed articles considered the internal consistency reliability with at least one suitable criterion, and the majority of models used both composite reliability and Cronbach's alpha. However, no one reported rho$_A$. Also, convergent validity has always been adequately treated with the Average Variance Extracted (AVE) or other measures. Twenty-four out of 30 models tested discriminant validity, mainly with the outdated Fornell-Larcker criterion. Only the last two studies considered the new Heterotrait-Monotrait (HTMT) criterion [86]; this is not surprising as it is a relatively new technique.

When we consider formative measurement models, almost all of the studies display the value of their items (weights), and half of them show the level of significance of such items to the model. Collinearity is not discussed in most cases; only three studies report Variance Inflation Factor (VIF) values, and no one used the condition index criterion. Redundancy analysis to assess convergent validity is not considered, nor are any other techniques to that end. Although some papers mentioned "convergent validity analysis," they refer to that one concerning reflective constructs (since they only report the AVE values), and not any proper techniques such as redundancy analysis.

4.5 Structural Model Evaluation

An overview of how structural models are evaluated in our sample of reviewed studies is presented in Table 7 (based on Hair et al. [78]). Nearly all studies report the explained variance (87%), path coefficients (93%), and the significance of path coefficients (93%). However, none of the studies tested their relevance with the Q^2, q^2, or Q^2_{predict}. This harms the studies' predictive quality and general applicability since little is known about how good data points of the items of reflective measurement endogenous models are predicted. One study reported the confidence interval. However, to gain more information about a coefficient estimate's stability, bootstrapping confidence intervals may be a suitable strategy [77].

Considerations about the magnitude of the exogenous constructs' impact on the endogenous ones (measured through effect sizes) are included only in 23% of the models. Total effects are just marginally reported (7%), although they might provide useful insights into a construct's antecedents.

Finally, only the two most recent papers [7, 8], made some considerations concerning heterogeneity. Barcomb et al. [7] used REBUS-PLS to identify four different segments within the sample of free/open source software developers. Each subgroup had a different 'profile' of motivations to continue participating. Similarly, Batra [8] uncovered heterogeneity through PLS-POS. Neglecting heterogeneity means ignoring the existence of possible subgroups in the surveyed population. Samples may contain subgroups that, if not identified and considered, may skew final results. As such possible biases caused by subgroups segmentation are hardly considered, neither regarding observed heterogeneity with Multi Group Analysis nor with unobserved heterogeneity through procedures such as FIMIX-PLS or REBUS-PLS.

4.6 Reporting

Many of the analyzed studies exhibit shortcomings in reporting, which impedes a comprehensive evaluation of the computed models. One issue on which many papers fell short is an indication of whether constructs are reflective or formative. We point out that recent methodological advances have now reconsidered this traditional view of reflective vs. formative measurement, and in that light, this issue is becoming less critical [85]. Model computation should be clearly stated, such as weighting scheme, stop criterion, sampling weights, bootstrapping, and blindfolding specifications, which are poorly considered in the studies included in our analysis. On a positive note, all studies

report the software used to run the PLS-SEM analyses. More advanced assessment criteria, such as heterogeneity, would also be essential to exclude possible sampling biases.

There is considerable room for improvement in reporting PLS-SEM studies within the software engineering community. Insufficient or unclear reporting has a direct impact on results evaluation, leading to inaccurate theory contribution. This may bias future research efforts, orienting the community's focus on misleading assumptions.

Table 6. Evaluation of measurement models

	Criterion	#	%
Reflective measurement models			
Items reliability	Items loading	23	82%
Internal consistency reliability	Only composite reliability	3	10%
	Only Cronbach's alpha	10	33%
	Both	17	57%
	Confidence interval	1	3%
	rho_A	0	0%
Convergent validity	AVE	26	87%
	Other	8	27%
Discriminant validity	Only Fornell-Larcker criterion	17	57%
	Only cross-loadings	0	0%
	Cross-loadings & Fornell-Larcker	5	17%
	HTMT	2	7%
Formative measurement models			
Items absolute contribution	Items weights	21	70%
Weights significance	p-value, t-value, significance level, others	14	47%
Collinearity	Only VIF	3	10%
	Only condition index	0	0%
	Both	0	0%
Convergent validity	Redundancy analysis	0	0%
	Other	0	0%

5 ANALYSIS OF STUDIES USING PLS-SEM IN SOFTWARE ENGINEERING RESEARCH

This section discusses the topics that have been studied using PLS-SEM as a research tool. The 29 papers in our review address a wide range of research goals, especially in the software engineering professional practice area. Table 8 presents a summary of our analysis.

The first consideration belongs to the adequateness of the selected sample size. An insufficient sample size reduces the power of the PLS-SEM computation, increasing the margin of error, thus biasing the model, and thus the outcome of the paper which may lead to incorrect theoretical models. Assessing *a priori* the minimal sample size by means of a power test, as explained in Sec. 4.2 is the best way to reduce Type II errors. We also highlight the minimum sample size considering the number of structural model path relations in each model. According to the sample description of each study, we assessed whether the size was adequate or not. To do so, we considered a default power of 0.8, medium effect sizes of 0.15, with an error probability of 0.05. Since most of the

Table 7. Evaluation of structural model

Criterion		#	%
Endogenous constructs' explained variance	R^2	26	87%
Effect size	f^2	7	23%
Predictive relevance	Q^2	0	0%
Relative predicted relevance	q^2	0	0%
Path coefficients	Absolute values	28	93%
Significance of path coefficients	t-values, p-values, others	28	93%
Confidence intervals		2	7%
Total effects		2	7%
Out-of-sample predictive power	Q^2_{predict}	0	0%
Heterogeneity	Multi Group Analysis	0	0%
	PLS-POS	1	3%
	REBUS-PLS	1	3%

analyzed papers did not specify the power of their analysis, we used these standard values for every study.

Most of the studies work with adequate sample size, i.e., 18 (62%). Only 11 (38%) do not use for their model computation a sufficient sample size. Moreover, four (14%) articles claimed to use a sample size, which is half or less than the minimum size required. In particular, these four studies, which have been highlighted in Table 8, display severe biases in their models and should probably not be considered as reliable for future research.

Previous scholars used PLS-SEM especially to investigate human-related aspects of software engineering. For a classification purpose, we refer to the IEEE Software Engineering Body of Knowledge (SWEBOK) Version 3.0 taxonomy [18], namely:

- Software Requirements
- Software Design
- Software Construction
- Software Testing
- Software Maintenance
- Software Configuration Management
- Software Engineering Management
- Software Engineering Process
- Software Engineering Models and Methods
- Software Quality
- Software Engineering Professional Practice
- Software Engineering Economics
- Computing Foundations
- Mathematical Foundations
- Engineering Foundations

As shown in Table 5, the majority of the papers (41%) deal with **software engineering professional practice** aspects. In particular, papers focus on the group dynamics and psychology aspects of developers, as their interaction with stakeholders. For example, they identify organizational trust factors in enterprise open source vendors to assess the impact of system trust on adopters' attitudes and intentions [149] or study the relationship between a team's external social capital and team flexibility [23].

Software engineering process aspects represent 21% of our sample. Such articles investigate the management aspects of the software life cycle and process improvement. In this regard, previous scholars examined the impact of software process standardization on software flexibility and, the final project performance [112]; or the effect of software process maturity in the selection of critical information systems implementation strategies, as its impact on software quality and project performance [179].

Also, several aspects regarding the development approach, within the **software engineering models and methods** (14%) have been investigated. Particular attention has been reserved for development methods, especially Agile ones. For instance, one research tried to understand the impact of tailoring criteria on the adoption or selection of agile practices [19]. Similarly, another explored the factors driving the use of agile practices among the adopters of such development method [187].

Similarly, **software engineering management** aspects got the attention, with 14% of the papers. Here, researchers were interested in issues related to scope definition, project planning, or project enactment, such as the identification of the impact of organizational factors on the overall performance of a software product line [5]. Or the exploration of the relationship among management control, system-user interaction quality, and project performance [188], to take another example.

Finally, a few articles (10%) addressed **software engineering economics** topics. Relevant aspects covered by scholars regarded life cycle economics, risk and uncertainty, as also economics fundamentals. Studies in this area regarded the evaluation of the ways co-production relationships between software developers and users improve the outcomes of a development project [167]. As well as the advancement of a measurement model to quantify the level of process harmonization in an organization [143].

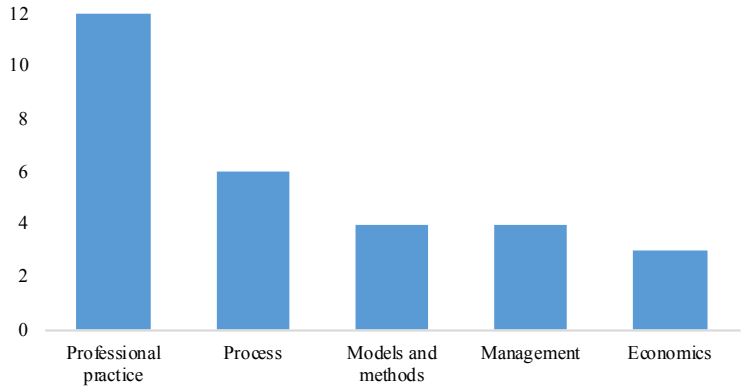

Fig. 5. Software engineering areas covered by PLS-SEM articles, according to the SWEBOK V.3 taxonomy

A detailed analysis for each paper is presented in Table 8, where interested readers can also find additional details, such as the published venue and the publication year.

Table 8. Analysis of PLS-SEM studies in software engineering

Study	Min. sample	Adequate size?	Sample description	SWEBOK Area	Research Goal
Green et al. (2005) [69]	92	No	63 SEI-Personal Software Process trained software developers	Process	Examination of factors that motivate developers to adopt and sustain the use of software process improvement innovations

Continued on next page

Table 8 continued from previous page

Study	Min. sample	Adequate size?	Sample description	SWEBOK Area	Research Goal
Wang et al. (2006) [188]	77	Yes	196 IT Professionals	Management	Exploration of the relationship among management control, system-user interaction quality, and project performance
Wang et al. (2006) [189]	77	Yes	128 Project leaders in charge of the ERP implementation	Professional practice	Advance a success model of ERP based on the social capital theory that serves to explain the importance of the group, rather than the individual for the successful implementation of enterprise systems
Meso et al. (2006) [119]	114	No - Severe bias	57 Students	Models and methods	Identification whether the use of strong-typed methodologies provides for superior facilitation of knowledge management processes, and whether this leads to the production of higher quality information systems solutions
Ahmed & Capretz (2007) [3]	114	No - Severe bias	44 IT Managers and Developers	Process	Analysis of the association among various key business factors and to study the relationships between them and the business performance of software product lines
Ahmed et al. (2007) [5]	98	No - Severe bias	40 (generic) Employees	Management	Identify the impact of organizational factors on the overall performance of a software product line
Subramanian et al. (2007) [179]	77	Yes	212 IT Professionals	Process	Examination of the effect of software process maturity in the selection of critical information systems implementation strategies, as its impact on software quality and project performance
Ahmed et al. (2008) [4]	98	No - Severe bias	33 (generic) Employees	Process	Definition of the impact of architecture process activities on the overall performance of software product line
Jung & Goldenson (2008) [99]	55	Yes	948 CMM assessments by SEI-authorized lead assessors	Process	Identification of the dimensions underlying a set of Capability Maturity Model for Software key process areas and estimation of the internal consistency of each dimension of the capability maturity concept
Liu et al. (2008) [112]	77	Yes	221 IT Professionals	Process	Examination of the impact of software process standardization on software flexibility and, the final project performance
Sohn & Mok (2008) [174]	98	No	77 Developers	Professional practice	Identify what affects Open Source Software utilization

Continued on next page

Table 8 continued from previous page

Study	Min. sample	Adequate size?	Sample description	SWEBOK Area	Research Goal
Chen et al. (2009) [27]	103	Yes	323 IT Professionals	Professional practice	Examination of the relationships among the management interventions of pre-project partnering, coordination structure, the perception gap, and IT development project performance
Jiang et al. (2009) [98]	127	Yes	151 IT Professionals	Professional practice	Development and validation of a research model that shows how perception gaps explain residual performance risks in a project
Chang et al. (2010) [22]	92	Yes	128 IT Professionals	Management	Examination of the impact of user commitment on system development processes as well as final project outcomes
Keil & Park (2010) [100]	77	Yes	661 Students	Professional practice	Provide empirical support for the partial mediation of the whistleblowing model
Li et al. (2010) [111]	98	Yes	119 Developers	Professional practice	Provide a theoretical explanation of how team flexibility explains the relationship between Agile tools and methods in practice and a software development team's flexibility
Shim et al. (2010) [167]	85	Yes	128 IT Professionals	Economics	Evaluation of how the co-production relationship between software developers and users improves the outcomes of a development project
Chang et al. (2011) [23]	109	Yes	118 Developers	Professional practice	Examination of the relationship between a team's external social capital and team flexibility
Chen et al. (2011) [26]	123	Yes	151 IT Project Managers	Management	Examination of how user influence and user responsibility affect information systems project performance
Ifinedo (2011) [95]	139	No	109 Managers of IT companies	Professional practice	Discover the effects of external expertise and in-house or internal computer/IT knowledge on the success of ERP packages
Vijayasarathy & Turk (2012) [187]	103	No	98 Developers	Models and methods	Identification of the factors driving the use of agile practices among the adopters of such development methodology
Parolia et al. (2013) [127]	131	No	92 IT Managers	Professional practice	Examine whenever the presence of self and social competences critical program competences can be developed with effective participation

Continued on next page

Table 8 continued from previous page

Study	Min. sample	Adequate size?	Sample description	SWEBOK Area	Research Goal
Parolia et al. (2015) [126]	109	No	92 Managers of IT companies	Professional practice	Analysis of the effectiveness of conflict resolution on the implementation efficiency and fulfillment of business objectives
Romero et al. (2015) [143]	109	No	100 BPM Professionals	Economics	Development and validation of a measurement model to quantify the level of process harmonization in an organization
Mayeh et al. (2016) [117]	143	Yes	184 ERP Users	Economics	Investigation of the factors affecting the intention to use ERP systems
Roumani et al. (2017) [149]	92	Yes	192 IT Managers	Professional practice	Identification of organizational trust factors in enterprise open source vendors to assess the impact of system trust on adopters' attitudes and intentions
Batra (2018) [8]	114	Yes	124 IT Managers and Developers	Models and methods	Investigation of which factor, agile values or plan-driven aspects, contributes more toward the success of data warehousing, business intelligence, and analytics; and what are the significant antecedents of agile values and plan-driven aspects
Campanelli et al. (2018) [19]	98	Yes	308 Developers	Models and methods	Understanding of the impact of tailoring criteria on the adoption or selection of agile practices
Barcomb et al. (2019) [7]	92	Yes	101 Developers	Professional practice	Development and assessment of a theoretical model for retention of episodic volunteers in FLOSS communities through the moderating effects of age, gender, tenure, and contribution type

6 DISCUSSION AND CONCLUSION

PLS-SEM is a leading statistical technique for conducting structural equation modeling. As the software engineering research community is increasingly paying attention to human factors, many of which are not directly observable but rather are latent variables, the use of SEM is now highly relevant. Many constructs found in the social and behavioral sciences are now also becoming relevant in an SE context [177]. Thus, it makes sense to also adopt the methods from the social sciences to analyze these constructs. For example, many software organizations are running software process improvement initiatives, and the ability to measure concepts such as 'perceived organization support' and 'trust' is becoming essential in such studies. This endeavor has, for scholars such as Connolly [37], also a wider implication for the entire computer science community, advocating the idea to:

> "integrate the analytic lenses supplied by social science theories and methodologies [...] to move out of the methodologically singular natural/engineering sciences and moving tentatively into the methodological pluralism of the social sciences."

PLS-SEM is one fruitful approach that facilitates this. Unfortunately, while the SE community is not new to this method, as evidenced by the 29 articles published in leading SE venues, this paper's opening quote suggests that there is no widespread familiarity with this method. Thus, this paper seeks to help address this by providing a high-level introduction and overview of best practices.

6.1 Recommendations

Our review shows that software engineering scholars rarely went through a thoughtful analysis process. Partly because the PLS-SEM literature made several significant advances in the last few years, researchers were not aware of new practices and techniques. Thus, we propose a step-wise review protocol to standardize both writing and reviewing, based on the most recent recommendations:

(1) **Theory and hypotheses formulation**. In the first stage of analysis, it is important to consider the literature on the investigated topic to justify and ground research hypotheses in prior literature. Also, a research model induced through a theory-driven empirical investigation (such as a field study) could be used for this purpose. Based on such insights, the structural model can be developed.
(2) **Measurement specification**. After Step 1, scholars should determine how to measure the constructs of interest. PLS-SEM offers different measurement models (i.e., Mode A or Mode B); the development of the consistent PLS (PLSc) estimator addresses some of the shortcomings that have been observed in the traditional PLS algorithms. Now, items can be defined according to previous studies, or new ones can be developed. This phase is crucial since poorly specified and ill-developed items are likely to lead to an inappropriate model.
(3) **Data collection and characteristics**. Before collecting data, a power analysis should be carried out to understand the minimum sample size, after which data can be gathered, ensuring a good data sampling strategy. The researcher should also have a clear idea about the target population (e.g., demographics) to use the wPLS algorithm. Skewness, kurtosis, outliers, and observed heterogeneity should be reported. Missing data and inconsistent data should be treated properly not to bias the subsequent model computation.
(4) **Measurement model evaluation**. *Reflective* model evaluation should test the internal consistency reliability, convergent validity, and discriminant validity with the proposed techniques. Similarly, *formative* model evaluation should include an assessment of the content validity, convergent validity, collinearity, significance, and relevance of formative items.
(5) **Structural model evaluation**. This step includes testing the collinearity, significance and relevance of the relationships, R^2 (and the Adjusted – R^2), f^2, Q^2, Q^2_{predict}, and q^2.
(6) **Advanced analysis**. Heterogeneity tests, such as MGA or FIMIX, may better explain (observed or unobserved) subgroup behaviors, as also Robustness checks. Generally, any analysis which may be helpful to understand better the research results should be included.
(7) **Reporting**. All elements that may be helpful to enable reviewers to validate the findings should be reported. All steps of this analysis have to be explained in detail to grasp the research outcome's substance and their generalizability, along with model computation details. Where possible, raw data, code, and computation tables should be stored on a permanent repository (e.g., Zenodo) to support replication studies.

6.2 Future Research Opportunities

Digital transformation processes and software's pervasiveness will lead to emerging research challenges that cannot be predicted. Accordingly, future research opportunities for the software engineering field are particularly widespread. As the SE field is expanding its scope and is increasingly

focused on human and organizational factors—many of which are behavioral concepts [80]—the use of PLS-SEM is a promising approach to consider. Besides, design concepts or artifacts [81] are also central to the software engineering discipline. Indeed, research within the SE domain can broadly be categorized as being knowledge-seeking or solution-seeking [175]. PLS-SEM can serve as a fruitful and accessible tool to SE researchers to build the necessary knowledge for developing grounded and useful solutions.

It is essential in going forward that researchers become mindful of the differences between PLS-SEM and CB-SEM. While it is beyond this article's scope to present a detailed comparison, these approaches rely on different assumptions and computation models; Rigdon delivers a comprehensive overview [135].

This article is positioned as a starting point for researchers interested in studying latent variables in software engineering research. From this point of departure, we envisage the following to be useful directions for further work.

- Many advances have been made in the PLS-SEM field in recent years. While many of these advances fall beyond the current article's introductory scope, we believe it will be useful to explore how these advances are relevant to software engineering studies and how they can be fruitfully leveraged and applied, for example, through showcase studies.
- A key goal of this article was to present a retrospective of PLS-SEM studies in software engineering. We believe it is now time to look forward and establish clear and accessible guidelines leveraging the latest methodological advances and guidelines to conduct new studies within the SE discipline. Given the increased level of attention for human and organizational factors, diversity, and understanding human behavior in the field, we believe that PLS-SEM will be a useful tool in the SE researcher's toolbox.
- Having said that, we also caution that researchers do not blindly adopt PLS-SEM as a "silver bullet"; indeed, alternative methods may be more appropriate, including the use of CB-SEM. Establishing when to use PLS-SEM and CB-SEM within the specific context of software engineering research is undoubtedly a welcome exercise. We believe more in-depth knowledge and more variety in methods within the SE domain will enrich the field and contribute to making SE research more relevant to industry – an issue highlighted in recent years.

Generally speaking, every time empirical software engineering research deals with opinions, perceptions, or behaviors that are not directly measurable but are represented as latent variables, PLS-SEM may be a useful method. In sum, scholars whose scope is to provide a preliminary understanding of an emergent phenomenon will find in PLS-SEM a valuable research tool.

6.3 Conclusion

PLS-SEM is a suitable analysis method to estimate and assess complex theoretical models with relatively few data restrictions [65]. Software engineering researchers have concrete opportunities to explore new and emergent software phenomena with this method. PLS-SEM is more widely used in other fields, including the related field of Information Systems, and consequently, those fields exhibit a high level of maturity of reporting [141]. We observe a limited but stable number of studies in the SE field that use PLS-SEM. Given its potential in the software engineering field, we evaluate these studies to reflect on those studies and provide support and suggestions for future studies in the SE field.

This paper is a first review of the use of PLS-SEM in software engineering research literature. It offers a brief introduction to PLS-SEM. While detailed descriptions of many advanced techniques fall beyond this paper's scope, interested readers are invited to follow up on the recommended literature. We propose a step-wise review protocol to standardize both the writing and reviewing

of PLS-SEM articles. Another focus of this article is a review of the use of PLS-SEM in software engineering studies. We synthesize evaluation guidelines from other disciplines and use those to review and assess 30 PLS-SEM models published in SE publication venues. One of our findings is that there is considerable room to improve computation, evaluation, and reporting standards. Many of the analyzed papers present some shortcomings from a methodological and reporting perspective. We mapped the topics that researchers have addressed with these techniques to the Software Engineering Body of Knowledge (SWEBOK) and found that the papers covered the areas of professional software practice, process, models and methods, management, and economics.

ACKNOWLEDGMENTS

We are grateful to the extensive and constructive feedback of the anonymous reviewers, whose comments have led to a better article. This work was supported by Science Foundation Ireland grant 13/RC/2094 P2 and 15/SIRG/3293.

REFERENCES

[1] M.I. Aguirre-Urreta and G.M. Marakas. 2008. The use of PLS when analyzing formative constructs: Theoretical analysis and results from simulations. In *International Conference on Information Systems*.
[2] M. Aguirre-Urreta and M. Rönkkö. 2015. Sample Size Determination and Statistical Power Analysis in PLS Using R: An Annotated Tutorial. *Communications of the Association for Information Systems* 36 (2015), 3.
[3] F. Ahmed and L. F. Capretz. 2007. Managing the business of software product line: An empirical investigation of key business factors. *Information and Software Technology* 49, 2 (2007), 194–208.
[4] F. Ahmed and Luiz F. Capretz. 2008. The software product line architecture: An empirical investigation of key process activities. *Information and Software Technology* 50, 11 (2008), 1098–1113.
[5] F. Ahmed, L. Fernando Capretz, and Shahbaz A. S. 2007. Institutionalization of software product line: An empirical investigation of key organizational factors. *Journal of Systems and Software* 80, 6 (2007), 836–849.
[6] A Banerjee et al. 2009. Hypothesis testing, Type I and Type II errors. *Industrial Psychiatry Journal* 18, 2 (2009), 127.
[7] A. Barcomb, K.-J. Stol, D. Riehle, and B. Fitzgerald. 2019. Why Do Episodic Volunteers Stay in FLOSS Communities?. In *Proceedings of the 41th International Conference on Software Engineering*. ACM/IEEE, New York, NY, USA, 948–954.
[8] D. Batra. 2018. Agile values or plan-driven aspects: Which factor contributes more toward the success of data warehousing, business intelligence, and analytics project development? *Journal of Systems and Software* 146 (2018), 249–262.
[9] J.-M. Becker and I. R. Ismail. 2016. Accounting for sampling weights in PLS path modeling: Simulations and empirical examples. *European Management Journal* 34, 6 (2016), 606–617.
[10] J.-M. Becker, K. Klein, and M. Wetzels. 2012. Hierarchical latent variable models in PLS-SEM: guidelines for using reflective-formative type models. *Long Range Planning* 45, 5-6 (2012), 359–394.
[11] J-M Becker, A. Rai, C. Ringle, and F. Volckner. 2013. Discovering Unobserved Heterogeneity in Structural Equation Models to Avert Validity Threats. *MIS Quarterly* 37 (09 2013), 665–694.
[12] J.-M Becker, C. M. Ringle, and M. Sarstedt. 2018. Estimating moderating effects in PLS-SEM and PLSc-SEM: interaction term generation* Data treatment. *Journal of Applied Structural Equation Modeling* 2, 2 (2018), 1–21.
[13] J.-M. Becker, C. M Ringle, M. Sarstedt, and F. Völckner. 2015. How collinearity affects mixture regression results. *Marketing Letters* 26, 4 (2015), 643–659.
[14] J. Benitez, J. Henseler, A. Castillo, and F. Schuberth. 2020. How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research. *Information & Management* 57, 2 (2020), 16.
[15] P. Bentler and D. Bonett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* 88, 3 (1980), 588.
[16] K. Bollen and R. Lennox. 1991. Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin* 110, 2 (1991), 305.
[17] K. A. Bollen. 2011. Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly* 35, 2 (2011), 359–372.
[18] P. Bourque et al. 2014. *Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0.* IEEE Computer Society, Washington, DC.
[19] A. S. Campanelli, R. D. Camilo, and F. S. Parreiras. 2018. The impact of tailoring criteria on agile practices adoption: A survey with novice agile practitioners in Brazil. *Journal of Systems and Software* 137 (2018), 366–379.

[20] D. T Campbell and D. W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56, 2 (1959), 81.
[21] Gabriel Cepeda-Carrion, Juan-Gabriel Cegarra-Navarro, and Valentina Cillo. 2019. Tips to use partial least squares structural equation modelling (PLS-SEM) in knowledge management. *Journal of Knowledge Management* (2019).
[22] K.-C. Chang, T. S. Sheu, G. Klein, and J. J. Jiang. 2010. User commitment and collaboration: Motivational antecedents and project performance. *Information and Software Technology* 52, 6 (2010), 672–679.
[23] K.-C. Chang, J.-H. Wong, Y. Li, Y.-C. Lin, and H.-G. Chen. 2011. External social capital and information systems development team flexibility. *Information and Software Technology* 53, 6 (2011), 592–600.
[24] Jun-Hwa Cheah, José L Roldán, Enrico Ciavolino, Hiram Ting, and T Ramayah. 2020. Sampling weight adjustments in partial least squares structural equation modeling: guidelines and illustrations. *Total Quality Management & Business Excellence* (2020), 1–20.
[25] J.-H. Cheah, M. Sarstedt, C. M. Ringle, T. Ramayah, and H. Ting. 2018. Convergent validity assessment of formatively measured constructs in PLS-SEM. *International Journal of Contemporary Hospitality Management* 30, 11 (2018), 3192–3210.
[26] C. C. Chen, J. Liu, and H.-G. Chen. 2011. Discriminative effect of user influence and user responsibility on information system development processes and project management. *Information and Software Technology* 53, 2 (2011), 149–158.
[27] H.-G. Chen, J. J. Jiang, G. Klein, and J. V. Chen. 2009. Reducing software requirement perception gaps through coordination mechanisms. *Journal of Systems and Software* 82, 4 (2009), 650–655.
[28] W.W. Chin. 1998. The partial least squares approach to structural equation modeling. *Modern Methods for Business Research* 295, 2 (1998), 295–336.
[29] W. Chin and R. P. Newsted. 1996. *Structural Equation Modeling Analysis with Small Samples Using Partial Least Square.* Sage, Thousand Oaks, CA.
[30] W. W. Chin. 1998. Issues and Opinion on Structural Equation Modeling. *MIS Quarterly* 22, 1 (1998), vii–xvi.
[31] W. W Chin. 2003. PLS-Graph.
[32] W. W. Chin and J. Dibbern. 2010. An Introduction to a Permutation Based Procedure for Multi-Group PLS Analysis: Results of Tests of Differences on Simulated Data and a Cross Cultural Analysis of the Sourcing of Information System Services Between Germany and the USA. In *Handbook of Partial Least Squares: Concepts, Methods and Applications*, V. Esposito Vinzi, W. W. Chin, J. Henseler, and H. Wang (Eds.). Springer, Berlin, 171–193.
[33] Chao-Min Chiu, Meng-Hsiang Hsu, and Eric T.G. Wang. 2006. Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision Support Systems* 42, 3 (2006), 1872–1888.
[34] G. A Churchill. 1979. A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research* 16, 1 (1979), 64–73.
[35] J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Routledge, New York, NY.
[36] J. Cohen. 1992. A power primer. *Psychological Bulletin* 112, 1 (1992), 155.
[37] R. Connolly. 2020. Why computing belongs within the social sciences. *Commun. ACM* 63, 8 (2020), 54–59.
[38] L. J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (01 Sep 1951), 297–334.
[39] A. C. Davison and D. V. Hinkley. 1997. *Bootstrap methods and their application.* Cambridge University Press, Cambridge.
[40] F. G. de Oliveira Neto, R. Torkar, R. Feldt, L. Gren, C. A. Furia, and Z. Huang. 2019. Evolution of statistical analysis in empirical software engineering research: Current state and steps forward. *J. Syst. Softw.* 156 (2019), 246 – 267.
[41] A. Diamantopoulos, M. Sarstedt, C. Fuchs, P. Wilczynski, and S. Kaiser. 2012. Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. *Journal of the Academy of Marketing Science* 40, 3 (2012), 434–449.
[42] A. Diamantopoulos and J. A Siguaw. 2006. Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management* 17, 4 (2006), 263–282.
[43] A. Diamantopoulos and H. M. Winklhofer. 2001. Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research* 38, 2 (2001), 269–277.
[44] T. Dijkstra and J. Henseler. 2015. Consistent partial least squares path modeling. *MIS Quarterly* 39, 2 (2015), 297–316.
[45] T. K. Dijkstra. 2010. Latent Variables and Indices: Herman Wold's Basic Design and Partial Least Squares. In *Handbook of Partial Least Squares: Concepts, Methods and Applications*, V. Esposito Vinzi, W. W. Chin, J. Henseler, and H. Wang (Eds.). Springer, 23–46.
[46] T. K. Dijkstra. 2017. A perfect match between a model and a mode. In *Partial least squares path modeling.* Springer, 55–80.
[47] T. K. Dijkstra and J. Henseler. 2015. Consistent and asymptotically normal PLS estimators for linear structural equations. *Computational Statistics & Data Analysis* 81 (2015), 10–23.
[48] T. K. Dijkstra and K. Schermelleh-Engel. 2014. Consistent Partial Least Squares for Nonlinear Structural Equation Models. *Psychometrika* 79, 4 (01 Oct 2014), 585–604.

[49] D. A Dillman, J. D Smyth, and L. Christian. 2014. *Internet, phone, mail, and mixed-mode surveys: the tailored design method.* John Wiley & Sons, Hoboken, NJ.
[50] A. L. Drolet and D. G. Morrison. 2001. Do we really need multiple-item measures in service research? *Journal of service research* 3, 3 (2001), 196–204.
[51] J. A Durlak. 2009. How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology* 34, 9 (2009), 917–928.
[52] T. Dybå, V. B. Kampenes, and D. Sjøberg. 2006. A systematic review of statistical power in software engineering experiments. *Information and Software Technology* 48, 8 (2006), 745–755.
[53] J. R. Edwards and R. P. Bagozzi. 2000. On the nature and direction of relationships between constructs and measures. *Psychological Methods* 5, 2 (2000), 155.
[54] B. Efron and R. Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* 1, 1 (1986), 54–75.
[55] V. Esposito Vinzi, T. Fahmy, Y. M. Chatelin, and M. Tenenhaus. 2007. PLS path modeling: some recent methodological developments, a software integrated in XLSTAT and its application to customer satisfaction studies. In *Proceedings of the Academy of Marketing Science Conference "Marketing Theory and Practice in an Inter-Functional World"*. 11–14.
[56] J. Evermann and M. Tate. 2016. Assessing the predictive performance of structural equation model estimators. *Journal of Business Research* 69, 10 (2016), 4565–4582.
[57] F. Faul, E. Erdfelder, A. Buchner, and A.G. Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4 (2009), 1149–1160.
[58] R. A. Fisher. 1992. *Statistical methods for research workers.* Springer, New York, NY.
[59] C. Fornell, M. D. Johnson, E W. Anderson, J. Cha, and B. E. Bryant. 1996. The American customer satisfaction index: nature, purpose, and findings. *Journal of Marketing* 60, 4 (1996), 7–18.
[60] C. Fornell and D. F. Larcker. 1981. Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research* 18, 1 (1981), 39–50.
[61] N. Forsgren, J. Humble, and G. Kim. 2018. *Accelerate: Building and Scaling High Performing Technology Organizations.* IT Revolution Press, Portland, Oregon.
[62] S. Fosso Wamba and L. Trinchera. 2014. Assessing unobserved heterogeneity in SEM using REBUS-PLS: A case of the application of TAM to social media adoption. In *Proceedings of the 20th Americas Conference on Information Systems.* AIS, Savannah, GA, 1-11.
[63] G. Franke and M. Sarstedt. 2019. Heuristics versus statistics in discriminant validity testing: a comparison of four procedures. *Internet Research* 29, 3 (2019), 430–447.
[64] C. Fuchs and A. Diamantopoulos. 2009. Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft* 69, 2 (2009), 195.
[65] D. Gefen, E.E. Rigdon, and D. Straub. 2011. Editor's comments: an update and extension to SEM guidelines for administrative and social science research. *MIS Quarterly* 35, 2 (2011), iii–xiv.
[66] S. Geisser. 1974. A predictive approach to the random effect model. *Biometrika* 61, 1 (1974), 101–107.
[67] D. L. Goodhue, W. Lewis, and R. Thompson. 2012. Does PLS have advantages for small sample size or non-normal data? *MIS Quarterly* 36, 3 (2012), 981–1001.
[68] O. Götz, K. Liehr-Gobbers, and M. Krafft. 2010. Evaluation of structural equation models using the partial least squares (PLS) approach. In *Handbook of partial least squares: Concepts, Methods and Applications.* Springer, Berlin, 691–711.
[69] G. C. Green, A. R. Hevner, and R. W. Collins. 2005. The impacts of quality and productivity perceptions on the use of software process improvement innovations. *Information and Software Technology* 47, 8 (2005), 543–553.
[70] S. P. Gudergan, C. M. Ringle, S. Wende, and A. Will. 2008. Confirmatory tetrad analysis in PLS path modeling. *Journal of Business Research* 61, 12 (2008), 1238–1249.
[71] J. F. Hair, M. C Howard, and C. Nitzl. 2020. Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. *Journal of Business Research* 109 (2020), 101–110.
[72] J. F. Hair, G. T. Hult, C. Ringle, and M. Sarstedt. 2016. *A primer on partial least squares structural equation modeling (PLS-SEM).* Sage, Thousand Oaks, CA.
[73] J. F. Hair, L. M. Matthews, R. L. Matthews, and M. Sarstedt. 2017. PLS-SEM or CB-SEM: updated guidelines on which method to use. *International Journal of Multivariate Data Analysis* 1, 2 (2017), 107–123.
[74] J. F. Hair, C. M Ringle, and M. Sarstedt. 2011. PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice* 19, 2 (2011), 139–152.
[75] J. F. Hair, J. J. Risher, M. Sarstedt, and C. M. Ringle. 2019. When to use and how to report the results of PLS-SEM. *European Business Review* 31, 1 (2019), 2–24.
[76] J. F. Hair and M. Sarstedt. 2019. Factors versus Composites: Guidelines for Choosing the Right Structural Equation Modeling Method. *Project Management Journal* 50, 6 (2019), 619–624.

[77] J. F. Hair, M. Sarstedt, C. M. Ringle, and S. P. Gudergan. 2017. *Advanced issues in partial least squares structural equation modeling*. Sage, Thousand Oaks, CA.
[78] J. F. Hair, M. Sarstedt, C. M. Ringle, and J. A. Mena. 2012. An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science* 40, 3 (2012), 414–433.
[79] T. Hastie, R. Tibshirani, and J. Friedman. 2008. *The elements of statistical learning: data mining, inference and prediction*. Springer, New York, NY.
[80] J. Henseler. 2017. Bridging design and behavioral research with variance-based structural equation modeling. *Journal of Advertising* 46, 1 (2017), 178–192.
[81] J. Henseler. 2017. Bridging Design and Behavioral Research With Variance-Based Structural Equation Modeling. *Journal of Advertising* 46, 1 (2017), 178–192.
[82] J. Henseler. 2018. Partial least squares path modeling: Quo vadis? *Quality & Quantity* 52, 1 (2018), 1–8.
[83] J. Henseler et al. 2014. Common beliefs and reality about PLS: Comments on Rönkkö and Evermann (2013). *Organizational Research Methods* 17, 2 (2014), 182–209.
[84] J. Henseler and T. Dijkstra. 2015. ADANCO 2.0 Professional.
[85] J. Henseler, G. Hubona, and P. A. Ray. 2017. Partial Least Squares Path Modeling: Updated Guidelines. *Partial Least Squares Path Modeling: Basic Concepts, Methodological Issues and Applications* (2017), 19.
[86] J. Henseler, C. M. Ringle, and M. Sarstedt. 2015. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science* 43, 1 (2015), 115–135.
[87] J. Henseler, C. M. Ringle, and M. Sarstedt. 2016. Testing measurement invariance of composites using partial least squares. *International Marketing Review* 33, 3 (2016), 405–431.
[88] J. Henseler, C. M. Ringle, and R. R. Sinkovics. 2009. The use of partial least squares path modeling in international marketing. In *New challenges to international marketing*. Emerald Group Publishing Limited, Bingley, UK, 277–319.
[89] J. Henseler and M. Sarstedt. 2013. Goodness-of-fit indices for partial least squares path modeling. *Computational Statistics* 28, 2 (2013), 565–580.
[90] J. Henseler and F. Schuberth. 2020. Using confirmatory composite analysis to assess emergent variables in business research. *Journal of Business Research* 120 (2020), 147 – 156.
[91] L. Hu and P. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling* 6, 1 (1999), 1–55.
[92] J. Hulland. 1999. Use of partial least squares (PLS) in strategic management research: A review of four recent studies. *Strategic Management Journal* 20, 2 (1999), 195–204.
[93] G. T. M. Hult et al. 2018. Addressing endogeneity in international marketing applications of partial least squares structural equation modeling. *Journal of International Marketing* 26, 3 (2018), 1–21.
[94] H. Hwang, M. Sarstedt, J. H. Cheah, and C. M. Ringle. 2020. A concept analysis of methodological research on composite-based structural equation modeling: bridging PLSPM and GSCA. *Behaviormetrika* 47, 1 (2020), 219–241.
[95] P. Ifinedo. 2011. Examining the influences of external expertise and in-house computer/ IT knowledge on ERP system success. *Journal of Systems and Software* 84, 12 (2011), 2065–2078.
[96] G. James, D. Witten, T. Hastie, and R. Tibshirani. 2013. *An introduction to statistical learning*. Springer, New York, NY.
[97] C. B. Jarvis, S. B. MacKenzie, and P. M. Podsakoff. 2003. A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research* 30, 2 (2003), 199–218.
[98] J. J. Jiang, G. Klein, S. Wu, and T.-P. Liang. 2009. The relation of requirements uncertainty and stakeholder perception gaps to project management performance. *Journal of Systems and Software* 82, 5 (2009), 801–808.
[99] H-W. Jung and D. R. Goldenson. 2008. The internal consistency and precedence of key process areas in the capability maturity model for software. *Empirical Software Engineering* 13, 2 (2008), 125–146.
[100] M. Keil and C. Park. 2010. Bad news reporting on troubled IT projects: Reassessing the mediating role of responsibility in the basic whistleblowing model. *Journal of Systems and Software* 83, 11 (2010), 2305–2316.
[101] R. E Kirk. 2001. Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement* 61, 2 (2001), 213–218.
[102] B. Kitchenham et al. 2017. Robust statistical methods for empirical software engineering. *Empirical Software Engineering* 22, 2 (2017), 579–630.
[103] B. Kitchenham and S. Charters. 2007. *Guidelines for performing systematic literature reviews in software engineering*. Technical Report. Technical report, Ver. 2.3 EBSE Technical Report. EBSE.
[104] N. Kock. 2019. From composites to factors: Bridging the gap between PLS and covariance-based structural equation modelling. *Information Systems Journal* 29, 3 (2019), 674–706.
[105] N. Kock. 2020. WarpPLS User Manual, version 7.
[106] N. Kock and P. Hadaya. 2018. Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods. *Information Systems Journal* 28, 1 (2018), 227–261.

[107] M.V. Kosti, R. Feldt, and L. Angelis. 2014. Personality, emotional intelligence and work preferences in software engineering: An empirical study. *Information and Software Technology* 56, 8 (2014), 973–990.
[108] H.C. Kraemer et al. 2003. Measures of clinical significance. *Journal of the American Academy of Child & Adolescent Psychiatry* 42, 12 (2003), 1524–1529.
[109] C. Larsen, K.and Bong. 2016. A Tool for Addressing Construct Identity in Literature Reviews and Meta-Analyses. *MIS Quarterly* 40, 3 (2016), 529–551.
[110] P. Lenberg, R. Feldt, and L.-G. Wallgren. 2015. Behavioral software engineering: A definition and systematic literature review. *Journal of Systems and Software* 107 (2015), 15–37.
[111] Y. Li, K.-C. Chang, Houn-Gee Chen, and J. J. Jiang. 2010. Software development team flexibility antecedents. *Journal of Systems and Software* 83, 10 (2010), 1726–1734.
[112] J. Liu, V. J. Chen, C.-L. Chan, and T. Lie. 2008. The impact of software process standardization on software flexibility and project management performance: Control theory perspective. *Inform. Softw. Technol.* 50, 9-10 (2008), 889–896.
[113] J.-B. Lohmöller. 1989. Predictive vs. structural modeling: Pls vs. ml. In *Latent variable path modeling with partial least squares*. Springer, 199–226.
[114] G.A Marcoulides and C. Saunders. 2006. Editor's comments: PLS: a silver bullet? *MIS quarterly* (2006), iii–ix.
[115] G. A Marcoulides, W. W Chin, and C. Saunders. 2009. A critical look at partial least squares modeling. *MIS Quarterly* 33, 1 (2009), 171–175.
[116] J. A. Martilla and J. C. James. 1977. Importance-Performance Analysis. *Journal of Marketing* 41, 1 (1977), 77–79.
[117] M. Mayeh, T. Ramayah, and A. Mishra. 2016. The role of absorptive capacity, communication and trust in ERP adoption. *Journal of Systems and Software* 119 (2016), 58–69.
[118] C. N. McIntosh, J. R. Edwards, and J. Antonakis. 2014. Reflections on partial least squares path modeling. *Organizational Research Methods* 17, 2 (2014), 210–251.
[119] P. Meso, G. Madey, M. D Troutt, and J. Liegle. 2006. The knowledge management efficacy of matching information systems development methodologies with application characteristics—an experimental study. *J. Syst. Softw.* 79, 1 (2006), 15–28.
[120] Bjørn-Helge Mevik and Ron Wehrens. 2007. The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software* 18, 2 (2007), 1–24.
[121] J. Miles. 2014. *Tolerance and Variance Inflation Factor.* American Cancer Society, Hoboken, NJ.
[122] A. Monecke and F. Leisch. 2012. semPLS: structural equation modeling using partial least squares. *Journal of Statistical Software* 48, 3 (2012).
[123] J. Nunnally. 1978. *Psychometric Theory.* McGraw-Hill, New York, NY.
[124] R. M. O'Brien. 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity* 41, 5 (Oct 2007), 673–690.
[125] A. Oztekin, A. Nikov, and S. Zaim. 2009. UWIS: An assessment methodology for usability of web-based information systems. *Journal of Systems and Software* 82, 12 (2009), 2038–2050.
[126] N. Parolia, J. V. Chen, J. J. Jiang, and G. Klein. 2015. Conflict resolution effectiveness on the implementation efficiency and achievement of business objectives in IT programs: A study of IT vendors. *Inform. Softw. Technol.* 66 (2015), 30–39.
[127] N. Parolia, J. J. Jiang, and G. Klein. 2013. The presence and development of competency in IT programs. *Journal of Systems and Software* 86, 12 (2013), 3140–3150.
[128] D.X. Peng and F. Lai. 2012. Using partial least squares in operations management research: A practical guideline and summary of past research. *Journal of Operations Management* 30, 6 (2012), 467–480.
[129] P.M. Podsakoff, S.B. MacKenzie, L. Jeong-Yeon, and N.P. Podsakoff. 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology* 88, 5 (2003), 879–893.
[130] P. Pollard and J. T. Richardson. 1987. On the probability of making Type I errors. *Psychological Bulletin* 102, 1 (1987), 159.
[131] P. Ralph et al. 2020. Pandemic Programming: How COVID-19 affects software developers and how their organizations can help. *Empirical Software Engineering* 25 (2020), 4927–4961. https://doi.org/10.1007/s10664-020-09875-y
[132] R. F Richter, G. Cepeda, J. L Roldan, and C. M Ringle. 2016. European management research using partial least squares structural equation modeling (PLS-SEM). *European Management Journal* 33, 1 (2016), 1–3.
[133] E.E. Rigdon. 2016. Choosing PLS path modeling as analytical method in European management research: A realist perspective. *European Management Journal* 34, 6 (2016), 598–605.
[134] E. Rigdon, J.-M. Becker, and M. Sarstedt. 2019. Factor indeterminacy as metrological uncertainty: implications for advancing psychological measurement. *Multivariate Behavioral Research* 54, 3 (2019), 429–443.
[135] E.E. Rigdon, M. Sarstedt, and C.M. Ringle. 2017. On comparing results from CB-SEM and PLS-SEM: Five perspectives and five recommendations. *Marketing Zfp* 39, 3 (2017), 4–16.

[136] E. E. Rigdon. 2012. Rethinking partial least squares path modeling: In praise of simple methods. *Long Range Planning* 45, 5-6 (2012), 341–358.
[137] E. E. Rigdon, C. M. Ringle, and M. Sarstedt. 2010. Structural modeling of heterogeneous data with partial least squares. In *Review of Marketing Research*. Emerald Group Publishing Limited, Bingley, UK, 255–296.
[138] C. M. Ringle and M. Sarstedt. 2016. Gain more insight from your PLS-SEM results: The importance-performance map analysis. *Industrial Management & Data Systems* 116, 9 (2016), 1865–1886.
[139] C. M. Ringle, M. Sarstedt, R. Mitchell, and S. P. Gudergan. 2018. Partial least squares structural equation modeling in HRM research. *The International Journal of Human Resource Management* 31, 12 (2018), 1–27.
[140] C. M. Ringle, M. Sarstedt, and E. A. Mooi. 2010. Response-based segmentation using finite mixture partial least squares. In *Data Mining*. Springer, Cham, 19–49.
[141] C. M. Ringle, M. Sarstedt, and D. Straub. 2012. A critical look at the use of PLS-SEM in MIS Quarterly. *MIS Quarterly* 36, 1 (2012), iii–xiv.
[142] C. M Ringle, S. Wende, and J.-M. Becker. 2015. SmartPLS 3.
[143] H. L. Romero, R. M. Dijkman, P. Grefen, A. J. van Weele, and A. De Jong. 2015. Measures of process harmonization. *Information and Software Technology* 63 (2015), 31–43.
[144] M Rönkkö. 2020. matrixpls: Matrix-based partial least squares estimation.
[145] M. Rönkkö and J. Evermann. 2013. A critical examination of common beliefs about partial least squares path modeling. *Organizational Research Methods* 16, 3 (2013), 425–448.
[146] M. Rönkkö, J. Evermann, and M.I. Aguirre-Urreta. 2016. Estimating Formative Measurement Models in IS Research – Analysis of the Past and Recommendations for the Future. http://urn.fi/URN:NBN:fi:aalto-201605031907.
[147] M. Rönkkö, C. McIntosh, and J. Antonakis. 2015. On the adoption of partial least squares in psychological research: Caveat emptor. *Personality and Individual Differences* 87 (2015), 76–84.
[148] M. Rönkkö, C. McIntosh, J. Antonakis, and J. Edwards. 2016. Partial least squares path modeling: Time for some serious second thoughts. *Journal of Operations Management* 47 (2016), 9–27.
[149] Y. Roumani, J. K. Nwankpa, and Y. F. Roumani. 2017. Adopters' trust in enterprise open source vendors: An empirical examination. *Journal of Systems and Software* 125 (2017), 256–270.
[150] D. Russo, P.H.P. Hanel, S. Altnickel, and N. van Berkel. 2021. Predictors of Well-being and Productivity of Software Professionals during the COVID-19 Pandemic–A Longitudinal Study. *Empirical Software Engineering* (2021).
[151] G. Sanchez. 2013. *PLS Path Modeling with R.* Trowchez Editions.
[152] M. Sarstedt. 2008. A review of recent approaches for capturing heterogeneity in partial least squares path modelling. *Journal of Modelling in Management* 3, 2 (2008), 140–161.
[153] M. Sarstedt et al. 2020. Structural model robustness checks in PLS-SEM. *Tourism Economics* 26, 4 (2020), 531–554.
[154] M. Sarstedt, P. Bengart, A. Shaltoni, and S. Lehmann. 2018. The use of sampling methods in advertising research: A gap between theory and practice. *International Journal of Advertising* 37, 4 (2018), 650–663.
[155] M. Sarstedt, A. Diamantopoulos, and T. Salzberger. 2016. Should we use single items? Better not. *Journal of Business Research* 69, 6 (2016), 1224–1231.
[156] M. Sarstedt, J. Hair, J.-H. Cheah, J.-M. Becker, and C. M. Ringle. 2019. How to specify, estimate, and validate higher-order constructs in PLS-SEM. *Australasian Marketing Journal* 27, 3 (2019), 197–211.
[157] M. Sarstedt, J. F. Hair, C. Nitzl, C. M. Ringle, and M. C. Howard. 2020. Beyond a tandem analysis of SEM and PROCESS: Use of PLS-SEM for mediation analyses! *International Journal of Market Research* 62, 3 (2020), 288–299.
[158] M. Sarstedt, J. Henseler, and C. Ringle. 2011. Multigroup analysis in partial least squares (PLS) path modeling: Alternative methods and empirical results. In *Measurement and Research Methods in International Marketing*. Emerald Group Publishing Limited, Bingley, UK, 195–218.
[159] M. Sarstedt and E. Mooi. 2014. *A concise guide to market research: The process, data, and methods using IBM SPSS Statistics*. Springer, Berlin.
[160] M. Sarstedt, C.M. Ringle, J.-H. Cheah, H. Ting, O. I Moisescu, and L. Radomir. 2020. Structural model robustness checks in PLS-SEM. *Tourism Economics* 26, 4 (2020), 531–554.
[161] M. Sarstedt and C. M. Ringle. 2010. Treating unobserved heterogeneity in PLS path modeling: a comparison of FIMIX-PLS with different data analysis strategies. *Journal of Applied Statistics* 37, 8 (2010), 1299–1318.
[162] M. Sarstedt, C. M. Ringle, and J. F. Hair. 2017. Treating unobserved heterogeneity in PLS-SEM: A multi-method approach. In *Partial least squares path modeling*. Springer, Cham, 197–217.
[163] M. Sarstedt, M. Schwaiger, and C. M. Ringle. 2009. Do we fully understand the critical success factors of customer satisfaction with industrial goods? Extending Festge and Schwaiger's model to account for unobserved heterogeneity. *Journal of Business Market Management* 3, 3 (2009), 185.
[164] F. Schuberth. 2020. Confirmatory composite analysis using partial least squares: setting the record straight. *Review of Managerial Science* (2020), 1–35.

[165] F. Schuberth, J. Henseler, and T. K. Dijkstra. 2018. Confirmatory composite analysis. *Frontiers in Psychology* 9 (2018), 25–41.
[166] P.N. Sharma, G. Shmueli, M. Sarstedt, N. Danks, and S. Ray. 2018. Prediction-Oriented Model Selection in Partial Least Squares Path Modeling. *Decision Sciences* forthcoming (2018).
[167] J. T. Shim, T. S. Sheu, H.-G. Chen, J. J. Jiang, and G. Klein. 2010. Coproduction in successful software development projects. *Information and Software Technology* 52, 10 (2010), 1062–1068.
[168] G. Shmueli et al. 2010. To explain or to predict? *Statist. Sci.* 25, 3 (2010), 289–310.
[169] G. Shmueli, S. Ray, J. Estrada, and S. Chatla. 2016. The elephant in the room: Predictive performance of PLS models. *Journal of Business Research* 69, 10 (2016), 4552–4564.
[170] G. Shmueli, M. Sarstedt, J. F. Hair, J.-H. Cheah, H. Ting, S. Vaithilingam, and C. M. Ringle. 2019. Predictive model assessment in PLS-SEM: guidelines for using PLSpredict. *European Journal of Marketing* 53, 11 (2019), 2322–2347.
[171] K. Sijtsma. 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74, 1 (2009), 107.
[172] S. G Sireci. 1998. The construct of content validity. *Social Indicators Research* 45, 1-3 (1998), 83–117.
[173] N. Slack. 1994. The Importance-Performance Matrix as a Determinant of Improvement Priority. *International Journal of Operations & Production Management* 14, 5 (1994), 59–75.
[174] S. Y. Sohn and M. S. Mok. 2008. A strategic analysis for successful open source software utilization based on a structural equation model. *Journal of Systems and Software* 81, 6 (2008), 1014–1024.
[175] K.-J. Stol and B. Fitzgerald. 2018. The ABC of software engineering research. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 27, 3 (2018), 51.
[176] M. Stone. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* 36, 2 (1974), 111–147.
[177] M.-A. Storey, N. A. Ernst, C. Williams, and E. Kalliamvakou. 2020. The who, what, how of software engineering research: a socio-technical framework. *Empirical Software Engineering* 25 (2020), 4097–4129.
[178] D. Straub, M. Boudreau, and D. Gefen. 2004. Validation guidelines for IS positivist research. *Communications of the Association for Information Systems* 13, 1 (2004), 24.
[179] G. H Subramanian, J. J Jiang, and G. Klein. 2007. Software quality and IS project performance improvements from software development process maturity and IS implementation strategies. *J. Syst. Softw.* 80, 4 (2007), 616–627.
[180] G. Svensson et al. 2018. Framing the triple bottom line approach: direct and mediation effects between economic, social and environmental elements. *Journal of Cleaner Production* 197 (2018), 972–991.
[181] S. Talukder et al. 2020. Predicting antecedents of wearable healthcare technology acceptance by elderly: A combined SEM-Neural Network approach. *Technological Forecasting and Social Change* 150 (2020), 119793.
[182] M. Tenenhaus, S. Amato, and V. Esposito Vinzi. 2004. A global goodness-of-fit index for PLS structural equation modelling. In *Proceedings of the XLII SIS scientific meeting*, Vol. 1. 739–742.
[183] M. Tenenhaus, V. E. Vinzi, Y. M. Chatelin, and C. Lauro. 2005. PLS path modeling. *Computational Statistics & Data Analysis* 48, 1 (2005), 159–205.
[184] B. Thompson. 2002. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher* 31, 3 (2002), 25–32.
[185] J.B. Ullman. 2006. Structural Equation Modeling: Reviewing the Basics and Moving Forward. *Journal of Personality Assessment* 87, 1 (2006), 35–50.
[186] N. Urbach and F. Ahlemann. 2010. Structural equation modeling in information systems research using partial least squares. *Journal of Information Technology Theory and Application* 11, 2 (2010), 5–40.
[187] L. Vijayasarathy and D. Turk. 2012. Drivers of agile software development use: Dialectic interplay between benefits and hindrances. *Information and Software Technology* 54, 2 (2012), 137–148.
[188] E. T. G. Wang, S.-P. Shih, J. J. Jiang, and G. Klein. 2006. The relative influence of management control and user–IS personnel interaction on project performance. *Information and Software Technology* 48, 3 (2006), 214–220.
[189] E. T. G. Wang, T.-C. Ying, J. J. Jiang, and G. Klein. 2006. Group cohesion in organizational innovation: An empirical examination of ERP implementation. *Information and Software Technology* 48, 4 (2006), 235–244.
[190] R. Wehrens and B.-H. Mevik. 2007. The PLS package: principal component and partial least squares regression in R.
[191] C.E. Werts, R.L. Linn, and K.G. Jöreskog. 1974. Intraclass Reliability Estimates: Testing Structural Assumptions. *Educational and Psychological Measurement* 34, 1 (1974), 25–33.
[192] H. Wold. 1982. Soft modeling: the basic design and some extensions. In *Systems under indirect observation*, K. Jöreskog and H. Wold (Eds.). North-Holland, 1–54.
[193] E. J. Wolf, K. M. Harrington, S. L. Clark, and M. W. Miller. 2013. Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement* 73, 6 (2013), 913–934.