Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

The Teamwork Process Antecedents (TPA) Questionnaire: Developing and Validating a Comprehensive Measure for Assessing Antecedents of Teamwork Process Quality

George Marsicano[1,2], Fabio Q. B. da Silva[1,3], Carolyn B. Seaman[3], Breno Giovanni Adaid-Castro[4]
georgemarsicano@unb.br*, fabio@cin.ufpe.br*, cseaman@umbc.edu, brenoadaid@gmail.com
[1] Centre for Informatics, Federal University of Pernambuco Recife, PE, Brazil.
[2] University of Brasilia, Brasília, DF, Brazil.
[3] University of Maryland Baltimore County, Baltimore, USA.
[4] Higher Education Institute of Brasilia, Brasilia, DF, Brazil.

ABSTRACT

Context: Most models of teamwork describe team behavior and effectiveness using an Input-Process-Output approach. In software engineering, the use of such models has focused on understanding and operationalizing the Process-Output components while less research effort has been applied to define and measure the Input-Process component.

Goal: To develop and validate a measure of team process antecedents (inputs) that addresses specific characteristics of software teams in industrial practice.

Method: First, we reviewed the group work literature, identified and integrated previously described antecedents of work group process, and developed a measure to tap those antecedents. This measure is operationalized in the Team Process Antecedent (TPA) questionnaire, which we then validated with 375 Brazilian software engineers from 100 companies, using exploratory and confirmatory factor analysis.

Results: We created a survey to operationalize two multidimensional antecedents of teamwork process, Team Structure and Team Composition, based on well-established models from the literature on work teams. We tailored the response items to the software engineering context to increase construct face validity. We reached a parsimonious set of five dimensions for Team Composition (16 response items) and four dimensions for Team Structure (11 response items). Our results show that our measure of TPA has excellent internal reliability and convergent and discriminant validity.

Conclusions: We created a novel measure of antecedents of teamwork process tailored to software teams, that captures the perception of team members about the adequacy of team composition and structure to achieve team goals. Further, we present the development of the TPA measure in the form of a guideline that may be used in the construction of other measurement instruments in empirical software engineering research. We believe both results are important contributions of this work.

Keywords: *software engineering, measurement instrument, questionnaire, exploratory factor analysis, confirmatory factor analysis.*

1 Introduction

The literature on work teams includes several models of individual and team behavior, e.g. Gladstein (1984), Hackman (1987), and Cohen (1993), to cite just a few. These models are useful to interpret teamwork outcomes and, most importantly, to manage teams to produce desired outcomes. Such models often describe team behavior and effectiveness using the Input-Process-Output framework (IPO) originally introduced by McGrath (1964), and later refined by Gladstein (1984). In this framework, **inputs** at individual, team, and organizational levels are combined and used by team level **processes** to produce desired (or not) teamwork **outputs**.

In the study of software teams, the use of IPO models has mostly focused on the Process-Output components. For instance, The Teamwork Quality model (TWQ) (Hoegl and Gemuenden, 2001) focused on how the interactions between members of software development teams (team processes) are related to team outputs. The TWQ construct is composed of six team processes (also called facets): communication, coordination, balance of members' contribution to teamwork, mutual support, effort, and cohesion. They also showed empirically that the TWQ construct is related to software development team success.

However, less focus has been placed on the Input-Process component of the IPO models in software engineering. In fact, in 111 studies recently synthesized in a systematic literature review (De Oliveira, 2019), just under 8.5% (9/111) studied antecedents (inputs) of teamwork process at individual and team levels. Among them, only five

used a measure for some factor of Team Composition (4,5%) and one for Team Structure (under 1%), which are postulated as antecedents of teamwork processes in most IPO models. None of the studies proposed, built, or validated a measure of team process antecedents that consistently included both composition and structure.

This is an important research gap because understanding antecedents of teamwork processes may provide practical ways to build and develop teams that perform well. In practice, team structure and composition are (at least partially) under the control of project managers, team leaders, human resources (HR) staff, and the team itself. Thus, building and managing teams with structure and composition that facilitate better processes could be a path towards higher teamwork effectiveness. Further, a measure of team process antecedents, such as structure and composition, is essential for more holistic studies of software teams in practice. Using such a measure, combined with measures for process quality (like TWQ, for instance) and team effectiveness, will support the construction of more comprehensive and complete models of software team effectiveness.

To help close this research gap, we created a measure of team process antecedents for use in research as well as in managing teams in practice. We started by reviewing the work group literature, where we identified and integrated previously described antecedents of work group process, and then developed a measure to tap those antecedents. The measure is operationalized in the Team Process Antecedent (TPA) questionnaire, which we validated with 375 Brazilian software engineers from 100 companies. In this validation, we applied an extensive set of statistical tests, in particular exploratory and confirmatory factor analysis. Our results show that our measure of TPA has excellent internal reliability and convergent and discriminant validity. TPA may be applied in research about software teams and also used by project managers, team leaders, and HR personnel to capture the perception of software team members about the adequacy of structure and composition of their teams to achieve team goals. We discuss potential applications of TPA in Section 5.

As a second, but equally important contribution, we present our work in the form of a guideline that could be used by other researchers in the construction and validation of measurement instruments in software engineering research. To the best of our knowledge, there is no published guideline specifically addressing the construction and validation of measurement instruments in empirical software engineering.

The rest of this paper is structured as follows. In Section 2, we present the conceptual underpinnings of our study. In Section 3, we outline the research method used to develop and validate the TPA questionnaire. In Section 4, we describe how TPA was developed from relevant literature, and how we validated TPA. In Section 5, we discuss the implications of our results, including the potential applications of TPA in software engineering research and practice. Finally, in Section 6, we present some conclusions and directions for future research.

2 Conceptual Background and Related Work

In this section, we start with a brief overview of teamwork concepts and terminology. We then discuss some of the teamwork Input-Process-Output models in the literature, focusing on Gladstein's model in detail, as it is central to our work. Finally, we discuss the gaps in the related work in software engineering research.

2.1 Characterizing Work Teams

In the literature, there are varying definitions of 'team' and 'group', a discussion of which is outside the scope of our study. In our work, we shall use Gladstein's (1984) definition: "a set of interdependent individuals who view themselves as a group and perform a task defined by the organization". This definition applies to software teams working in organizational settings, which are the focus of our research. We, thus, are not directly addressing open source software communities.

Further, we use the term 'teams' or 'software teams', and 'teamwork' to express the work performed by software teams. The term 'software team' is more common than 'software group' both in research and practice, but 'group' does appear in the literature, and we preserve that term when citing other work when the authors used it.

2.2 Teamwork Input-Process-Output Models

In the literature, behavior and outcomes of teamwork have been described or explained using an Input-Process-Output (IPO) model or framework since the seminal work by McGrath (1964). In such models, inputs at individual, team, and organizational levels are combined and used by team level processes to produce desired (or not) teamwork outputs. The use of an IPO model to frame team effectiveness is particularly important for capturing the dynamic interactions and emergent states that constitute teamwork (Salas, et al., 2007). Figure 1 shows the generic structure of the IPO framework, adapted from Hackman (1987) and based on the original work of McGrath (1964).

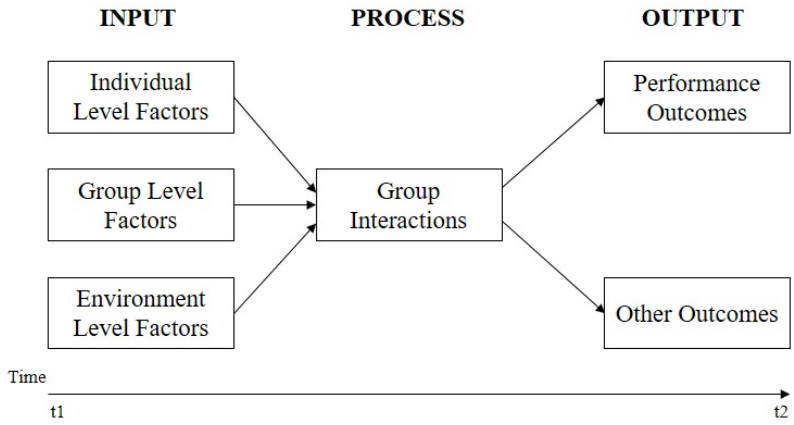

Fig. 1 *Input-Process-Output Framework (Adapted from Hackman (1987)*

A comprehensive literature review of IPO models is outside the scope of this article (we refer the reader to Mathieu et al. (2008)). In Table 1, we summarize and compare some of the most referenced models in the literature, listing the main factors addressed by each, excluding moderators and mediators for brevity.

Table 1. Summary of Input-Process-Output Models of Teamwork.

Model	INPUT	PROCESS	OUTPUT
Gladstein (1984)	Group Composition Group Structure Resources Available Organizational Structure	Open Communication Supportiveness Conflict Discussion of Strategy Weighting Individual Inputs Boundary Management	Group Performance Individual Satisfaction
Hackman (1987)	Structure of the task Composition of the group Group norms about performance processes Education System Reward System Information System	Group Synergy	Task Output Capability to work together as group in future Individual satisfaction with group experience
Campion, et al. (1993)	Job Design Interdependence Composition Context	Potency Social Support Workload Sharing Communication/Cooperation within groups	Productivity Satisfaction Manager Judgments
Cohen (1993)	Task Design Group Composition Organizational Context Environmental Factors	Conflict Communication Collaboration Task Process Social Integration	Team Performance Member Attitudes Quality of Work Life Withdrawal Behaviors

Inputs are factors that enable or constrain the performance of team processes. They are antecedents of the intra-team interactions that are part of team processes. In most models, input factors are grouped into at least two levels: the factors directly related to team characteristics, such as structure and composition; and factors related to the environment or context in which the team exists and works. The design of the task or job is also seen as an input factor in the models of Hackman (as part of group design), Campion, and Cohen. In Gladstein's Model, task design factors are addressed as moderators between team processes and outputs (see Section 2.3).

Team processes are interdependent actions performed by the team members to accomplish individual and collective goals (Marks, et al., 2001). Beyond the tasks related to the work itself (such as programming or testing software), team processes include a diverse set of interactions among team members that are necessary (or not) for the work itself. The quality of these interactions is directly related to the effectiveness of the team, expressed by the quality and levels of team outputs.

Outputs are the results generated by the activities performed by the team, which often include some type of product or service, or both. Outputs also include team member reactions to teamwork, such as individual satisfaction and commitment, burnout, and withdrawal behaviors. Most models use a multidimensional characterization of effectiveness to conceptualize the desired team outputs, beyond productivity (efficiency) and quality (efficacy).

To structure our initial findings and to support our initial conceptualizations, we chose Gladstein's model (1984) as a framework. As described in Section 3.1 (Phase 2), Gladstein's model fits our results more consistently than the other models presented in Table 1. As it plays a central role in our work, we describe the model in the next section, as part of our conceptual background.

2.3 Gladstein's Model of Team Effectiveness

As shown in Figure 2, Gladstein's model structures inputs into two groups: the team level (team structure and team composition) and the organization level (available resources and organizational structure). Each second order construct within each group is further refined into first order constructs that are then individually conceptualized and operationalized. The model explains the influence of these inputs on team processes as well as on team effectiveness, and the relationship between team processes and team effectiveness, moderated by the tasks performed by the team.

At the team level, structure is viewed as the relatively stable arrangement among people, expressed in terms of division and specialization of work and methods of coordination and control. In this sense, Gladstein's concept of structure includes Hackman's (1987) concept of 'structure of the task' and 'group norms about performance processes'. Measurable indicators of structure include role and goal clarity, specific work norms, task control, size, and formal leadership. Group composition has four dimensions: skills needed to perform the task, group heterogeneity that assures positive interaction, and experience with the job and organization that assures a group's knowledge of standard operating procedures.

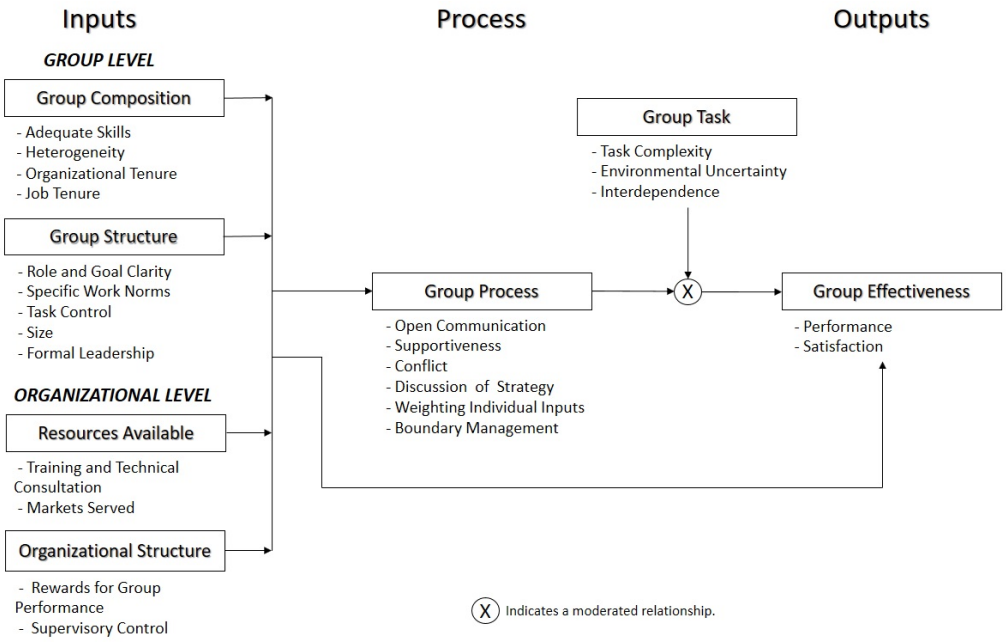

Fig. 2 *Gladstein's model (1984)*

The inputs at the organizational level, available resources (training availability and market growth) and organizational structure (supervisory behavior and rewards for group performance), are contextual variables that are also seen to affect team processes and effectiveness. While team level factors are likely to change or be adapted for different teams, organizational level factors tend to be more uniform across the organization.

Group process is characterized by two different types of behavior: maintenance behaviors and task behaviors. The first is articulated by the humanistic school and is assessed by three scales: open communication of ideas and feelings, supportiveness, and interpersonal conflict. The second is articulated by the decision-making and the boundary-management theorists and is assessed by weighting individual inputs by knowledge and skill, discussion of performance strategies in novel situations, and managing the boundary with other groups (Gladstein, 1984).

The model also posits that the relationship between group process and effectiveness is not constant, but is thought to vary with the nature of the task to be performed. The tasks are characterized by task complexity, task interdependence, and environmental uncertainty. This moderating relationship implies that, in order to be effective, the group must have an

information-processing capacity that matches the information-processing requirements of its task.

The last component of the model is group effectiveness, represented by two sets of outputs: group performance and group-member satisfaction. Group performance is related to the capacity of the group to meet the goals established for the period (e.g. revenue). Group-member satisfaction corresponds to satisfaction with being a team member, satisfaction with the job, the compensation system, the method of evaluation, and workload, and the satisfaction with dealing with the customer and meeting customer needs.

Our goal in this study is to capture relevant antecedents of teamwork in software development and create a measure to operationalize these antecedents. Therefore, we shall focus on the input side of the IPO framework and use Gladstein's model to frame our research. In Section 5.2, we discuss the potential limitations of this focus and show directions for future work that could address them. Further, we also show in Section 5.4 how the measure we developed with this focus can be applied in practice.

2.4 Empirical Research about Team Process Antecedents in Software Engineering Industrial Practice

A recent systematic mapping study identified 111 empirical studies of software teams conducted in industrial practice (De Oliveira, 2019). In this review, just under 8.5% (9/111) of the selected studies investigated antecedents of teamwork process at individual (3/111) and team levels (6/111). Table 2 summarizes the six studies that addressed team process antecedents at the team level, which is the focus of our research.

Table 2. Studies of Antecedents (team level) of Team Process (De Oliveira, 2019).

Antecedents	Process	Output	Data Collection Technique	Study
Task Interdependence, Role, Domain knowledge	Communication and Coordination	-	Interview	Damian, et al. (2013)
Goal Setting, Team-External Influence over Project Decisions, Team-Internal Equality of Influence over Project Decisions, Team Proximity, Task innovative	Teamwork Quality (TWQ): Communication, Coordination, Balance of Members' Contribution to Teamwork, Mutual Support, Effort, and Cohesion	Team Performance	Questionnaire and interview	Hoegl and Parboteeah (2003), Hoegl and Parboteeah (2006a), Hoegl and Proserpio (2004), Hoegl, et al. (2003)
Team climate, Team leadership	Cooperation and Competition	Team Performance, Team agility	Questionnaire	Liu, et al. (2014)
Specialized skills, Division of work	Dickinson and McIntyre's teamwork model: Communication, Monitoring, Feedback, Backup Behavior, Coordination.	-	Interview	Moe and Dyba (2010)
Shared leadership, Team Orientation, Redundancy (skills), Learning (Shared Mental Models), Autonomy	Teamwork process	-	Focus group	Ringstad, et al. (2011)
Diversity in Team Composition	Relationship Conflict	-	Questionnaire	Wickramasinghe and Nandula (2015)

It is important to note that the articles by Hoegl and his colleagues (Hoegl and Parboteeah, 2003; Hoegl, et al., 2003; Hoegl and Proserpio, 2004; Hoegl and Parboteeah, 2006a) were part of a larger study about Teamwork Quality (TWQ) (Hoegl and Gemuenden, 2001). From this initial study, different data analyses were performed, with different

goals, but using the same data set. We thus consider the articles derived from the work on TWQ (Hoegl and Gemuenden, 2001) as a single study.

Table 2 identifies important research gaps both at the conceptual and operational levels. At the conceptual level, each study addressed a different set of antecedents and processes, with very few intersections. Thus, it is difficult to compare their results or to aggregate them to increase validity and reliability, and to move towards a more holistic and comprehensive theory of team work in software engineering.

At the operational level, the development and validation of measures using questionnaires is not described in a way that shows that a rigorous and consistent process was followed. In general, the studies focused only on the presentation of an internal reliability index (e.g. Cronbach's alpha), without addressing other aspects, such as construct validity (convergent and discriminant). According to Clark and Watson (1995), internal reliability is important (necessary condition), but not sufficient to demonstrate that the research instrument measures a construct accurately. Further, none of them proposed, built, and validated a measure of team process antecedents that consistently included both composition and structure, which is the contribution of our work.

3 Method

The research procedures used in this study were based on the model proposed by Pasquali (2010), which addresses three aspects of instrument elaboration and validation: theoretical, empirical, and analytical. Pasquali's model was chosen because it makes each phase of the construction and validation of the measurement instrument explicit, thus increasing the confidence that the process was conducted consistently and completely. Further, the explicit phases and tasks of the model may be helpful to other researchers creating or evolving their own instruments. Figure 3 shows the phases and associated tasks in Pasquali's model. In the rest of this section, we provide the details of how these tasks were carried out.

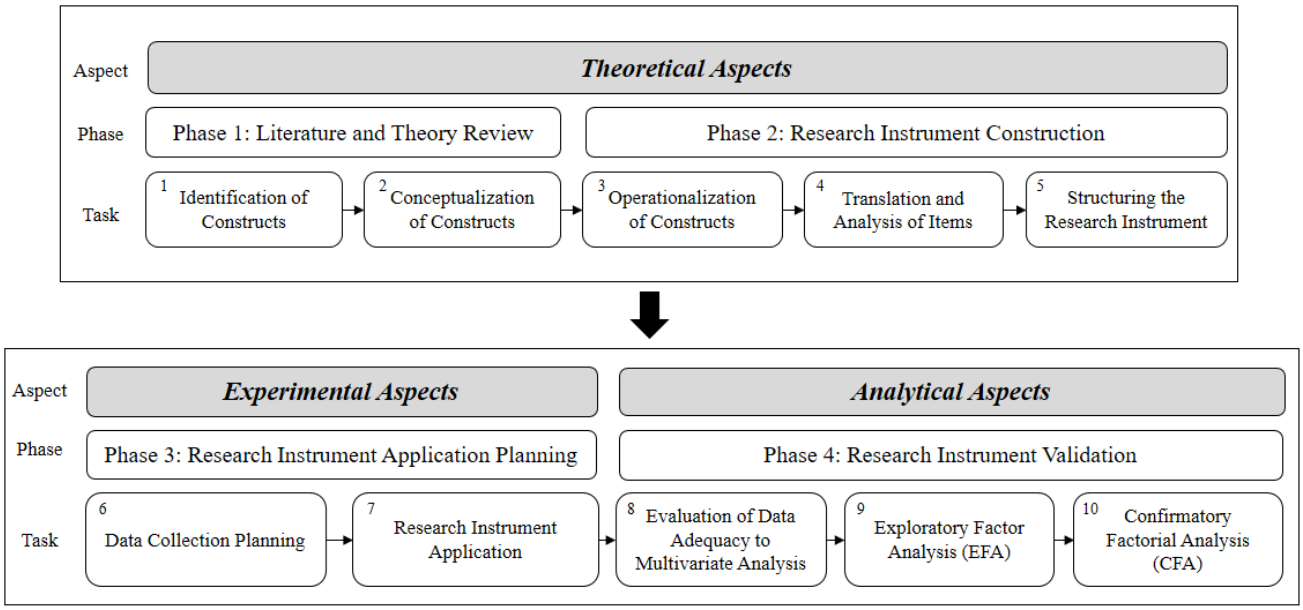

Fig. 3 *Detailed Process of Instrument Development. Adapted from Pasquali (Pasquali 2010)*

3.1 Theoretical Aspects

The theoretical aspects of designing and validating a measurement instrument (Pasquali 2010) include the choice of a theory about the constructs one wants to measure, as well as the operationalization of the constructs by a set of response items and a measuring scale. This is accomplished in two phases.

Phase 1: Literature and Theory Review

In this phase, empirical and conceptual work from the literature is reviewed and used to identify constructs of interest and provide consistent conceptualization of the constructs.

Task 1: Identification of Constructs

The objective is to identify relevant constructs to include in the measurement instrument. In general, the identification starts from relevant theory and previous conceptual and empirical research on the topic of interest. Typical sources of

information for this task are *ad hoc* or systematic literature reviews (or mapping studies). Further, exploratory studies, in particular in-depth qualitative enquiries, could be used to identify constructs from the target population. In both cases, it is essential to base the identification of constructs on solid theoretical foundations (Pasquali, 2010).

In this study, we started this task by performing an exploratory qualitative study. We then used Gladstein's Model to structure the initial findings. Finally, we reviewed theories and models from the literature on teamwork to consolidate the results.The exploratory qualitative study took place between July and December 2016. In depth, semi-structured interviews were performed with 26 members of 8 distinct software teams from four Brazilian software companies, in three locations in Brazil. We asked participants about their perceptions of teamwork and the antecedents of team effectiveness. We refer the reader to Pereira, et al. (2017) and Marsicano, et al. (2017) for details of the study that are outside the scope of this article. Table 3 presents the set of constructs (teamwork antecedents) that resulted from the analysis and interpretation of the interviews, which was performed without the use of an *a priori* model or theory.

Table 3. List of Antecedents of Teamwork Process from Pereira, et al. (2017) and Marsicano, et al. (2017).

List of Construct for Antecedents of Team Process
Team experience in the Organization
Team experience with work
Skills (interpersonal, managerial and technical)
Roles and responsibilities
Work organization (work processes)
Team size
Leadership style

We then chose Gladstein's Model (1984) to structure the results from the qualitative study. We started with the four groups of inputs of Gladstein's Model: Team Structure, Team Composition, Organizational Structure, and Resources Available, but our results populated only the first two, i.e. the team level factors. One plausible explanation is that our interviewees were software team leaders and members, who may be more likely to recall antecedents that are under their influence. So, we decided to focus on team level factors because the resulting measurement instrument would be more directly applicable to practice. We shall discuss the potential limitations of this focus and future research to close possible gaps in Section 5.2.

To resolve the differences between our set of constructs (Table 3) and those from Gladstein's model (Figure 2), we conducted an ad-hoc literature review looking for other models of teamwork and team effectiveness (APPENDIX A), which resulted in the consolidated set of constructs shown in Figure 5. Figure 4, extracted from Gladstein's model is presented here to facilitate comparison.

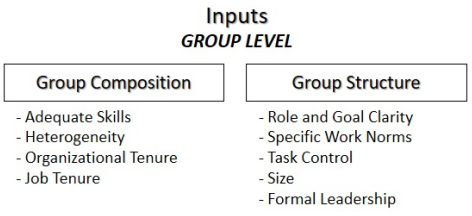

Fig. 4 *Team Structure and Composition from Gladstein's Model*

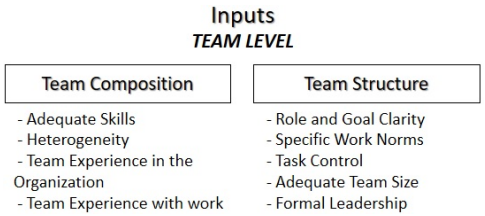

Fig. 5 *Set of Antecedent Constructs of Teamwork*

Task 2: Conceptualization of Constructs

The objective of this task is to establish clear-cut and consistent definitions of the constructs identified in the previous task, sharpening and improving construct definition, and raising the theoretical level of constructs. Theories and models from the literature, as well as the opinions of specialists and the experience of the researchers are used to clarify definitions and to choose between competing conceptualizations. Tasks 1 and 2 are often conducted in parallel in practice. To perform Task 2, we started with Gladstein's conceptualization for the initial set of constructs and complemented them with definitions from the related work reviewed in Task 1. The resulting glossary of terms is presented in Table 4 and Table 5.

Table 4. Constructs (1st level) of Team Composition.

Construct	Description
Adequate Skills	Refers to the required skills that the software development team must possess to execute their tasks. It is measured through the team member's perspective about the adequacy of the team competencies (Gladstein, 1984).
Heterogeneity	It refers to the degree of heterogeneity of the team in terms of knowledge, skills, attitudes, and experiences, aiming to ensure a positive interaction among its members (Gladstein, 1984).
Team Experience in the Organization	Refers to the work experiences in terms of roles and teamwork performed by the team members in the organization (Pereira, et al., 2017; Marsicano, et al., 2017).
Team Experience with work	Refers to the team members' experiences in terms of similarity of past and current activities (Pereira, et al., 2017; Marsicano, et al., 2017).

Table 5. Constructs (1st level) of Team Structure.

Construct	Description
Role and Goal Clarity	Refers to the degree to which the goals and roles of the team are specified, understood and accepted by the team (Gladstein, 1984).
Specific Work Norms	Refers to how the team members should behave, their routines, and work procedures (Gladstein, 1984; Levine, 1990).
Task Control	Refers to the degree of control or authority that a team has over its internal work processes (Hackman, 1980).
Adequate Team Size	Refers to whether or not the team is large enough to meet the objectives and goals proposed to it (Wagman, et al., 2005).
Formal Leadership	Refers to the leader's behavior in relation to the team and the organization (Gladstein, 1984).

Phase 2: Research Instrument Construction

The first task in this phase is operationalization of the constructs (i.e. developing response items) identified and conceptualized in the previous phase. Then, whenever necessary, the instrument is translated to the target language in which it will be validated. Finally, the data collection instrument is built in electronic or paper format, or both.

Task 3: Operationalization of Constructs

Response items empirically operationalize the conceptual definitions that define the theoretical constructs. After response items are developed (or reused from other studies), the rating scales are defined. We developed our response items in two steps. First, we revisited the studies used in Phase 1 looking for existing items for each construct. When existing items could not be found, we still used these studies to support the creation of our own response items and to adapt them to the software engineering context. We used these guidelines to build response items:

- When creating new items from the results of our exploratory study, we based their syntactic and semantic structure as close as possible on how interviewees expressed the related constructs.
- When reusing items or creating new items from studies in the literature, we tried to keep the wording of the items as close as possible to wording used in the studies.
- When mulitiple definitions were found, we chose those that more closely matched the definitions of the constructs presented in Table 4 and Table 5.
- Finally, we reviewed the wording of the items to make them more consistent with the jargon and technical terms used in software development practice, sometimes refering back to the exploratory qualitative study results.

This task resulted in an initial set of 51 response items that operationalize the five constructs related to team composition and four related to team structure, shown in Table 6.

Table 6. Item Quantity and Reference Overview by Construct.

Construct (2nd Order)	Latent Construct (1st Order)	# of Items	Literature
Team Composition	Adequate Skills	25	New items based on Aladwani, 2002; (Hoegl and Parboteeah, 2006a; 2006b; 2006c); Pereira, et al., 2017; Marsicano, et al., 2017.
	Heterogeneity	3	Items adapted from Campion, et al. (1993).[A]
	Team Experience in the Organization	3	New items based on Gladstein, 1984; Aladwani, 2002; Pereira, et al., 2017; Marsicano, et al., 2017.
	Team Experience with work	5	
Team Structure	Role and Goal Clarity	4	New items based on Gladstein, 1984; Aladwani, 2002; Hoegl and Parboteeah, 2003.
	Specific Work Norms	4	New item based on Marsicano, et al. (2017). Items adapted from Wageman, et al. (2005).[B]
	Task Control	1	Item adapted from Wageman, et al. (2005).[C]
	Adequate Team Size	1	Item adapted from Wageman, et al. (2005).[D]
	Formal Leadership	5	New items based on Gladstein, 1984; Cohen, 1993; Campion, et al., 1993; Hoegl and Parboteeah, 2006a; Yang, et al., 2011; Pereira, et al., 2017; Marsicano, et al., 2017; Ishak, et al., 2018.

We then used a 5-point Likert scale for the values of each response item, as shown in Figure 6.

[A] The items were adapted to refer directly to the context of software development teams. For example: ‘The members of my team vary widely in their areas of expertise’ (original item) to ‘My team members have diverse areas of expertise in software development’ (adapted item).

[B] The items were adapted to focus on the team. For example: 'Standards for member behavior in this team are vague and unclear.' (original item) to 'My team’s behavior patterns are vague and unclear' (adapted item).

[C] The items were adapted to suit the context of software development and the use of the Likert scale. Originally, items are answered via 'yes' or 'no'.

[D] The items were adapted to suit the context of software development and to use the Likert scale. Originally, items are calculated and generate a result on the appropriateness of team size.

In a scale of 1 to 5, with 1 meaning "Totally Disagree" and 5 meaning "Totally Agree", please mark the value that best express your level of agreement with the following affirmatives with respect to your current software development team.

1	2	3	4	5
▲	▲	▲	▲	▲
Totally Disagree				Totally Agree

Affirmatives	1	2	3	4	5
…	□	□	□	□	□
My team's goals are shared and accepted by all its members.	□	□	□	□	□
The roles and responsibilities of my team members are accepted by all.	□	□	□	□	□
…	□	□	□	□	□

Fig. 6 *Example response items*

Finally, we gave each item a unique identifier, e.g. INT1, CGR2, NOR3, MNG3, TEC2, which were used during execution of EFA and CFA. The TPA's initial set of response items is thus a mix of adapted items (16%), and new items (84%) (APPENDIX B).

Task 4: Translation and Analysis of Items

In our study, we needed to translate the items from English (the original language from the literature) to Brazilian Portuguese (the language used by our validation study subjects). Following the translation method proposed by Dias Júnior (2016), which specifically addresses the translation of instruments to Brazilian Portuguese, the response items were translated into Portuguese by three researchers: the first author and two senior researchers. All three are native Portuguese speakers but one of the senior researchers has also been residing and working in English speaking countries for several years. The three versions were then compared and integrated by the first and second authors. Differences in the translation were then resolved in a consensus meeting.

After the translation, we piloted the research instrument to identify potential problems with software engineering terminology, redundant items, and lack of clarity. The 23 subjects included 10 software engineers (four developers, three testers, one project manager, one scrum master, and one usability analyst); nine academic resarchers, and four graduate students (three doctoral and one master student). As a result, three redundant response items were identified and removed, and12 response items were revised to increase clarity and remove ambiguities.

Task 5: Structuring the Research Instrument

This task involves how to order and group the items, as well as the decision of whether the instrument will be administered online or on paper, or both. Although no changes are made to the content of the response items, this task is important to achieve a final structure that will assist the participants to provide consistent and complete answers to the questionnaire Brace (2018).

The research instrument was divided into three subsections. The first contained a brief description about the research and its objective. The second was composed of the participants' demographic information. The third contained all the response items resulting from the previous tasks. An online version of the instrument was built and used in the validation. No paper version was used.

3.2 Experimental Aspects

The experimental aspects deal with planning and execution of the data collection using the measurement instrument developed in Phases 1 and 2. Collected data is used as the input to the data analysis in the Analytical Aspects (Section 3.3)

Phase 3: Research Instrument Application

Task 6: Data Collection Planning

The goal is to define the target population, the sampling method, and the target sample size appropriate for the statistical tests that will be applied to the collected data. Our population is Brazilian software engineering professionals (team leaders and team members) working in industry (in Brazil or in other countries), in either new development or maintenance.

Because this was a validation study only (Kitchenham and Pfleeger, 2002), we used non-probabilistic sampling methods, in particular convenience and snowball sampling, with the aim of obtaining a heterogeneous sample of sufficient size. Heterogeneity refers to different roles represented (managers, leaders, developers, testers, etc.), from

teams of different sizes and approaches to software development (Agile, traditional, etc.), in different types of organizations (public and private), and locations (different Brazilian states or countries). Sample heterogeneity helps to improve the results when using factor analysis to explain and reduce data, and to confirm the factorial structure of the measurement instrument (Laros, 2012).

As for the sample size, we followed the recommendations of Pasquali (2010) and Cattell (2012), and set our goal for at least five subjects per item in the instrument, for a minimum of 250 participants.

Task 7: Research Instrument Application

This task included defining the period of data collection and how the research instrument would be distributed (Brace, 2018). Data collection was carried out in January and February 2019. The instrument was made available through a web address, accessible 24 hours a day, allowing all invitees to access it using computers, tablets, and smartphones. In the end, we collected data from 375 participants.

3.3 Analytical Aspects

The analytical aspects establish the statistical analysis procedures to be performed to check and improve instrument validity (Pasquali, 2010). The statistical procedures used in this research focused on evaluation of data adequacy for multivariate analysis (Task 8), exploratory factor analysis (EFA) (Task 9), and confirmatory factor analysis (CFA) (Task 10)[5].

Phase 4: Research Instrument Validation

Task 8: Evaluation of Data Adequacy for Multivariate Analysis

Before beginning a multivariate data analysis, one must identify the potential problems related to missing data, outliers, normality, and multicollinearity (Hair, et al., 2009; Kline, 2010; Marôco, 2010), as described below.

Missing Data and Outliers. According to Hair, et al. (2009), missing data occurs when valid values for one or more variables are not available for analysis. In practice, the missing data represent a reduction in the size of the sample available for analysis. We checked our data for incorrect[6] and/or incomplete[7] responses. In order to check for outliers, we used the Mahalanobis distance measure for the identification of extreme multivariate values (Marôco, 2010).

Normality. Because we used Maximum Likelihood (ML) for model estimation, it is necessary that the manifested variables present a normal multivariate distribution (Marôco, 2010). We verified normality using histograms with the Gaussian curve (normal curve) and the Kolmogorov-Smirnov and Shapiro-Wilk tests (Marôco, 2010). In order to evaluate the hypothesis of normal distribution of the variables, the measures of distribution form, skewness (sk), and kurtosis (ku) were used. The values of sk and ku should be below 3 and 10, respectively (Kline 2010).

Absence of Multicollinearity. Multicollinearity identifies if two or more variables measure the same phenomenon (Kline, 2010). We used two statistical procedures to test multicollinearity. The first is related to tolerance, indicating the proportion of total standardized variance that is unique (not explained by other variables). A tolerance value smaller than 0.10 may indicate extreme multivariate collinearity (redundancy) (Kline, 2010). The second is the variance inflation factor (VIF), which presents the ratio of total standardized variance over single variance (tolerance). VIF value smaller than 10.0 demonstrates that the variable is not redundant (Kline 2010).

Task 9: Exploratory Factor Analysis

If the results of Task 8 (see Section 4.1) are favorable, measure validation can proceed with exploratory factor analysis (EFA). EFA can be understood as a set of multivariate techniques that aims to find an underlying structure in a data matrix and to define the factors that best explain the covariance of a certain set of variables demonstrating what the instrument is measuring, as well as the items that compose each factor (Hair, et al., 2009; Pasquali, 2010; Damásio, 2012). EFA is usually conducted when the researcher does not yet have a prior theory or evidence that supports a grouped set of items (latent constructs). In this sense, EFA is an "exploratory" analysis because no restrictions are initially placed on the patterns of relations between the observed measures and the latent variables (Brown, 2006).

Factoring conditions. Before initiating the EFA, it is necessary to verify if the correlation matrix is factorable (Pasquali, 2010. First, we visually check that at least 50% of the correlations are greater than 0.30 (Pasquali, 2012). If this is true then two tests are performed:

[5] For the statistical analysis, we used IBM® SPSS® Statistics, version 25 and IBM® SPSS® Amos, version 25.

[6] Incorrect refers to the participation of people with profiles that are outside the scope of this research.

[7] Incomplete, refers to unanswered response items in the research instrument.

(1) Kaiser-Meyer-Olkin (KMO) sample adequacy test, for which values > 0.70 are good, > 0.80 are great, and > 0.90 are excellent (Hair, et al., 2009; Pasquali, 2012; Osborne, et al., 2014);
(2) Bartlett's sphericity test that evaluates to what extent the covariance matrix is an identity matrix. When this hypothesis is rejected ($p < 0.05$), the matrix is factorable. (Figueiredo Filho and Silva Júnior, 2010; Pasquali, 2012; Osborne et al., 2014).

Pasquali (2012) emphasizes that this matrix factorization analysis should be performed using Principal Components Analysis (PCA) and not Factor Analysis, because a variance-covariance matrix is needed with 1 in the main diagonal.

If the correlation matrix is factorable, then the following procedures are performed, until a final factor structure is obtained.

Factor Retention. This procedure establishes the number of factors that will remain during the EFA (Laros, 2012). The goal is to minimize both overestimation and underestimation of the retained factors (Laros, 2012; Damásio, 2012). For this, two criteria are used in a complementary way: (1) Kaiser-Guttman criterion (eigenvalue > 1.00), where the eigenvalue corresponds to the amount of variance explained by a factor (Pasquali, 2012), and (2) Cattell's scree test, which is a graphical representation of eigenvalues (Laros, 2012; Pasquali, 2012; Osborne et al., 2014). We used a threshold of 50% of cumulative explained variance (Marôco 2018).

Extraction and Rotation of Factors. We used Principal Factor Analysis (PFA) because it is one of the main methods of extraction (Pasquali, 2012) and offers the best results (Damásio, 2012). This method assumes that each variable consists of a part common to the factorial structure and a specific part of the variable. As reported by Damásio (2012) factorial rotations aim to facilitate the interpretation of the factors, with the objective of achieving a simple factorial structure, where each variable presents a high factorial load on only one factor (Laros, 2012; Osborne, et al., 2014). Since a correlation between the factors is assumed, an oblique rotation is more adequate (Nascimento, 2014). One of the most appropriate rotation methods is the PROMAX rotation (Damásio, 2012).

Interpretation of the Factor Matrix. Interpretation is based on the factorial load and the commonalities (Hair, et al., 2009). The factor load evaluates the correlation between initial variables and the factors, allowing observation of the factorial contribution of the variables on the retained factors. Factor load values greater than 0.40 can be considered to have a practical significance (Hair, et al., 2009). Otherwise, the variable is considered statistically independent, and should be removed (Hair, et al., 2009). Also, the items that saturate in more than one factor, whose difference is less than 0.10, are candidates to be excluded from the analysis (Neiva, et al., 2007). The commonalities represent the amount of variance explained by the factorial solution for each variable (Hair, et al., 2009; Damásio, 2012). Variables with commonality lower than 0.30 are candidates for elimination, because they do not contribute to the construction of the factor (Figueiredo Filho and Silva Júnior 2010; Nascimento, 2014).

Reliability of the measurement instrument. This procedure verifies whether the items that make up a scale reflect the construct that they are intended to measure (Field, 2012). Cronbach's Alpha (α) is the most commonly used method for evaluating internal reliability (Damásio, 2012). George and Mallery (2003) present $\alpha > 0.90$ as excellent, $\alpha > 0.80$ as good, $\alpha > 0.70$ as acceptable, $\alpha > 0.60$ as questionable, $\alpha > 0.50$ as poor, and $\alpha < 0.50$ as unacceptable. Hair, et al. (2009) state that in exploratory studies an α around 0.60 is acceptable.

Naming Factors. After reaching a satisfactory factorial solution, it may be necessary to name and/or rename factors (Hair et al., 2009). This involves identifying factors with the greatest significance and interpreting their meaning. This process is somewhat subjective and different researchers might assign different names due to their experiences (Hair et al., 2009). In order to minimize possible divergences regarding the naming of factors, Hair et al. (2009) suggest that a logical name be designated that represents the latent nature of the factors.

Higher-Order Factor Analysis in EFA. At this point, it may be necessary to advance in the analysis when the 1st order factors are highly correlated (> 0.50) (Laros, 2012; Tabachnick and Fidell, 2001). Higher-order (or hierarchical) factor analysis is used to group 1st order factors into 2nd or 3rd order factors (Brown, 2006; Laros, 2012; Osborne, et al., 2014). This involves examining the factor correlation matrix to identify possible higher-order groupings of factors.

Task 10: Confirmatory Factor Analysis – CFA

Confirmatory Factor Analysis (CFA) is a type of Structural Equation Modeling (SEM) (Raykov, 2012) that deals specifically with measurement models (Brown, 2006). CFA is a way to test how well the measured variables represent a smaller number of constructs (Hair et al., 2009). It is commonly used in the process of developing a scale to examine the latent structure of a research instrument (e.g., a questionnaire) (Brown, 2006).

CFA provides a more stringent test of the underlying factor structure for a survey instrument than exploratory

factor analysis (Avoid, et al., 1999). Further, CFA provides a more rigorous test of construct validity than multi-trait/multi-methods approaches (Spreitzer, 1995). When a CFA model is adjusted and demonstrates construct validity, the measurement theory[8] is supported (Hair et al., 2009). The steps and corresponding tests in the CFA are described below.

Model Specification. This consists of the formal design of the theoretical model in which the following decisions are taken: (1) which manifest variables operationalize which latent variables, (2) which causal relationships between latent variables and/or manifest variables should be included/excluded, (3) which (non-casual) associations should be included/omitted from the model, and (4) what errors, or residues, should be correlated (Marôco, 2010).

Model Estimation. This procedure aims at obtaining estimates of the model's parameters that are capable of reproducing the data observed in the sample in the best possible way (Marôco, 2010), taking into account:

i. *The data matrix:* the variance-covariance matrix will be used to maximize the probability of observing the correlational structure of the manifested variables found in the sample.
ii. *The sample size:* the common criteria that establishes between 5 to 10 subjects per item of the instrument and a sample of over 250 subjects (Pasquali, 2010; Cattell, 2012);
iii. *The estimation method:* the Maximum Likelihood (ML) is the traditional method used in SEM (Hair, et al., 2009; Marôco, 2010).

Evaluation of the quality of the model adjustment. This procedure aims to evaluate how well a theoretical model can reproduce a correlational structure of manifest variables in the study sample (Marôco, 2010). In other words, the perception of how well a theory adjusts to the data (Hair, et al., 2009). The Goodness of Fit (GOF) model is obtained through the use of indices classified as: absolute, relative, parsimony, population discrepancy, and information theory (APPENDIX C, Table 23). Although dozens of model quality indices can be found in the literature, it is unusual to report all of them since they are often redundant (Marôco, 2010). Table 24 (APPENDIX C) shows the quality indices we understood as adequate and sufficient for the evaluation of model adjustment in this work (Brown, 2006; Hair, et al., 2009; Marôco, 2010; Kline, 2010).

Evaluation and Adjustment of the Model. If the adjustment adequacy indices are consistent with a good fit model, this will provide initial support for the correctness of the model specification (Brown, 2006). Three statistics are frequently used complementarily to verify the "localized areas of strain":

- *Standardized residuals* present the absolute values on how each variance and covariance were reproduced by the model parameter estimates, for each pair of indicators (items) (Brown, 2006). The values should be lower than 2.58, otherwise, they may indicate adjustment problems (Byrne, 2010). In this case, one or both matrix items should be eliminated in order to correct the model (MacCallum, 1986).
- *Modification Index (MI)* reflects the approximation of how much the general model χ^2 (Chi-square) would decrease if the fixed or restricted parameter were freely estimated (Brown, 2006). This analysis must be done sequentially, starting with those that have high values of MI (> 11) in more than one item (Marôco, 2010).
- *Critical Ratio* (C.R.) represents the estimated parameter divided by the standard error, which should be greater than 1.96 for the parameter to be considered significant, based on a probability of 0.50 (Byrne, 2010). Variables that do not meet this criterion should be taken out of the model (Byrne, 2010; Marôco, 2010).

It should be noted that the evaluation and adjustment of the model should take into consideration not only statistical issues, but must also observe the theoretical and conceptual context that supports the model (Brown 2006; Hair, et al., 2009; Marôco 2010). Finally, regarding the validation of the adjusted model, Marôco (2010) states that, when there is no possibility of collecting a new sample, the ECVI index can be used. Thus, the model that presents the lowest ECVI value will be the most stable in the population (Browne and Cudeck, 1989).

Reliability and Validity of the Measurement Instrument. The indices and conditions to verify internal reliability and construct validity (convergent and discriminant) used in our work are presented in APPENDIX C (Table 25).

3.4 Ethics

We followed the norms of Resolution 466/12 – CNS-MS of the Brazilian National Health Council that regulates research with human subjects. The norm establishes the ethical principles that must be followed to avoid harm and increase the benefits of the research to participants. It also covers aspects related to voluntary participation, confidentiality, and the right to withdraw from the research.

An Informed Consent Form was provided in the introduction of the research instrument that explained the overall

[8] The measurement theory specifies a series of relationships that suggest as variables, measurements that represent a latent construct that was not directly measured (Hair, et al., 2009).

objective and relevance of the research, guaranteed data confidentiality, the anonymity of the participation, the non-obligatory nature of the participation, and the right to withdraw from the research at any moment. All participants accepted the terms of the consent form and, thus, freely agreed to participate. No participant withdrew from the research.

The research instrument did not collect personal information that would allow a participant to be identified. Therefore, the researchers did not have access to the source of the information, maintaining full participant confidentiality.

4 Results

In this section, we present the results of the Analytical Aspect described in Section 3.3. We start with the characterization of the data sample and then present the results of the EFA and CFA.

Using the procedures for evaluation and adjustments of the data (Section 3.3, Task 8), the initial set of 375 records was reduced to 326 valid records: 17 incomplete records, 8 non software development professionals, and 24 outliers were removed. The descriptive statistics of these 326 participants is presented in Table 7 through Table 10.

Table 7. Descriptive Statistics of Participants.

Description		Total	Percentage	
Total valid cases		326	100%	
Total Male		240	73,62%	
Total Female		86	26,38%	
Item	**Min.**	**Max.**	**M.**	**S.D.**
General Age	19	62	41	8,18
Male Age	19	59	39	8,18
Female Age	21	62	42	8,18
Experience in Software Engineering (years)	0,50	37,00	11,95	7,36
Organization Experience Time (years)	0,10	35,00	5,44	5,92

Legend: Min: Minimum; Max: Maximum; M: Mean; SD: Standard Deviation

Table 8. Characteristics of Participants in Team.

Description	Total	Percentage
Role		
Team Member	194	59,51%
Leadership	132	40,49%
Dedication (Member or Leader) in Team		
Full-time	285	87,42%
Part-time	41	12,58%
SE Approach		
Traditional	16	4,91%
Agile	195	59,82%
Mix	115	35,28%

Table 9. Team Characteristics.

Item	Min.	Max.	M.	S.D.
How long the participant has been on the team? (years)	0,10	16,00	2,63	2,57
How long ago was the team created? (years)	0,10	25,00	4,47	4,69
Number of people in the team	3,00	100,00	11,00	10,90

Legend: Min: Minimum; Max: Maximum; M: Mean; SD: Standard Deviation

Table 10. Range of Team Size.

Range of Team Size	Total	Percentage
3 to 8 people	178	54.60%
9 to 16 people	95	29.14%
17 to 30 people	38	11,66%
> 30 people	15	4.60%

Contributing to the heterogeneity of the sample, participants worked in diverse organizational contexts: 215 participants were in the private sector (65.95%), 53 in public companies (16.26%), 11 in mixed-economy companies (3.37%), 36 were in the public service (11.04%), and 11 declared themselves as being in other types of organizations (3.37%). These organizations develop software for clients as their core business (58.28%) and for themselves (41.72%). Finally, 317 research participants (97.24%) were working in Brazil at the time of the survey, with representation in all regions of the country. The other 9 participants (2.76%) were Brazilians who were working in other countries (Angola, Australia, Austria, Canada, United States, Portugal, and Sweden).

We tested normality of the final data set and consider that the data were within the limits indicated by Kline (2010) ($sk < 3$ and $ku < 10$). Thus, the results demonstrated a satisfactory data quasi-normality for the subsequent factor analysis. Finally, the results obtained regarding the absence of multicollinearity were within the parameters reported by Kline (2010), characterizing them as non-redundant. Therefore, the final data set was considered adequate for multivariate analysis.

4.1 Exploratory Factor Analysis – EFA

This section presents the results of Exploratory Factor Analysis (EFA), following the procedures described in Section 3.3, Task 9. The EFA was initiated with 51 items. The factorization conditions were verified and showed that more than 50% of the correlations between the variables are greater than 0.30. Further, an excellent KMO value (0.930) and statistically significant Bartlett value ($p < 0.000$) were observed. As a result, the factorization conditions were met.

Then, with the application of the Kaiser-Guttman criterion (eigenvalue > 1), up to 11 factors to be retained were identified, with an explanatory power of 63.188%. Using the scree graph, up to 11 factors were also identified. Table 11 and Figure 7 present these results.

Table 11. Application of Kaiser-Guttman criterion.

Total Variance Explained

Factor	Initial Eigenvalues		
	Total	Variance %	Cumulative %
1	15.867	31.112	31.112
...	...	...	...
11	1.099	2.154	63.188
12	0.953	1.869	65.057
...	...	...	...
51	0.145	0.285	100

Method: Principal Component Analysis.

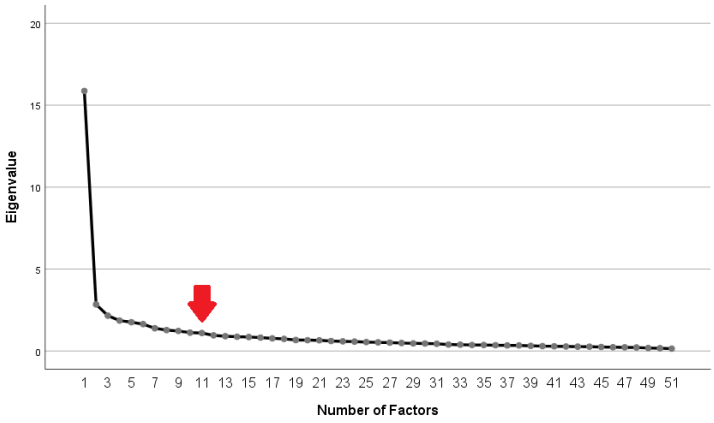

Fig. 7 Scree Graph

Considering these results, the extraction was performed by means of the Principal Axis Factoring (PAF), the OBLIQUE rotation, using the PROMAX method, for the extraction of up to 11 factors. Six rotations were performed, until only the statistically and conceptually significant items and factors remained. In such rotations, factorial loads, commonalities, and item saturations were verified in more than one factor. Items INT4, INT11, MNG3, LEAD3, TEW1, TEC1, MNG4, MNG5 and MNG6 were discarded because they presented a factorial load inferior to 0.40. Items TEAM1, TSC1, TEW2, TEW3, TEW5, TEC2 and EOR3 were also removed, because they presented a commonality inferior to 0.30, not contributing to the factor construction. As for saturation, there were no items removed.

Thus, as a result of the 6th rotation, a KMO of 0.919 and a Bartlett with $p < 0{,}000$ were verified. Regarding the factor retention, both the Kaiser-Guttman criterion and the Scree Graph indicated 9 factors, with an explanatory power of 67.454%. Therefore, it was observed that, in comparison with the initial analysis, the last rotation improved the model's explanation power, as well as reduced the number of items to 35 and of factors to 9, also increasing its significance.

Reliability of the Measurement Scale

Cronbach's Alpha for the 35-item scale was 0.93, considered excellent (George and Mallery, 2003). The reliability of the scale of all but one latent construct was between 0.74 and 0.87. One of the constructs had a result of 0.68 but is still considered acceptable for exploratory studies (Hair, et al., 2009).

Considering this, it is important to emphasize that the EFA results retained items and factors that presented greater statistical and theoretical compatibility.

Table 12 presents the EFA results, containing the code for each item, its associated factor and commonality, as well as the number of items, Cronbach's Alpha value and the factorial loads for each factor.

Table 12. Matrix Factor of the Measurement Scale.

Item	1	2	3	4	5	6	7	8	9	Commonality
INT1	0.50									0.54
INT2	0.86									0.62
INT3	0.67									0.58
INT5	0.91									0.65
INT6	0.54									0.51
INT7	0.57									0.40
INT8	0.64									0.36
INT9	0.44									0.44
INT10	0.55									0.54
CGR1		0.75								0.58
CGR2		0.70								0.66
CGR4		0.50								0.57
NOR1		0.53								0.61
LEAD1			0.86							0.77
LEAD2			0.90							0.77
LEAD4			0.47							0.48
LEAD5			0.58							0.55
TEC3				0.80						0.60
TEC4				0.87						0.67
TEC5				0.43						0.44
TEC6				0.68						0.58
TEC7				0.45						0.52
HET1					0.80					0.57
HET2					0.79					0.60
HET3					0.51					0.47
CGR3						0.60				0.49
TEW4						0.46				0.50
NOR2						0.57				0.47
TEC8						0.65				0.40
MNG1							0.83			0.73
MNG2							0.77			0.70
NOR3								0.66		0.58
NOR4								0.77		0.73
EOR1									0.67	0.48
EOR2									0.86	0.74
Nº of Items	9	4	4	5	3	4	2	2	2	
Cronbach's Alpha	0.87	0.84	0.86	0.82	0.75	0.68	0.84	0.82	0.74	
Min. FL	0.44	0.49	0.47	0.43	0.51	0.46	0.77	0.66	0.67	
Max. FL	0.91	0.75	0.89	0.87	0.80	0.65	0.83	0.77	0.86	
Mea. FL	0.63	0.62	0.70	0.64	0.70	0.57	0.80	0.71	0.77	

Legend: FL – Factorial Load; Mín. – Minimum; Max. – Maximum; Mea. – Means.

Presentation of the Factors Identified in the EFA

It is possible to observe that the quantity of factors, initially proposed in Section 3.1, remained the same. However, there was a redistribution and decrease of response items. In addition, not all of the proposed factors remained. The factors that are no longer present in the measurement scale are: the team size (adequacy of the team size in terms of the work to be executed), and task control (team's authority over processes and tasks). A justification for this may be associated with the fact that these factors had only one evaluation item linked to them, which may have made such measurements fragile. During the EFA rotations these items were grouped with other items.

Next, each one of the factors introduced in the measurement scale is presented with its proper nomenclature,

description, and associated comments (Table 13). It should be noted that, whenever there was a need to rename factors the guidelines proposed by Hair, et al. (2009) were followed.

Table 13. Factors and Descriptions.

Factor ID	Factor Name	Description
Factor 1 (F1)	Interpersonal Skills	Initially, this factor was a part of the latent construct 'team skills'. The EFA results split skills into three factors. Thus, F1 became 'interpersonal skills', and it concerns a set of team skills within the framework of human relations established among its members, regarding: collaborating for the accomplishment of the activities, constructing relationships of respect and trust, feeling a part of the team, willingness to ask, offer and receive help from one another, conducting internal feedback, establishing conversations (work and personal) openly, separating technical and personal discussions, and managing conflicts among team members.
Factor 2 (F2)	Role and Goal Clarity	Remains with the original name, and referring to the establishment, sharing, and acceptance of the work goals, roles and responsibilities, by team members. In addition, an item was added to this factor, in which teamwork process must be established and aligned to team's needs. The analysis reveals that the creation of roles and responsibilities is closely linked to the work process, which must be established in alignment with the needs of the team, that is, the fulfillment of the team's goals.
Factor 3 (F3)	Formal Leadership	Keeps almost in its entirety the set of initially established items. This factor refers to the behavior of the team leader (or manager), in terms of giving freedom and encouraging the team to make its own decisions about the work, stimulating the team to work autonomously, being a facilitator in the decision-making and in the work execution, and encouraging the team to solve its own problems.
Factor 4 (F4)	Technical Skills	Initially this factor was included in the latent construct 'team skills'. After the EFA, it became a new grouping, which refers to team members possessing a set of knowledge about technologies and patterns (coding, architecture, etc.), methods, practices and tools required to carry out the work, and enabling the team to understand and to establish technical solutions, in an appropriate manner for the client and the project, as well as resolving technical problems inherent in software development.
Factor 5 (F5)	Heterogeneity	Remains as 'heterogeneity', and maintains all initial set of items. This factor refers to the team's degree of heterogeneity in terms of knowledge, skills, attitudes, and experiences.
Factor 6 (F6)	Team Maturity	These items are related to the experience of the team in working more autonomously, coupled with the clarity in the definition of roles and responsibilities of the team members, and clarity regarding behavior standards that should be observed by team members. These items align with some characteristics of software development team maturity proposed by Marsicano, et al. (2017). For this reason, Factor 6 was renamed.
Factor 7 (F7)	Management Skills	Items in this factor were initially included in 'team skills', later becoming a new grouping. This factor is related to the team's ability to maintain visibility on the progress of their work for all stakeholders (for example, manager, clients, etc.) and to monitor their work continuously.
Factor 8 (F8)	Rules of Behavior	Initially referred to as 'specific work norms', this factor was renamed. The items that compose this factor refer to the clarity about what is and what is not an acceptable behavior for team members as well as to the agreement on how they should behave.
Factor 9 (F9)	Experience in the Organization	Remains as 'experience in the organization', and maintains all initial response items. This factor denotes that team members have work experience in the organization in which they work, operating together or on different teams.

Matrix Factor of Higher Order (2nd Order)

As a final step of the EFA, the factor correlation matrix was verified in order to identify possible groupings of 1st order factors into higher-order factors (Laros, 2012). The results of this verification may contribute to the maintenance of the factor distribution identified by Gladstein (1984), as well as new groupings, or none at all.

For this procedure, the factor matrix (1st order) generated in the 6th EFA rotation was analyzed, which established the set of 35 items distributed in 9 factors. The matrix (Table 14) indicates that some of the factors have significant correlations (> 0.50) (Tabachnick and Fidell, 2001).

Table 14. Matrix of Factor Correlations (6th rotation of EFA).

Factor	F1	F2	F3	F4	F5	F6	F7	F8	F9
Interpersonal Skills (F1)	1.00								
Role and Goal Clarity (F2)	0.51	1.00							
Formal Leadership (F3)	0.60	0.49	1.00						
Technical Skills (F4)	0.64	0.39	0.49	1.00					
Heterogeneity (F5)	0.49	0.31	0.40	0.48	1.00				
Team Maturity (F6)	0.42	0.42	0.44	0.44	0.28	1.00			
Management Skills (F7)	0.59	0.45	0.50	0.53	0.27	0.36	1.00		
Rules of Behavior (F8)	0.57	0.46	0.44	0.49	0.42	0.33	0.44	1.00	
Experience in the Organization (F9)	0.24	0.19	0.21	0.34	0.31	0.17	0.24	0.21	1.00

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.

Based on this matrix, the syntax presented by Wolff and Preising (2005) was executed in the IBM® SPSS® Statistics, version 25 tool to generate 2nd order matrices, in order to investigate the possibility of factor clusters. Thus, three 2nd order matrices were generated: (1) with only one 2nd order factor, (2) with two 2nd order factors and (3) with three 2nd order factors. Theses matrices and our observations are presented in APPENDIX D, Table 26, Table *27* and

Table *28*, respectively. The results provided statistical support for the verification in the CFA of two models of second-order: one factor and two factors of higher-order (matrices (1) and (2)).

4.2 Confirmatory Factor Analysis – CFA

Confirmatory factor analysis (CFA) of the resulting measurement scale was performed to verify the adequacy of the relation between items and factors, as well as the convergent and discriminant validity of the constructs, following the procedures described in Section 3.3, Task 10. In total, four models were analyzed:

- Model 1: composed of the 9 identified factors in the EFA and their relationships with each other (only 1st order factors);
- Model 2: presenting the relation of the 9 identified factors in the EFA with a factor of 2nd order;
- Model 3: grouping the 1st order factors into two higher order factors; and
- Model 4: using the structure based on Gladstein (1984) with two 2nd order constructs (team structure and team composition) (see Figure 5).

In order for the higher-order model to be identified, it is necessary to first identify and validate the 1st order model Marôco (2010), so we started our CFA with Model 1.

Model 1: Measurement Model Adjustment

The CFA of Model 1 presented indices that are close to satisfactory. However, the values of GFI and CFI did not meet the established parameters (see Table 15). In this type of situation, with χ^2 / df > 2.0, and GFI and CFI below the reference values, Marôco (2010) recommends to observe, primarily, the modification indices. In addition, the residues and the values of C.R. must be verified (Brown, 2006; Byrne, 2010).

Therefore, it is necessary to proceed with the model adjustment towards its improvement (Marôco, 2010). To do this, items (CGR3, CGR1, INT1) were removed since they presented modification indices above 11 in more than one item, as well as items with standardized residual covariance above 2.58 (NOR2, INT8, TEC3, INT7, INT9). Thus, a very significant adjustment of Model 1 was reached, being considered excellent (Table 15).

Table 15. Quality Indices of Model 1: before and after adjusts.

Sample N = 326	**Indices**	**Model 1** *(before adjusts)*	**Model 1** *(after adjusts)*
Absolute Indices	χ^2 /df	2.132	1.584
	SRMR	0.060	0.044
	GFI	0.828	0.906
Relative Indices	CFI	0.890	0.958
Parsimony Indices	PGFI	0.690	0.688
	PCFI	0.785	0.783
Population Discrepancy Indices	RMSEA	0.059	0.042
Information Theory Indices	ECVI	4.090	1.959

Model 1: Reliability and Validity of the Measurement Scale

To verify the reliability of the measurement scale, Jöreskog's Rho (ρ) was used in the CFA. Jöreskog's ρ ranged between 0.70 and 0.87, with the exception of 'work experience', which obtained a value equal to 0.65. However, it was decided to keep this factor due to its value being close to satisfactory, as well as this being an exploratory study, which creates space for future research that can improve this factor (Hair, et al., 2009; Marôco, 2010; Pasquali, 2012).

Convergent validity, which aims to evaluate the degree to which two measures of the same concept are correlated, was verified using Jöreskog's Rho of convergent validity (Rho_{cv}). All factors presented $Rho_{cv} > 0.50$, ranging from 0.50 to 0.72, verifying an adequate convergent validity (Marôco, 2010).

In order to verify the discriminant validity, that is, the degree to which a construct is truly different from the others (Hair, et al., 2009), two complementary methods were used. In the first one, it was verified that the Rho_{cv} is superior to the square of the estimate between two factors in all but one case. The second method was performed using the χ^2 difference test. This ensured that the total discriminant validity of the model was confirmed[9].

APPENDIX E shows the Rho (ρ) of reliability and convergence of each factor, as well as the factorial loads of each of the items that make up the scale (Table 29) and other values that show the high significance of all relations between factors and items (Table 30). Based on these results, the construct reliability and validity (convergent and discriminant) of the scale of measurement established by Model 1 was confirmed. This model was named The Model of Teamwork Process Antecedents (TPA) Questionnaire and is graphically represented in Figure 8.

[9] $\alpha = 0.05$, $(glr - glu) = 1$, $\chi^2_{dif} = 40.376$ and $\chi^{20,95\,(1)} = 3.84$. The expression $\chi^2_{dif} > \chi^2_{1-\alpha}\,(glr - glu)$, is confirmed, rejecting the null hypothesis: $\chi^2_u = \chi^2_r$.

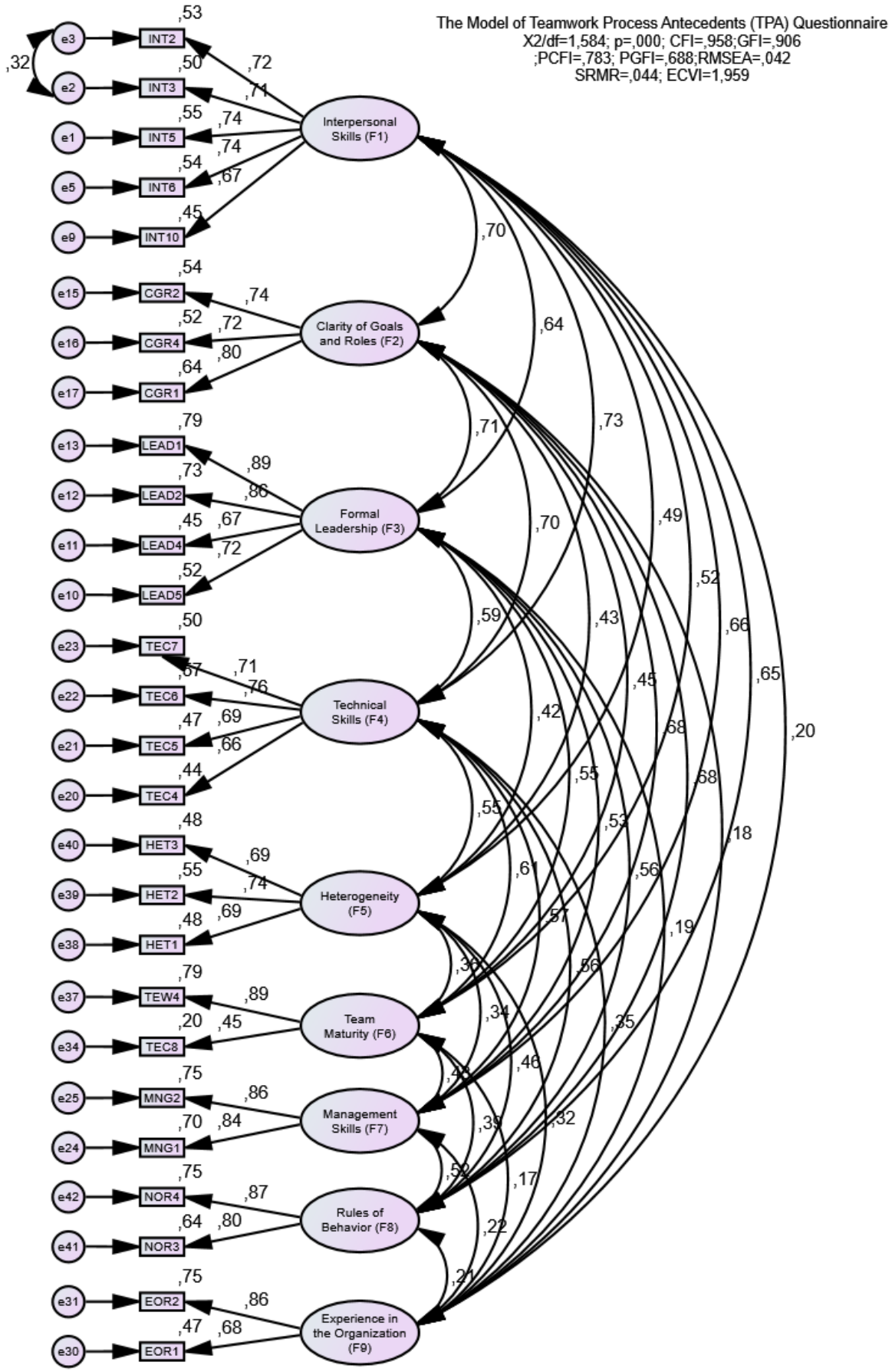

Fig. 8 The Model of Teamwork Process Antecedents (TPA) Questionnaire

Higher-Order Model – CFA

After identifying and validating the 1st order model, we proceeded to verify the higher-order models. Higher-order factors (2nd or 3rd order) in the CFA are usually indicated by (i) considerable correlations between 1st order factors, (ii) correlations between errors of items that saturate in different factors and/or (iii) theoretical justification (Marôco, 2010). A higher-order factor must be composed of at least three lower order factors as indicators (Hair et al., 2009 and Kline, 2010). If not, the direct effects of the 2nd order factor on the 1st order factors may be under-identified (Kline, 2010). In addition, each 1st order factor must have at least two indicators (Kline, 2010).

Our data meet these criteria. Both the EFA (see Table 12 and Table 14) and the CFA results (see Table 16) show significant (>0.50) correlations between factors, with the required number of 1st order factors as indicators. Table 17 shows the results of evaluating the quality of the original Model 1, as well as 3 higher-order models.

Table 16. Matrix of Correlations between Factors - Model 1 (CFA).

Factor	F1	F2	F3	F4	F5	F6	F7	F8	F9
Interpersonal skills (F1)	1.00								
Role and Goal Clarity (F2)	0.70	1.00							
Formal Leadership (F3)	0.64	0.71	1.00						
Technical Skills (F4)	0.74	0.70	0.59	1.00					
Heterogeneity (F5)	0.49	0.43	0.42	0.55	1.00				
Team Maturity (F6)	0.52	0.45	0.55	0.61	0.36	1.00			
Management Skills (F7)	0.66	0.68	0.53	0.57	0.35	0.44	1.00		
Rules of Behavior (F8)	0.65	0.68	0.56	0.56	0.46	0.39	0.52	1.00	
Experience in the Organization (F9)	0.20	0.18	0.19	0.35	0.32	0.17	0.22	0.21	1.00

Table 17. Comparative: Quality Indices of Model 1, Model 2, Model 3 and Model 4.

Sample N = 326	Indices	Model 1	Model 2	Model 3	Model 4
Absolute Indices	χ^2 /df	1.584	1.648	1.615	1.625
	SRMR	0.044	0.051	0.049	0.050
	GFI	0.906	0.892	0.895	0.894
Relative Indices	CFI	0.958	0.949	0.952	0.951
Parsimony Indices	PGFI	0.688	0.741	0.741	0.740
	PCFI	0.783	0.849	0.849	0.848
Population Discrepancy Indices	RMSEA	0.042	0.045	0.044	0.044
Information Theory Indices	ECVI	1.959	1.986	1.956	1.965

Table 17 shows that the three models (2 to 4) presented indices very close to satisfactory and to each other. However, they did not meet the established parameters for the absolute Goodness of Fit Indices (GFI) indices. In addition, in Model 4, it is important to note that the covariance between the two 2nd order constructs is 0.93, strongly indicating the existence of a single construct.

Model 1, which has the best absolute, relative, and population discrepancy rates, is also the only one that met the parameters of GFI. Models 2, 3 and 4 obtained parsimony values equal or better than those of Model 1. In general, higher-order models are more parsimonious because they consume fewer degrees of freedom (Hair, et al., 2009). Models 1 and 3 are borderline on the Information Theory indices, although Model 3 has the best value.

Finally, in view of the EFA and CFA results, as well as based on the theory used and the objectives of this work (analyze the influence of each 1st level construct on teamwork process), Model 1 is the best representation of the measurement scale. In addition, taking into account all indices of model quality, internal reliability, convergent validity, and discriminant validity, Model 1 proved to be the most appropriate.

4.3 Summary of the Results

We have described the results of the development and validation of a comprehensive and integrative measure, using the existing literature and adapting or scaling it to measure a set of teamwork process antecedents in software development teams. We started with 51 response items built from previous work and from literature reviews, operationalizing nine 1st order factors. We applied the initial questionnaire and obtained 375 answers, from which we extracted 326 valid data records.

We used exploratory and confirmatory factor analysis (EFA and CFA) to refine the factor structure and the response items. EFA resulted in the retention of nine factors operationalized by 35 response items. The CFA confirmed the nine factors, although indicating a decrease of items to 27. The results of the CFA and the psychometric analyses performed support the internal reliability and construct validity (convergent and discriminant) of the measurement scale. Finally, the possibility of higher-order constructs was verified. However, where this possibility was indicated, there was a greater statistical and theoretical support for keeping a set of 1st order constructs.

Therefore, the construction and validation process of TPA produced a measure of nine constructs using 27 response items that can be found in Table 18.

Table 18. Constructs and response items of TPA.

Team Composition		
Factor	**Items**	**Description**
Interpersonal Skills (F1)	INT2	My team can establish relationships of respect and trust among its members.
	INT3	My team members feel that they are, in fact, part of the team.
	INT5	The members of my team are always willing to ask, offer and receive help, voluntarily, among themselves.
	INT6	My team is willing to give and receive feedbacks consistently among themselves.
	INT10	My team is capable to solve interpersonal conflicts among its members.
Technical Skills (F4)	TEC4	My team has the necessary knowledge about the methods, practices and tools used to accomplish their activities.
	TEC5	My team can establish technical solutions without the need for direct interference from the manager and the software architect.
	TEC6	My team can understand and give a solution to a technical problem in a manner appropriate to the client and the project.
	TEC7	My team can solve the technical problems inherent in software development with the participation of all its members.
Management Skills (F7)	MNG1	My team can keep the visibility on the progress of their work to all stakeholders (for example: manager, clients, etc.).
	MNG2	My team can follow and monitor their work continuously.
Heterogeneity (F5)	HET1	My team members have diverse areas of expertise in software development.
	HET2	My team members have different backgrounds and software development experiences.
	HET3	My team members have competencies that complement each other.
Team Maturity (F6)	TEW4	My team members lack experience in working in a more autonomous way, without the interference of the manager and the software architect, for example.
	TEC8	My staff needs constant interference from the manager and the software architect to be able to perform their activities.
Experience in the Organization (F9)	EOR1	My team members have already worked together on other software development projects in the Organization where we work.
	EOR2	My team members have already participated in other software development projects, in other teams, in the Organization where we work.
Team Composition		
Factor	**Items**	**Description**
Role and Goal Clarity (F2)	CGR2	My team's goals are shared and accepted by all its members.
	CGR4	The roles and responsibilities of my team members are accepted by all.
	NOR1	My team's work process is established and aligned with its needs
Formal Leadership (F3)	LEAD1	The manager (or immediate manager) gives freedom and encourages my team to make their own decisions about work.
	LEAD2	The manager (or immediate manager) stimulates my team to work autonomously.
	LEAD4	The manager (or immediate manager) plays a facilitating role in the decision-making and in the execution of my team's work.
	LEAD5	The manager (or immediate manager) encourages my team to solve their own problems.

Team Composition		
Factor	**Items**	**Description**
Rules of Behavior (F8)	NOR3	It is clear what is and what is not an acceptable behavior for the members of my team.
	NOR4	My team members agree on how they should behave.

5 Discussion

In this section, we start by discussing how to interpret TPA results (Section 5.1). After this, we present the main limitations of our study and possible ways to improve it in future research (Section 5.2). We follow with a discussion of the implications of our results. First, we discuss the implications and potential uses of the TPA questionnaire in empirical studies in software engineering, related to the study of software teams (Section 5.3). We then discuss potential applications of the TPA questionnaire in software engineering practice (Section 5.4) which could assist managers, human resources personnel, and team members to better understand and develop their software teams. Finally, we discuss the importance of using a consistent method for measure development and validation in empirical software engineering research, such as the one presented in this work (Section 5.5).

5.1 Interpretation of TPA Results

Consistent with our use of the IPO framework, TPA is a measure of a set of inputs to team processes. The contention is that, in the presence of these inputs, the team will be able to perform processes at or above desirable levels, contributing to the achievement of team goals. TPA measures the perception of each team member about the adequacy of the process inputs for the team to perform. For instance, Table 19 shows the response items about "Technical Skills" and "Heterogeneity".

Table 19. Examples of Antecedents and Response Items.

Antecedent	Item #	Response Item
Technical Skills (F4)	TEC4	My team has the necessary knowledge about the methods, practices and tools used to accomplish their activities.
	TEC5	My team can establish technical solutions without the need for direct interference from the manager and the software architect.
	TEC6	My team can understand and give a solution to a technical problem in a manner appropriate to the client and the project.
	TEC7	My team can solve the technical problems inherent in software development with the participation of all its members.
Heterogeneity (F5)	HET1	My team members have diverse areas of expertise in software development.
	HET2	My team members have different backgrounds and software development experiences.
	HET3	My team members have competencies that complement each other.

Each response item is operationalized by a Likert attitude scale, ranging from "Totally Agree" to "Totally Disagree". In this study, we used a 5-point scale. Thus, an individual score of "Technical Skill" reflects the perception of the individual about the adequacy of the technical skills of all team members with respect to the tasks and problems at hand. For instance, a low score for "Technical Skills" means that the individual perceives her team as lacking the adequate technical skills for its tasks. The "Heterogeneity" score reflects the individual perception of how diverse and complementary are team members expertise and competencies.

The internal reliability of the scale supports the aggregation of each response item to the construct level (e.g., using the median of the scores of the response items). Thus, TPA produces scores for each of the nine first level constructs. As we shall discuss in Sections 5.3 and 5.4, TPA can be used in empirical studies and also to assist the management of software teams in practice.

5.2 Limitations and Future Work

The first important limitation is that we only validated the Portuguese version of the questionnaire. We started with a set of response items in English, translated the items into Portuguese, and applied the translated version to collect data. Although we are confident that the translation process was performed following widely accepted standards, yielding semantically equivalent response items, cultural aspects may change the understanding of the response items. Therefore, researchers using the English version of the questionnaire may wish to follow our validation method on a smaller scale study and validate the English response items as well. Further, researchers wishing to develop a version of TPA for any language other than Portuguese or English may also use our process to develop and validate the new instrument. Following this process will increase the comparability of results contributing to replication and synthesis of studies using the TPA measure.

The second limitation is related to the completeness of TPA regarding the constructs used to model team structure and composition. Great effort was made to create a comprehensive and integrative measure of structure and composition, including using results from an extensive qualitative study and searching the general literature on teamwork for constructs and their operational definition. However, we cannot assure that TPA exhausts the full set of relevant factors that are antecedents of teamwork processes in software development. Nevertheless, TPA is a good starting point, covering most constructs found in the established literature and that may be extended in future research. Further, the process used to develop and validate TPA can be applied in future extensions.

A third limitation is that TPA only measures team level factors of structure and composition, leaving out organizational factors, as discussed above. We consciously made the choice to focus on team level constructs for two reasons. First, we wanted to create a measurement instrument that could be used to manage software teams in practice. The team level factors are more likely to be directly influenced by project managers, human resources personnel, team leaders, and the team members themselves, thus making the measurement instrument actionable in practice. Second, using more constructs (thus, more response items) would mean a longer and more complex questionnaire. This could affect the quality of the answers in unpredictable ways and could confound the results of the questionnaire validation. For instance, participants could answer the later questions with less quality due to fatigue. Further, we would need to increase the size of our sample, which is always problematic in field studies with professionals. Therefore, we started with the team level factors to build a comprehensive and valid measure, leaving the organizational level factors for future studies. Once more, the process used to build TPA could also be applied in the development of the measure for the organizational level factors.

Finally, the validation of the questionnaire may have a cultural bias because all participants were from the same country and thus might have a more uniform interpretation of the response items than a more culturally heterogeneous sample. Although related, it is not the same problem as validating only the Portuguese questionnaire, because culture could still affect the interpretation of the response items independently from language. This is a known and complex problem in psychometrics and we believe that only with the consistent development and validation of measurement instruments and their application in different contexts will we achieve higher confidence of the instrument's validity.

5.3 Implications for Research

To the best of our knowledge, TPA is the first validated measure of team structure and composition developed specifically for software development teams. Further, TPA has shown excellent validity and reliability characteristics. Therefore, we believe that TPA can have multiple applications in the study of software teams, some of which we discuss below.

The first area of research that can apply TPA is in understanding the potential correlations between composition and structure in software teams. It seems that some structure constructs are developed as the team members interact and the way structure is developed may be affected by team composition. TPA can be used to investigate such relationships, in particular in longitudinal, cohort studies. The results presented in Section 4.2 (Task 10) are a starting point for such investigations.

Second, TPA can be used to investigate how factors of structure and composition relate to process quality. TPA was developed to measure antecedents of team process quality. In a complementary way, the Teamwork Quality (TWQ) model is a measure of teamwork process quality. The application of both TPA and TWQ would allow the identification of the potential effects of team structure and composition on the processes modeled by TWQ.

Third, TPA would facilitate the investigation of direct relationships between structure and composition, and individual and team level outcomes. Gladstein's model contends that composition and structure are related to team processes and also, directly, to team outcomes such as performance and satisfaction. Future research can study the relationship of TPA constructs with these and other outcomes, such as turnover and job burnout.

Fourth, the conceptualization of the constructs and the operationalization in terms of response items were done mostly based on qualitative studies performed with software development professionals. In particular with the construction of response items, we must try to use the terminology and even the phrases as they are spoken and understood by the intended participants who will answer the questions. Therefore, even though the constructs could be generalized, we do not want to claim that the operationalization could be used in other domains without test, because this would introduce a threat to construct validity.

Finally, TPA can be used to investigate the levels of agreement or disagreement among team members with respect to structure and composition. Further, we could use TPA to identify agreement (or disagreement) between team members and team leaders or managers. This is also an important potential application of TPA in practice, as will be

discussed in Section 5.4.

5.4 Implications for Practice

TPA is a measure of the perceptions of individuals about their work in teams. Understanding this subjective perception is important for all levels of the software organization, including human resources personnel, project managers, team leaders, and the team members themselves. Measures of TPA may be used in practice in at least two complementary ways.

First, the measure can be used to investigate how the perceptions of levels of TPA are more or less uniform within the team and between the team and other external stakeholders, which may reveal several important issues related to software team management. A non-exhaustive list of direct application of TPA in the industrial practice is:

- Within team perception:
    - Heterogeneous perceptions within a team may be related to dysfunctional behavior, conflict, and, ultimately, ineffective performance. TPA could be applied to all members of a software team and inter-rater reliability indices could be used to identify levels of disagreement among team members. Subsequently, qualitative techniques could be used to identify reasons or sources of disagreement, contributing to a deeper understanding of team and individual behavior.
    - Homogeneous within-team perceptions reveal that team members share a consistent view of the team process antecedents. Homogeneous high scores indicate that team members agree the team possesses required levels of team process antecedents, and this should correlate with high levels of team process performance. Conversely, homogeneous low levels of the scores show the shared perception that the team does not have the necessary levels of antecedent to perform well.
- Between team perceptions: different teams in the same software organization may have different perceptions regarding TPA constructs. By recording these differences, the organization could investigate how different structures and compositions are related to other organizational indicators, such as project success.
- Between team and other stakeholders' perceptions: managers (in general) and team members may have distinct perceptions about structure and composition, which could indicate miscommunication and lack of appropriate feedback, among other factors. TPA can be used to measure the alignment of perceptions between managers and team members.

TPA may also be used as part of a more general evaluation of the team and other organizational aspects. As a valid and stable measure, TPA levels can be measured and correlated with other factors in the organizational context, such as job satisfaction, job retention or turnover, job burnout, and several others related to individual and team effectiveness.

5.5 Measure Development in Empirical Software Engineering Research

In this section, we discuss each task of our measurement scale development process in terms of its impact on scale validity (content, construct, internal reliability, etc.). We show the potential benefits for using each task as well as the potential threats to validity that are introduced if the task is not properly performed during the measure development (Table 20).

Table 20. Impacts of the Tasks of Measure Development on Instrument Validity.

Task	**Potential benefits** *(when task is properly performed)*	**Potential Threats to Validity** *(when task is not properly performed)*
Task 1: Identification of Constructs	✓ Establishes and clarifies the relevant theoretical and conceptual foundations for the constructs. ✓ Help to structure the traceability between the theory and the constructs used.	✓ It can contribute to casting doubt about the conceptualization of constructs. ✓ The origin and foundations of the constructs are lost.
Task 2: Conceptualization of Constructs	✓ Establishes clear, consistent definitions of the constructs identified in task 1. ✓ Traceability between the theory, constructs, and definitions used.	✓ Makes it hard to verify the consistency between constructs and definitions. ✓ May introduce inconsistent construct definitions.
Task 3: Operationalization of Constructs	✓ Establishes clear, consistent response items for each construct from task 2. ✓ Traceability between the theory, constructs, definitions, and	✓ Makes it hard to verify the consistency between theory, constructs, definitions, and operationalization items. ✓ The origin of the operationalization items

Task	Potential benefits *(when task is properly performed)*	Potential Threats to Validity *(when task is not properly performed)*
	operationalization items used.	is lost, which may generate doubts regarding the relationship with the construct (construct validity)
Task 4: Translation and Analysis of Items	✓ Selects translation method according to research focus and goals. ✓ Increases content validity by conducting a pilot test. ✓ Ensures that the selection of scale items addresses not only empirical issues but also includes practical and cultural considerations.	✓ Consistency and replicability of the item translation process is lost. ✓ Threatens content validity with the possibility of including inconsistent items in the scale.
Task 5: Structuring of Research Instrument	✓ A consistent structure of the research instrument. ✓ Decreases risk of bias introduction due to poor instrument structure.	✓ May inadvertently add bias due to sequencing of the questions. ✓ Poor structure my reduce attention and focus of respondent.
Task 6: Data Collection Planning	✓ Clearly and consistently establishes the target population and sampling method. ✓ Data collection within a population consistent with the research objectives. ✓ Increases possibility of reaching higher numbers of participants.	✓ Sampling method not aligned with the scope of the survey. ✓ Data collection within a population not aligned with the research objectives. ✓ Biased sampling leading to invalid results.
Task 7: Research Instrument Application	✓ Planning and organization for research execution. ✓ Establishment of appropriate means of making the research instrument available to the participating sample.	✓ Inadequate data collection with, for example, an insufficient sample size to perform the required analyses.
Task 8: Evaluation of Data Adequacy to Multivariate Analysis	✓ Demonstrates consistency with the statistical techniques used (e.g. missing data, outliers, normality, and multicollinearity).	✓ Skewed results in terms of model adjustment statistics, estimates, and parameter significances.
Task 9: Exploratory Factorial Analysis	✓ Demonstrates internal reliability. ✓ Confirms or refutes the structure of a research instrument. ✓ Reduces and optimizes a factorial structure.	✓ Unverified factorization conditions threatens the validity of the EFA procedures used. ✓ Difficulty in replicating the procedures performed to keep or remove items. ✓ An inconsistent or inadequate factorial structure.
Task 10: Confirmatory Factorial Analysis	✓ Presentation of model quality ✓ Presentation and use of criteria to adjust the models ✓ Greater internal reliability and internal consistency by using more than one index. ✓ A more rigorous test of construct validity. ✓ Greater convergent validity, i.e. the degree to which two measures of the same concept are correlated. ✓ Greater discriminant validity, demonstrating that a scale is sufficiently different from another similar concept.	✓ Difficulty in comparing results. ✓ Lack of information about how the model adjustments were performed. ✓ Lack of confidence in the internal reliability of the scale. ✓ Lack of confidence that a set of established items represents the theoretical latent construct that they are designed to measure. ✓ Cannot demonstrate the validity of a measurement model without proving the quality of fit and evidence of construct validity (convergent and discriminant).

Thus, the execution of Tasks 1 to 10, as reported in this research, help to establish a uniform, systematic, replicable, and rigorous process for the development and validation of measurement instruments in empirical software engineering research.

6 Concluding Remarks

In this article, we describe the development of a measure of team process antecedents that captures the perception of

software team members about different aspects of team composition and structure. We used a well-established model for measurement instrument development and validation (Pasquali, 2010) that addresses the theoretical, empirical, and analytical aspects of instrument construction. The development and validation of the TPA questionnaire had the participation of 375 team leaders and team members of software development teams from 100 different organizations.

To validate our measure, we applied a series of statistical tests with emphasis on Exploratory (EFA) and Confirmatory Factor Analysis (CFA). EFA promoted the reduction of the total number of items initially established for the measurement instrument, from 51 to 35, in addition to reorganizing those items in 9 factors. The results of the EFA retained the items and factors that presented greater statistical and theoretical compatibility. We also demonstrated the internal reliability of the scale through the use of Cronbach's alpha. Subsequently, the Factors Matrix (1st order) resulting from the EFA was analyzed, in which significant correlations were observed indicating the existence of two higher-order factors (2nd order).

Starting from the EFA results, CFA was performed in order to analyze three models based on their quality indices, standardized residues, modification indices and the value of the critical ratio. These values indicated that Model 1, containing nine 1st order factors, was the most adequate to express the constructs of interest. The reliability of this model was verified using the Jöreskog's Rho indices (ρ) and Cronbach's alpha. In addition, the convergent and discriminant validity of the construct of the scale were confirmed.

In Section 5.1, we discussed several areas for improving this study and directions for future research on the topic. In particular, we are currently investigating the application of TPA and TWQ (including Team Performance and Personal Success scales) in a broader set of individuals to verify the influence and relationships between antecedents, TWQ, and outcomes of software development teams (individual and collective).

We also discussed the implications of our results for research and practice in Sections 5.3 and 5.4, respectively. In research about software teams, TPA can be used to investigate the potentially complex relationships among all components in the IPO framework: between composition and structure, and other antecedents; between composition and structure, and team processes; and, supported by Gladstein's Model (1998), the direct relationships between antecedents and outputs.

In practice, TPA can provide a better understanding of the perceptions of software engineers regarding the adequacy of composition and structure of their software teams. In particular, TPA can be used to measure how heterogeneous these perceptions are within and between teams in an organization. Within team heterogeneity is important because heterogeneous perceptions among team members may lead to dysfunctional behavior, conflict, and, ultimately, ineffective performance. Between team heterogeneity is important to understand how different structures and compositions are related to other organizational indicators, such as project success.

To the best of our knowledge, a validated measure of team structure and composition did not exist in the empirical software engineering literature. Further, we present the development of the TPA measure in the form of a guideline that may be used in the construction of other measurement instruments in empirical software engineering research. Thus, the use of TPA (scale, construction and validation strategy, and results) can generate benefits, both in the context of research and software engineering practice. We believe that these results are important contributions of this article.

Acknowledgements

Fabio Q. B. da Silva holds a research grant from CNPq 306856/2017-4. The authors would like to thank the anonymous reviewers and the EMSE editors for their feedback on the first version of this article, which helped to greatly improve this final version.

References

Aladwani, Adel M. (2002). An Integrated Performance Model of Information Systems Projects. Journal of Management Information Systems 19 (1): 185–210. https://doi.org/10.1080/07421222.2002.11045709.

Bagozzi, Richard P., Youjae Yi, and Lynn W. Phillips. (1991). Assessing Construct Validity in Organizational Research. Administrative Science Quarterly 36 (3): 421. https://doi.org/10.2307/2393203.

Brace, Ian. (2018). Questionnaire design: How to plan, structure and write survey material for effective market research. Kogan Page Publishers.

Brown, TA. (2006). Confirmatory Factor Analysis for Applied Research. New York, NY, US: Guilford Publications.

Browne, M. W., and R. Cudeck. (1989). Single Sample Cross-Validation Indices for Covariance Structures. Multivariate Behavioral Research. https://doi.org/10.1207/s15327906mbr2404_4.

Byrne, Barbara M. (2010). Structural Equation Modeling With AMOS Basic Concepts, Applications, and Programming

(Multivariate Applications Series). Structural Equation Modeling. 2nd Ed. Ottawa, Ontario, Canada: Routledge. https://doi.org/10.1080/10705511.2014.935842.

Campion, M. A., Medsker, G. J., & Higgs, A. C. (1993). Relations between work group characteristics and effectiveness: Implications for designing effective work groups. Personnel psychology, 46(4), 823-847.

Campion, Michael A., Ellen M. Papper, and Gina J. Medsker. (1996). Relations between Work Team Characteristics and Effectiveness: A Replication and Extension. Personnel Psychology. https://doi.org/10.1111/j.1744-6570.1996.tb01806.x.

Carpenter, Mason A. (2002). The implications of strategy and social context for the relationship between top management team heterogeneity and firm performance. Strategic Management Journal, v. 23, n. 3, p. 275-284,

Carson, J. (2006). Internal team leadership: An examination of leadership roles, role structure, and member outcomes. Dissertation submitted to the Faculty of the Graduate School of the University of Maryland, College Park, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Cattell, R. (2012). The Scientific Use of Factor Analysis in Behavioral and Life Sciences. Edited by Springer Science & Business Media. Springer, Boston, MA. https://doi.org/https://doi.org/10.1007/978-1-4684-2262-7.

Cha, Jongseok, Youngbae Kim, Jeong Yeon Lee, and Daniel G. Bachrach. (2015). Transformational Leadership and Inter-Team Collaboration: Exploring the Mediating Role of Teamwork Quality and Moderating Role of Team Size. Group and Organization Management 40 (6): 715–43. https://doi.org/10.1177/1059601114568244.

Clark, Lee Anna, and David Watson. (1995). Constructing Validity: Basic Issues in Objective Scale Development The Centrality of Psychological Measurement. Psychological Assessment 7 (3): 309–12.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. New York: Academic Press (2a ed.).

Cohen, Susan G. (1993). Designing Effective Self Managed Work Teams. Advances in Interdisciplinary Studies of Work Teams, 1994: Theories of Self-Managing Work Teams 1: 67–102.

Damásio, Bruno F. (2012). Uso Da Análise Fatorial Exploratória Em Psicologia. Avaliação Psicologica.

Dayan, Mumin, and C. Anthony Di Benedetto. (2009). Antecedents and Consequences of Teamwork Quality in New Product Development Projects: An Empirical Investigation. European Journal of Innovation Management 12 (1): 129–55. https://doi.org/10.1108/14601060910928201.

Dickinson, T. L., and McIntyre, R. M. (1997). A conceptual framework for teamwork measurement. In M. T. Brannick & E. Salas (Eds.), Team performance assessment and measurement: Theory, methods, and applications (pp. 19–43). Mahwah NJ: NEA.

Dokko, Gina; Wilk, Steffanie L.; Rothbard, Nancy P. (2009). Unpacking prior experience: How career history affects job performance. Organization Science, v. 20, n. 1, p. 51-68,

Dreesen, Tim; Schmid, Thomas. (2018). Do As You Want Or Do As You Are Told? Control vs. Autonomy in Agile Software Development Teams. Proceedings of the 51st Hawaii International Conference on System Sciences.

Easley, Robert F., Sarv Devaraj, and J. Michael Crant. (2003). Relating collaborative technology use to teamwork quality and performance: An empirical analysis. Journal of Management Information Systems 19.4: 247-265. https://doi.org/10.1080/07421222.2003.11045747.

Field, Andy. (2012). Descobrindo a Estatística Usando o SPSS. 2nd ed. Porto Alegre, RS: Bookman, Artmed.

Figueiredo Filho, D. B., and Silva Júnior, J. A. da. (2010). Visão Além Do Alcance: Um Introdução à Análise Fatorial. Opin. Publica. 16: 160–85. https://doi.org/10.1590/s0104-62762010000100007.

Fornell, Claes and Larcker, David F. (2006). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error: A Comment. Journal of Marketing Research 18 (3): 36–50. https://doi.org/10.2307/3150979.

George, D., and Paul M. (2003). SPSS for Windows Step by Step : A Simple Guide and Reference, 11.0 Update. Allyn and Bacon.

Gladstein, Deborah L. (1984). Groups in Context: A Model of Task Group Effectiveness. Administrative Science Quarterly 29 (4): 499. https://doi.org/10.2307/2392936.

Hackman, J. R. (1987). The design of work teams. In J. W. Lorsch (Ed.), Handbook of organizational behavior (pp. 315 – 342). Englewood Cliffs, NJ: Prentice Hall.

Hair, J. F., Black, B., Babin, B., Anderson, R. E., Tatham, R. L. (2009). Análise Multivariada de Dados. 6ª Ed. Porto Alegre: Bookman.

Hashmi, Arshia, Shahibuddin Ishak, and Hazlinda Binti Hassan. (2018). Role of team size as a contextual variable for the relationship of transformational leadership and teamwork quality. Asian Journal of Multidisciplinary Studies 6: 5.

Hashmi, Arshia, Shahibuddin Ishak, Hazlinda Binti Hassan, and Muhammad Azeem Ahmad. (2017). A Conceptual Framework for Describing the Innovation in Teams. International Journal of Economic Research 14 (14PartII): 59–72.

Hoegl, Martin, and Hans Georg Gemuenden. (2001). Teamwork Quality and the Success of Innovative Projects: A

Theoretical Concept and Empirical Evidence. Organization Science 12 (4): 435–49. https://doi.org/10.1287/orsc.12.4.435.10635.

Hoegl, Martin, and K. Praveen Parboteeah. (2003). Goal Setting and Team Performance in Innovative Projects: On the Moderating Role of Teamwork Quality. Small Group Research 34 (1): 3–19. https://doi.org/10.1177/1046496402239575.

Hoegl, Martin, Parboteeah, K. Praveen, Gemuenden, Hans Georg. (2003). When teamwork really matters: task innovativeness as a moderator of the teamwork–performance relationship in software development projects. Journal of Engineering and Technology Management, v. 20, n. 4, p. 281-302.

Hoegl, Martin, and K. Praveen Parboteeah. (2006a). Autonomy and Teamwork in Innovative Projects. Human Resource Management 45 (1): 67–79. https://doi.org/10.1002/hrm.20092.

Hoegl, Martin, and K. Praveen Parboteeah. (2006b). Team Goal Commitment in Innovative Projects. International Journal of Innovation Management 10 (03): 299–324. https://doi.org/10.1142/s136391960600151x.

Hoegl, Martin, and K. Praveen Parboteeah. (2006c). Team Reflexivity in Innovative Projects. R and D Management 36 (2): 113–25. https://doi.org/10.1111/j.1467-9310.2006.00420.x.

Hoegl, Martin, and Luigi Proserpio. (2004). Team Member Proximity and Teamwork in Innovative Projects. Research Policy 33 (8): 1153–65. https://doi.org/10.1016/j.respol.2004.06.005.

Hoegl, Martin, Katharina Weinkauf, and Hans Georg Gemuenden. (2004). Interteam Coordination, Project Commitment, and Teamwork in Multiteam R&D Projects: A Longitudinal Study. Organization Science 15 (1): 38–55. https://doi.org/10.1287/orsc.1030.0053.

Jehn, Karen A.; Bezrukova, Katerina. (2004). A field study of group diversity, workgroup context, and performance. Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, v. 25, n. 6, p. 703-729,

Dias Júnior, J. J. L. (2016). Adaptação e Tradução de Escalas de Mensuração Para o Contexto Brasileiro : Um Método Sistemático Como Alternativa a Técnica Back-Translation. Métodos e Pesquisa Em Administração 1 (December): 4–12.

Kitchenham, B A, and S L Pfleeger. (2002). Principles of Survey Research Part 5: Populations and Samples. ACM SIGSOFT Software Engineering Notes 27 (5): 17–20.

Kline, R. (2010). Principles and Practice of Structural Equation Modeling. In Structural Equation Modeling, 534. New York, NY, US: The Guilford Press. https://doi.org/10.1038/156278a0.

Koufteros, Xenophon A. (1999). Testing a Model of Pull Production: A Paradigm for Manufacturing Research Using Structural Equation Modeling. Journal of Operations Management 17 (4): 467–88. https://doi.org/10.1016/S0272-6963(99)00002-9.

Kozlowski, S. W., & Bell, B. S. (2013). Work groups and teams in organizations: Review update.

Laros, J. A. 2012. O Uso Da Análise Fatorial: Algumas Diretrizes Para Pesquisadores. In Análise Fatorial Para Pesquisadores, 141–60. Brasilia, DF: LabPam Editora.

Levine, John M.; Moreland, Richard L. (1990). Progress in small group research. Annual review of psychology, v. 41, n. 1, p. 585-634.

Lindsjørn, Yngve, Dag I.K. Sjøberg, Torgeir Dingsøyr, Gunnar R. Bergersen, and Tore Dybå. (2016). Teamwork Quality and Project Success in Software Development: A Survey of Agile Development Teams. Journal of Systems and Software 122: 274–86. https://doi.org/10.1016/j.jss.2016.09.028.

Littlepage, Glenn; Robison, William; Reddington, Kelly. (1997). Effects of task experience and group experience on group performance, member ability, and recognition of expertise. Organizational behavior and human decision processes, v. 69, n. 2, p. 133-147.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. Psychological Bulletin, 100(1), 107-120. doi. 10.1037/0033-2909.100.1.107.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. Academy of management review, 26(3), 356-376.

Marôco, João. (2010). Análise de Equações Estruturais - Fundamentos Teóricos, Software e Aplicações. ReportNumber.

Marôco, João. (2018). Análise Estatística Com o SPSS Statistics. 7a ed. Pêro Pinheiro: ReportNumber. https://doi.org/10.1111/1440-1681.12086.

Marsicano, G., Pereira, D. V., da Silva, F. Q., & França, C. (2017). Team maturity in software engineering teams. In Proceedings of the 11th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (pp. 235-240). IEEE Press. https://doi.org/10.1109/ESEM.2017.36.

Mathieu, John, Travis M. Maynard, Tammy Rapp, and Lucy Gilson. (2008). Team Effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse into the Future. Journal of Management 34 (3): 410–76.

https://doi.org/10.1177/0149206308316061.

McGrath, Joseph Edward. 1964. Social psychology: A brief introduction. Holt, Rinehart and Winston,

Mehta, Nikhil; Hall, Dianne; Byrd, Terry. (2014). Information technology and knowledge in software development teams: The role of project uncertainty. Information & Management, v. 51, n. 4, p. 417-429,

Mom, Tom JM; Fourné, Sebastian PL; Jansen, Justin JP. (2015). Managers' work experience, ambidexterity, and performance: The contingency role of the work context. Human Resource Management, v. 54, n. S1, p. s133-s153,

Morgenson, F., Campion, M. A., & Bruning, P. F. (2012). Job and team design. Handbook of human factors and ergonomics (4th ed., pp. 441–474). Hoboken, NJ: John Wiley & Sons.

Nascimento, Thiago Gomes. (2014). Desempenho Profissional: Relações Com Valores, Práticas e Identidade No Serviço Policial. Tese de Doutorado, Universidade de Brasília.

Neiva, E. R., Abbad, G. e Tróccoli, B. T. (2007). Roteiro Para Análise Fatorial de Dados., Universidade de Brasília, Brasília, DF.

De Oliveira, Millena Lauyse Silva. (2019). O Estudo de Equipes de Desenvolvimento de Software na Indústria: Um Mapeamento Sistemático da Literatura. Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco. Dissertação de Mestrado.

Osborne, Jason, Jason W. Osborne, Anna B. Costello, and J. Thomas Kellow. (2014). Best Practices in Exploratory Factor Analysis. Best Practices in Quantitative Methods. Scotts Valley, CA: CreateSpace Independent Publishing. https://doi.org/10.4135/9781412995627.d8.

Pasquali, Luiz. (2010). Testes Referentes a Construto: Teoria e Modelo de Construção. In Instrumentação Psicológica: Fundamentos e Práticas, 165–98. Porto Alegre: ArtMed.

Pasquali, Luiz. (2012). Análise Fatorial Para Pesquisadores. Brasília, DF: LabPam Editora. www.scielo.br/reeusp.

Pereira, D. V., Corrêa, G. M., da Silva, F. Q., & Ribeiro, D. M. (2017). Team maturity in software engineering teams: a work in progress. In Proceedings of the 10th International Workshop on Cooperative and Human Aspects of Software Engineering (pp. 70-73). IEEE Press. https://doi.org/10.1109/CHASE.2017.2.

Quiñones, Miguel A. (2004). Work experience: A review and research agenda. International review of industrial and organizational psychology, v. 19, p. 119-138,

Raykov, Tenko. (2012). Scale Construction and Development Using Structural Equation Modeling. In In R. H. Hoyle (Ed.), Handbook of Structural Equation Modeling, 472–92. New York, NY, US: The Guilford Press.

Salas, E., Stagl, K. C., Burke, C. S., & Goodwin, G. F. (2007). Fostering team effectiveness in organizations: Toward an integrative theoretical framework. In Nebraska symposium on motivation (Vol. 52, p. 185).

Spreitzer, G. M. (1995). Psychological, Empowerment In The Workplace: Dimensions, Measurement And Validation. Academy Of Management Journal, 38(5), 1442–1465. Doi:10.2307/256865

Stray, Viktoria; Faegri, Tor Erlend; Moe, Nils Brede. (2016). Exploring norms in agile software teams. In: International Conference on Product-Focused Software Process Improvement. Springer, Cham, p. 458-467.

Tabachnick, Barbara G., and Linda S. Fidell. (2001). Using Multivariate Statistics. Allyn and Bacon.

Tannenbaum, Scott I., Rebecca L. Beard, and Eduardo Salas. (1992). Team building and its influence on team effectiveness: An examination of conceptual and empirical developments. Advances in psychology. Vol. 82. North-Holland, 117-153.

Wageman, Ruth, J. Richard Hackman, and Erin Lehman. (2005). Team Diagnostic Survey : Development of an Instrument. Journal of Applied Behavioral Science 41 (4): 373–98. https://doi.org/10.1177/0021886305281984.

Wickramasinghe, V., & Nandula, S. (2015). Diversity in team composition, relationship conflict and team leader support on globally distributed virtual software development team performance. Strategic Outsourcing: An International Journal, 8(2/3), 138-155. doi: 10.1108/SO-02-2015-0007,

Wolff, Hans-georg, and Katja Preising. (2005). Exploring Item and Higher Order Factor Structure with the Schmid-Leiman Solution: Syntax Codes for SPSS and SAS. Behavior Research Methods 37 (1): 48–58.

Yang, Li Ren, Chung Fah Huang, and Kun Shan Wu. (2011). The Association among Project Manager's Leadership Style, Teamwork and Project Success. International Journal of Project Management 29 (3): 258–67. https://doi.org/10.1016/j.ijproman.2010.03.006.

APPENDIX A

Table 21 presents in the first three columns the set of constructs obtained by Pereira, et al., (2017) and Marsicano, et al. (2017) (first column), Gladstein (1984) (second column) and the ad-hoc literature review (third column). Finally, the last column presents the constructs we chose to be used in the research questionnaire. The decision was based on our experience and knowledge supported by the relevance of each construct in the literature reviewed.

Table 21. Comparing and decision related to the set of constructs to be used.

Empirical Study (Pereira, et al., 2017) (Marsicano, et al., 2017)	**Gladstein's Model** (Gladstein, 1984)	**Literature Review**	**Decision**
Team Experience in the Organization	-	Experiences[A]	Team Experience in the Organization
Team Experience with work	-	Experience with work[B]	Team Experience with work
-	Heterogeneity	Heterogeneity[C] Functional Diversity[D]	Heterogeneity
-	Organizational Tenure	-	Will be collected as demographic data.
-	Job Tenure	-	
Skills (interpersonal, managerial and technical)	Adequate Skills	Interpersonal (social), Management and Technical Skills[E] Technical Competence[F]	Adequate Skills (Interpersonal, Management and Technical)
Roles and responsibilities	Role and Goal Clarity	Goal Setting[G] Clarity of Roles And Goals[H] Team Goal Commitment[I]	Role and Goal Clarity
Work Organization (work processes)	Specific Work Norms	Work Norms[J] Team Orientation[K]	Specific Work Norms
-	Task Control	Team Autonomy[L] Task Control[M]	Task Control
Team Size	Size	Team Size[N]	Will be collected as demographics data and "adequate team size" (Wageman, et al., 2005) will be part of the questionnaire.
Leadership Style	Formal Leadership	Leadership[O]	Formal Leadership

[A] (Quiñones, 2004; Wageman, et al., 2005; Hoegl and Parboteeah 2006a).

[B] (Littlepage, et al., 1997; Aladwani, 2002; Dokko, et al. 2009; Mom, et al. 2015; Dreesen e Schmid, 2018).

[C] (Campion, et al., 1993; Carpenter, 2002; Jehn e Bezrukova, 2004; Hoegl and Parboteeah, 2006a).

[D] (Dayan and Di Benedetto 2009; Wageman et al., 2005; Lee e Xia, 2010; Morgenson, et al. 2012; Mehta, et al., 2014).

[E] (Hoegl and Proserpio, 2004; Wageman, et al., 2005; Hoegl and Parboteeah, 2006a; Hoegl and Parboteeah, 2006c).

[F] (Aladwani, 2002; Hoegl and Parboteeah, 2006b; Dayan and Di Benedetto, 2009).

[G] (Hoegl and Parboteeah, 2003; Hashmi, et al., 2018).

[H] (Hoegl and Proserpio, 2004).

[I] (Hoegl and Parboteeah, 2006b).

[J] (Levine e Moreland, 1990; Tannenbaum, et al., 1992; Hoegl and Proserpio, 2004; Wageman, et al., 2005; Stray, et al., 2016).

[K] (Sala,s et al., 2005).

[L] (Hoegl, et al., 2004; Hoegl and Parboteeah, 2006a).

[M] (Hackman e Oldham, 1980; Lee e Xia, 2010).

[N] (Levine e Moreland, 1990; Hoegl and Proserpio, 2004; Wageman, et al., 2005; Hoegl, 2005; Hoegl e Parboteeah, 2006b; Kozlowski e Bell, 2013; Cha, et al. 2015; Hashmi, et al. 2017; Hashmi, et al., 2018).

[O] (Dickinson and McIntery, 1997; Salas, et al., 2005; Morgeson, 2005; Carson, 2006; Hoegl and Parboteeah, 2006a; Salas, et al., 2007; Cha, et al., 2015; Hashmi, et al., 2017; Hashmi, et al., 2018).

Empirical Study (Pereira, et al., 2017) (Marsicano, et al., 2017)	**Gladstein's Model** (Gladstein, 1984)	**Literature Review**	**Decision**
-	-	Team Composition[P]	Will be collected as a multidimensional construct as described in Gladstein's model.

APPENDIX B

Table 22. List of Response Items, per 1st Level Constructs, of TPA.

Identifier	**Description**	**Item**	**References**
Adequate Skills			
INT1	My team is always willing to collaborate to perform the activities assigned to it.	New	New items built based on studies Aladwani, 2002; Hoegl and Parboteeah, 2006a; Hoegl and Parboteeah, 2006b; Hoegl and Parboteeah, 2006c; Pereira, et al., 2017; Marsicano, et al., 2017).
INT2	My team can establish relationships of respect and trust among its members.	New	
INT3	My team members feel that they are, in fact, part of the team.	New	
INT4	My team can work proactively.	New	
INT5	The members of my team are always willing to ask, offer and receive help, voluntarily, among themselves.	New	
INT6	My team is willing to give and receive feedbacks consistently among themselves.	New	
INT7	My team members do not possess barriers to talk about work.	New	
INT8	My team members have the liberty and openness to make personal conversations (private) among themselves.	New	
INT9	My team knows how to separate technical discussions from personal matters.	New	
INT10	My team is capable to solve interpersonal conflicts among its members.	New	
INT11	My team can deal with pressure and adversities, not allowing it to negatively affect their results.	New	
MNG1	My team can keep the visibility on the progress of their work to all stakeholders (for example: manager, clients, etc.).	New	
MNG 2	My team can follow and monitor their work continuously.	New	
MNG 3	My team can make decisions with the participation of all its members.	New	
MNG 4	My team can establish the objectives and goals of their work autonomously.	New	
MNG 5	My team assumes the responsibility of meeting their commitments.	New	
MNG 6	My team can stay focused on their goals and objectives.	New	
TEC1	My team can establish and adapt their processes, methods and software engineering practices autonomously.	New	
TEC2	My team knows the business domain for which they will develop the solution.	New	
TEC3	My team has the necessary knowledge about the technology and standards used to accomplish their activities.	New	
TEC4	My team has the necessary knowledge about the methods, practices and tools used to accomplish their activities.	New	
TEC5	My team can establish technical solutions without the need for direct interference from the manager and the software architect.	New	
TEC6	My team can understand and give a solution to a technical problem in a manner appropriate to the client and the project.	New	
TEC7	My team can solve the technical problems inherent in software development with the participation of all its members.	New	

[P] (Hackman, 1987; Hoegl and Parboteeah, 2003; Wageman, et al., 2005; Easley, et al., 2017).

Identifier	Description	Item	References
TEC8	My staff needs constant interference from the manager and the software architect to be able to perform their activities. (R)	New	
	Heterogeneity		
HET1	My team members have diverse areas of expertise in software development.	Adapted	Items adapted from study (Campion, et al., 1993).
HET2	My team members have different backgrounds and software development experiences.	Adapted	
HET3	My team members have competencies that complement each other.	Adapted	
	Team Experience in the Organization		
EOR1	My team members have already worked together on other software development projects in the Organization where we work.	New	New items built based on studies of (Gladstein, 1984; Aladwani, 2002; Pereira, et al., 2017; Marsicano, et al., 2017).
EOR2	My team members have already participated in other software development projects, in other teams, in the Organization where we work.	New	
EOR3	My team members have already played more than one role in software development projects in the Organization where we work.	New	
	Team Experience with work		
TEW1	My team members have worked on other projects with the same business domain as our current project.	New	New items built based on studies of (Gladstein, 1984; Aladwani, 2002; Pereira, et al., 2017; Marsicano, et al., 2017).
TEW2	My team members have worked with the same technologies as those used in our current project.	New	
TEW3	My team members have used the same processes and practices in previous projects that will be used in our current project.	New	
TEW4	My team members lack experience in working in a more autonomous way, without the interference of the manager and the software architect, for example. (R)	New	
TEW5	My team members have always played the same role in other projects as they play in our current project.	New	
	Role and Goal Clarity		
CGR1	My team's goals are clearly established.	New	New items built based on studies of (Gladstein, 1984; Aladwani, 2002; Hoegl and Parboteeah, 2003).
CGR2	My team's goals are shared and accepted by all its members.	New	
CGR3	There is a lack of clarity in the definition of the roles and responsibilities of my team members. (R)	New	
CGR4	The roles and responsibilities of my team members are accepted by all.	New	
	Specific Work Norms		
NOR1	My team's work process is established and aligned with its needs	New	New item built based on (Marsicano, et al., 2017). Items adapted from study (Wageman, et al., 2005).
NOR2	My team's behavior patterns are vague and unclear. (R)	Adapted	
NOR3	It is clear what is and what is not an acceptable behavior for the members of my team.	Adapted	
NOR4	My team members agree on how they should behave.	Adapted	
	Task Control		
TSC1	During the software development process, the authority (through decision-making, monitoring, and management of the team, process and performance) is centered on the external manager, making the team responsible only for the operational execution of the work.	Adapted	Item adapted from study (Wageman, et al., 2005).
	Adequate Team Size		
TEAM1	Considering the work as a whole, most of the time my team has the appropriate number of people to perform the task that needs to be accomplished.	Adapted	Item adapted from study (Wageman, et al., 2005).
	Formal Leadership		
LEAD1	The manager (or immediate manager) gives freedom and	New	New items built based on

Identifier	Description	Item	References
	encourages my team to make their own decisions about work.		studies (Gladstein, 1984; Cohen, 1993; Campion, et al., 1993; Hoegl and Parboteeah, 2006a; Yang, et al., 2011; Pereira, et al., 2017; Marsicano, et al., 2017; Ishak, et al., 2018).
LEAD2	The manager (or immediate manager) stimulates my team to work autonomously.		
LEAD3	My team shares responsibilities and leadership.		
LEAD4	The manager (or immediate manager) plays a facilitating role in the decision-making and in the execution of my team's work.		
LEAD5	The manager (or immediate manager) encourages my team to solve their own problems.		

Note: These items are translations of the original Portuguese language items used.

APPENDIX C

Tables to use in Confirmatory Factorial Analysis (CFA).

Table 23. Description of the Quality Indices Classifications (Marôco, 2010).

Indices	Description
Absolute Indices	They evaluate the quality of the model itself, without comparison with other models.
Relative Indices	They evaluate the quality of the model of the study in comparison to the model with the worst possible adjustment (restrictive) and/or to the model with the best possible adjustment.
Parsimony Indices	They aim to compensate for the artificial improvement of the model, which is only achieved by the inclusion of more free parameters approaching the studied model to the saturated model.
Population Discrepancy Indices	They compare the adjustment of the obtained model with the sample moments (averages and sample variances) in relation to the adjustment of the model that would be obtained with the population moments (population averages and variances).
Information Theory Indices	They are based on the χ^2 and penalize the model in function of its complexity. These indices do not represent reference values (the lower the better) and are used to compare alternative models that fit the data. The best model will be the one that presents the lowest indices.

Table 24. Quality Indices, Reference Values, Description and Classification.

Indices	Reference Value[Q]	Description	Indices Classification[R]
χ^2 /df (Chi-square by degree of freedom)	$1 \leq \chi^2 / df \leq 5$	It allows to detect the model adjustment, measuring the degree of absolute parsimony of the model. Values between 2 and 5 are acceptable, less than 2 are good and above 5 are unacceptable (Marôco, 2010).	Absolute
SRMR (Standardized Root Mean Square Residual)	< 0.08	It is a measure of the residual mean absolute correlation, the global difference between the observed and predicted correlations (Kline, 2010).	Absolute
GFI (Goodness of Fit Index)	> 0.90	GFI explains the proportion of the observed covariance between the manifested variables, explained by the adjusted model (Marôco, 2010).	Absolute
CFI (Comparative Fit Index)	> 0.90	CFI is an incremental adjustment index (Hair et al., 2009). It measures the relative decrease of the lack of adjustment (Marôco, 2010).	Relative
PGFI (Parsimony Comparative Fit Index)	> 0.60	It penalizes the GFI by the relation of parsimony (Marôco 2010).	Parsimony
PCFI (Parsimony Goodness of Fit Index)	> 0.60	It penalizes the CFI by the relation of parsimony (Marôco, 2010).	Parsimony
RMSEA (Root Mean Square Error of Approximation)	< 0.10	It represents best how well a model fits a population and not just a sample used for estimation (Hair, et al., 2009)	Population Discrepancy

[Q] The reference values used in this research are proposed by Marôco (2010), with the exception of SRMR, which is suggested by (Brown, 2006; Kline, 2010).

[R] This classification is proposed by Marôco (2010).

Indices	Reference Value[Q]	Description	Indices Classification[R]
ECVI (Expected Cross-Validation Index)	The smallest possible.	Reflects the theoretical adjustment of the model in other samples similar to the one in which the model was adjusted from, based on a single sample (Marôco, 2010).	Based on Information Theory

Table 25. Indices and Conditions to verify Internal Reliability and Construct Validity.

Criterion	Description	Indices or Tests	Condition	Reference
Internal Reliability	The verification of how the items that make up a scale reflect the construct that it is measuring.	Jöreskog's Rho (ρ).	Rho (ρ) > 0.70	Field (2012), Nascimento (2014).
Convergent validity	It is the degree to which two measures of the same concept are correlated.	Rho of convergent validity (Rho_{cv}).	$Rho_{cv} \geq 0.50$	Hair, et al. (2009), Fornell and Larcker (2006), Marôco (2010).
Discriminant validity	It is the degree to which a construct is truly different from the others.	Rho of convergent validity (Rho_{cv}).	$Rho_{cv} \geq$ the square of the correlation estimate between pars of constructs.	Fornell, et al. (2006), Koufteros (1999), Hair, et al. (2009), Marôco (2010).
		The chi-square difference test (χ^2) between a fixed solution (χ^2_r) and a free solution (χ^2_u).	$\chi^2_{dif} \geq \chi^2_{1-\alpha\,(glr\,-\,glu)}$.	Bagozzi, et al. (1991), Marôco (2010).

If $\chi^2_{dif} \geq \chi^2_{1-\alpha\,(glr\,-\,glu)}$ where, $\chi^2_{dif} = \chi^2_r - \chi^2_u$ and $\alpha = 0{,}05$ and $(glr - glu)$ is the difference between degrees of freedom of the fixed model and the free model, it is confirmed that the factors are not perfectly correlated, demonstrating discriminant validity (Bagozzi, et al., 1991; Marôco, 2010).

APPENDIX D

Observing Table 26, it is possible to notice that all latent factors of the 1st order have a high level of significance (> 0.50) when related to a 2nd order latent factor, except for the factor of 'experience in the organization'. Indicating the possibility of having a multifaceted, operationalized, 2nd order factor, from the 1st order factors. Table 27 presents the relation of the 1st order factors with two 2nd order factors, suggesting the existence of two groupings. In this scenario, all factors have substantive factor loads (> 0.50), except for the factor of 'experience in the organization'. In this second context, the distribution of factorial loads between the two groups is more parsimonious when compared to the data presented in Table 26. An indication that the existence of two latent factors of 2nd order may be more adequate than just one.

Observing Table 28, three main points can be noted. The first one refers to the maintenance of the grouping of 1st order factors, interpersonal skills, role and goal clarity, formal leadership, work experience, management skills and rules of behavior, with a 2nd order factor, as was reported in Table 27. This reinforces the evidence of grouping these factors into a higher-order factor. Second, 1st order factors that relate to Factor 2, of 2nd order, have unbalanced factorial loads (with a difference greater than 0.40 between them); in addition, the factor of experience in the organization has a load lower than 0.30, and can be considered statistically independent, not contributing to the factor analysis (Hair, et al. 2009). Third, Factor 3 (2nd Order) is related to only one 1st order factor, therefore not justifying the possible existence of a higher-order factor. In view of these points, the possibility of having 3 or more higher-order factors related to the 9 first-order factors identified in the EFA is ruled out.

Table 26. Factors Matrix with One 2nd Order Factor [a].

Name of the Factor (1st Order)	2nd Order Factor
	1
Interpersonal Skills (F1)	0.83
Role and Goal Clarity (F2)	0.63
Formal Leadership (F3)	0.71
Technical Skills (F4)	0.75

Table 27. Rotated Factors Matrix with Two 2nd Order Factors [a].

Name of the Factor (1st Order)	2nd Order Factors	
	1	2
Interpersonal Skills (F1)	0.68	0.47
Role and Goal Clarity (F2)	0.65	0.19
Formal Leadership (F3)	0.66	0.31
Technical Skills (F4)	0.50	0.60
Heterogeneity (F5)	0.27	0.62

Heterogeneity (F5)	0.57
Team Maturity (F6)	0.56
Management Skills (F7)	0.67
Rules of Behavior (F8)	0.66
Experience in the Organization (F9)	0.35

Extraction Method: Principal Axis Factoring.

a. 1 extracted factor. 5 necessary interactions.

Team Maturity (F6)	0.51	0.24
Management Skills (F7)	0.63	0.28
Rules of Behavior (F8)	0.53	0.39
Experience in the Organization (F9)	0.14	0.41

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 3 interactions.

Table 28. Factors Matrix with Three 2nd Order Factors [a].

Name of the Factor (1st Order)	2nd Order Factors		
	1	2	3
Interpersonal Skills (F1)	0.64	0.43	0.27
Role and Goal Clarity (F2)	0.69	0.12	0.16
Formal Leadership (F3)	0.64	0.26	0.22
Technical Skills (F4)	0.39	0.74	0.27
Heterogeneity (F5)	0.23	0.23	0.86
Team Maturity (F6)	0.48	0.25	0.14
Management Skills (F7)	0.58	0.41	0.06
Rules of Behavior (F8)	0.53	0.28	0.27
Experience in the Organization (F9)	0.14	0.29	0.23

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 5 interactions.

APPENDIX E

Table 29 presents the Rho (ρ) of reliability and convergence (Rho_{cv}) of each factor, as well as the factorial loads of each of the items that make up the final measurement scale. Table 29 also shows the mapping of each factor with the aspects of the team (composition and structure).

Table 29. Reliability and Convergent Validity of Factors - Model 1.

Team Composition					
Factor	**ρ**	**Rho_{cv}**	**Items**	**Description**	**Factorial Load**
Interpersonal Skills (F1)	0.84	0.52	INT2	My team can establish relationships of respect and trust among its members.	0.74
			INT3	My team members feel that they are, in fact, part of the team.	0.71
			INT5	The members of my team are always willing to ask, offer and receive help, voluntarily, among themselves.	0.73
			INT6	My team is willing to give and receive feedbacks consistently among themselves.	0.74
			INT10	My team is capable to solve interpersonal conflicts among its members.	0.67

Team Composition					
Factor	**ρ**	**Rho_{cv}**	**Items**	**Description**	**Factorial Load**
Technical Skills (F4)	0.80	0.50	TEC4	My team has the necessary knowledge about the methods, practices and tools used to accomplish their activities.	0.66
			TEC5	My team can establish technical solutions without the need for direct interference from the manager and the software architect.	0.69
			TEC6	My team can understand and give a solution to a technical problem in a manner appropriate to the client and the project.	0.76
			TEC7	My team can solve the technical problems inherent in software development with the participation of all its members.	0.71
Management Skills (F7)	0.84	0.72	MNG1	My team can keep the visibility on the progress of their work to all stakeholders (for example: manager, clients, etc.).	0.84
			MNG2	My team can follow and monitor their work continuously.	0.87
Heterogeneity (F5)	0.75	0.50	HET1	My team members have diverse areas of expertise in software development.	0.70
			HET2	My team members have different backgrounds and software development experiences.	0.74
			HET3	My team members have competencies that complement each other.	0.69
Team Maturity (F6)	0.65	0.50	TEW4	My team members lack experience in working in a more autonomous way, without the interference of the manager and the software architect, for example.	0.45
			TEC8	My staff needs constant interference from the manager and the software architect to be able to perform their activities.	0.89
Experience in the Organization (F9)	0.75	0.61	EOR1	My team members have already worked together on other software development projects in the Organization where we work.	0.68
			EOR2	My team members have already participated in other software development projects, in other teams, in the Organization where we work.	0.86

Team Composition					
Factor	**ρ**	**Rho_{cv}**	**Items**	**Description**	**Factorial Load**
Role and Goal Clarity (F2)	0.80	0.57	CGR2	My team's goals are shared and accepted by all its members.	0.74
			CGR4	The roles and responsibilities of my team members are accepted by all.	0.72
			NOR1	My team's work process is established and aligned with its needs	0.80
Formal Leadership (F3)	0.87	0.62	LEAD1	The manager (or immediate manager) gives freedom and encourages my team to make their own decisions about work.	0.72
			LEAD2	The manager (or immediate manager) stimulates my team to work autonomously.	0.67
			LEAD4	The manager (or immediate manager) plays a facilitating role in the decision-making and in the execution of my team's work.	0.86
			LEAD5	The manager (or immediate manager) encourages my team to solve their own problems.	0.89

Team Composition					
Factor	**ρ**	**Rho_{cv}**	**Items**	**Description**	**Factorial Load**
Rules of Behavior (F8)	0.82	0.70	NOR3	It is clear what is and what is not an acceptable behavior for the members of my team.	0.80
			NOR4	My team members agree on how they should behave.	0.87

In Table 30, the values that show high significance of all relations between factors and items are shown, where the values of C.R. > 1.96 and p-value values are all below 0.001 (***).

Table 30. Reliability and Validity of Factors - The Model of Teamwork Process Antecedents (TPA) Questionnaire.

		Non-Standardized Coefficient	**Standard Error**	**C.R.**	**P-Value**	**Label**
INT5	Interpersonal Skills	1.000				
INT3	Interpersonal Skills	0.848	0.071	11.887	***	par_1
INT2	Interpersonal Skills	0.805	0.065	12.453	***	par_2
INT6	Interpersonal Skills	1.206	0.095	12.757	***	par_3
INT10	Interpersonal Skills	0.889	0.078	11.381	***	par_4
LEAD5	Formal Leadership	1.000				
LEAD4	Formal Leadership	1.040	0.089	11.661	***	par_5
LEAD2	Formal Leadership	1.364	0.095	14.427	***	par_6
LEAD1	Formal Leadership	1.395	0.094	14.878	***	par_7
CGR2	Role and Goal Clarity	0.867	0.066	13.099	***	par_8
CGR4	Role and Goal Clarity	0.831	0.065	12.795	***	par_9
NOR1	Role and Goal Clarity	1.000				
TEC4	Technical Skills	1.000				
TEC5	Technical Skills	1.446	0.141	10.224	***	par_10
TEC6	Technical Skills	1.179	0.103	11.418	***	par_11
TEC7	Technical Skills	1.279	0.123	10.436	***	par_12
MNG1	Management Skills	1.000				
MNG2	Management Skills	0.985	0.069	14.231	***	par_13
EOR1	Experience in the Organization	1.000				
EOR2	Experience in the Organization	1.127	0.231	4.887	***	par_14
TEC8	Team Maturity	1.000				
TEW4	Team Maturity	1.992	0.365	5.456	***	par_15
HET1	Heterogeneity	1.000				
HET2	Heterogeneity	0.965	0.091	10.630	***	par_16
HET3	Heterogeneity	0.787	0.088	8.942	***	par_17
NOR3	Rules of Behavior	1.000				
NOR4	Rules of Behavior	0.950	0.072	13.179	***	par_18

Legend: $^{***}p < 0{,}001$