A Comparative Analysis of Agile Teamwork Quality Measurement Models

Manuel Silva, Mirko Perkusich, Arthur Freire, Danyllo Albuquerque, Kyller Costa Gorgônio, Hyggo Almeida, Angelo Perkusich, *Member, IEEE,* and Everton Guimarães

***Abstract*—Multiple models (or instruments) for measuring Teamwork Quality (TWQ) for Agile Software Development can be found in the literature. Regardless, such models have different constructs and measures, with no empirical evidence for comparing them. This study analyzed two agile TWQ models, resulting in equivalent results. We mapped the models' variables given their definitions. We then collected data using both a Bayesian Network model, namely the TWQ-BN model, and Structural Equation Modeling, namely the TWQ-SEM model. We interviewed 162 team members from two software development companies. We analyzed the data using the Bland-Altman method. We obtained enough evidence to conclude that the results for *Communication*, *Coordination*, *Cohesion* and *Mutual Support* variables are not equivalent. On the other hand, we did not have enough evidence to claim that the models do not agree for measuring *Effort* and *Balance of member contribution* variables. The results of this study detail how two state-of-the-art agile TWQs compare in terms of their measures as well as potential research areas for further investigation.**

***Index Terms*—Teamwork Quality Models, Empirical Study, Comparative Analysis, Agile Software Development.**

I. Introduction

In Agile Software Development (ASD) software teams are the critical source of agility [1], [2]. Hence, teamwork quality (TWQ) is critical for agile projects' success [3], [4], [5]. The industry is rapidly adopting ASD [6], [7], and the need for systematic team development [8] compelled researchers to focus more on teamwork aspects. As pointed out by Cruz et al. in [9], more than 70% of the studies about personality in the last four decades were published after 2002.

To perform the TWQ assessment, different models have been used, such as: Teamwork Process Antecedents (TPA) questionnaire [10], agile Team Work Quality (aTWQ) [8], Bayesian Networks (BN) [11]; Structural Equation Modeling (SEM) [12], [3]; Radar Plot [13]; and System Dynamics [14]. The main differences among such models are the variables that make up the agile TWQ construct. A major issue is the absence of studies that analyze and empirically compare these existing models.

To address this gap, we performed a comparative analysis of the models proposed by Freire et al. [11], namely Team Work Quality based on Bayesian Networks (the TWQ-BN model), and Lindsjørn et al. [3], namely Team Work Quality based on Structural Equation Modeling, the TWQ-SEM model. We focused on these two models due to the cost and effort of conducting an empirical study such as this one. Besides, these two models have been empirically evaluated in the industry published in a high-standard venue, and the proposed models focus on the TWQ construct, considering the causality between factors.

We first mapped the variables, namely *Communication*, *Coordination*, *Cohesion*, *Mutual Support*, *Effort* and *Balance of member contribution* according to the authors' definitions. Then, we provided them with data collected using specific questionnaires as defined by the authors. We had 162 participants from various roles from 25 software teams, including but not limited to developers, Quality Assurance, Scrum Masters, technical leaders, and managers. We compared the results of the models for the previously associated variables using the Bland-Altman method [15].

We extend the work of Silva et al. [16] with additional contributions, including (i) details about how we identified the theoretical mapping between both instruments' variables; (ii) details about the questionnaires used to collect data; (iii) more detailed information about the teams and subjects; (iv) the results associated to each model; (v) more granular discussion at the level of metrics (questions); (vi) implications of the research and practice; (vii) more in-depth discussion about threats to the validity of the present study. Given this, this article eases replicating the study and presents new findings.

This paper elucidates the similarities between the TWQ-BN and TWQ-SEM models and examines the study's implications for research and practice. Section II presents the required background information for the TWQ-BN model and TWQ-SEM models, as well as the mapping of variables between models. Section III presents the design of the comparative analysis study performed. Section IV presents the results and discusses the study's research questions and the implications for research and practice. Section V discusses the study's threats to validity. Finally, section VI presents our concluding remarks and future work.

Manuscript received February 3, 2022; revised March 6, 2022. Date of publication May 10, 2022. Date of current version May 10, 2022. The associate editor prof. Dinko Begušić has been coordinating the review of this manuscript and approved it for publication.

The part of this paper was presented at the International Conference on Software, Telecommunications and Computer Networks (SoftCOM) 2021.

M. Silva, M. Perkusich, A. Freire, D. Albuquerque, K. Gorgônio, H. Almeida and A. Perkusich are with the Intelligent Software Engineering Group (ISE/VIRTUS), Federal University of Campina, Grande, Brazil, contact: perkusic@virtus.ufcg.edu.br.

E. Guimarães is with the Pennsylvania State University, Malvern, USA, ezt157@psu.edu.

Digital Object Identifier (DOI): 10.24138/jcomss-2021-0177

II. Background

This section presents background information related to the TWQ-SEM model (Section II-A) and the TWQ-BN model, Section II-B, as well the mapping carried out between the models' variables (Section II-C).

A. Structural Equation Modeling Based model (TWQ-SEM model)

Hoegl and Gemuenden [12] introduced the concept of team collaboration Teamwork Quality (TWQ) with six facets or variables, as shown in Table I: *Communication*, *Coordination*, *Balance of member contribution*, *Mutual support*, *Effort*, and *Cohesion*. They empirically analyzed how the proposed model relates to project success using SEM, and the variables' values were obtained based on a questionnaire. Three to 10 questions were associated with each variable with a total of 61 questions and responses as abased on a 5-points Likert scale. The results demonstrated that TWQ is significantly associated with team performance, and TWQ has a strong association with the personal success of team members.

Lindsjørn et al. [3] replicated Hoegl and Gemuenden's[12] study focusing on ASD instead of on a traditional software development paradigm. As a result, they concluded that the TWQ is a significant factor in improving team performance, especially for the product's quality. They proposed a way to measure TWQ based on SEM, having the high-order factor as the dependent variable, and the construct facets as independent variables. They empirically verified whether the proposed model produced a covariance matrix consistent with the sample covariance matrix. For this purpose, they collected data using the questionnaire proposed by Hoegl and Gemuenden. Figure 1 shows the resulting calculated relationship between TWQ and its corresponding variables. The arrows represent the standardized factorial loads for each construct and show the variation explained by the variable in the TWQ. In the SEM approach, as a general rule, a factorial load of 0.7 or higher represents that the factor extracts sufficient variation from this variable.

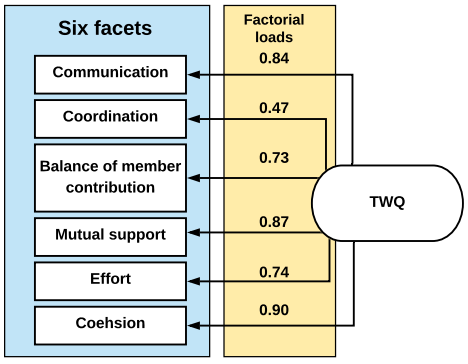


Fig. 1. Structural Equation Modeling-Based Teamwork Quality Model [16] (TWQ-SEM model).

We considered the variables directly related to the TWQ, and disregarded the project success variables. It is worth mentioning that Lindsjørn et al. concluded that there is a need for more research efforts to validate the TWQ construct and its measurement.

The data source for the TWQ-SEM model is a questionnaire, where the observable variables (OV) are considered metrics, and each one is linked to the questions in the questionnaire. However, the relation between questions to a given OV for the TWQ-SEM model is many-to-one. Given this, the authors defined that the value for a given OV is the arithmetic mean of the responses for the set of questions related to it.

The OV effort, for instance, has the following four questions related to it:

1) Every team member fully pushes the teamwork;
2) Every team member makes the teamwork their highest priority;
3) The team put(s) much effort into the teamwork;
4) There are conflicts regarding the effort that team members put into the teamwork.

Given that the possible responses for such questions are on a scale from 1 (strongly disagree) to 5 (strongly agree), if the responses for the previous questions are 4, 4, 3, and 4, the result for this variable would be 3.75.

B. Bayesian Networks Based Model (the TWQ-BN model)

Freire et al. [11] presented a Bayesian Network (BN) to assess and improve the TWQ in the context of ASD. BNs are probabilistic graph models that describe knowledge about an uncertain domain [17]. A BN, B, is a directed acyclic graph symbolizing a joint probability distribution over a set of random variables V. The network is defined by the pair $B = \{G, \Theta\}$. G is the directed acyclic graph in which the nodes $X_1, \ldots, X_n$ are random variables, and the arcs represent the direct dependencies between these variables. Thus, a BN is a directed acyclic graph and the probability distributions.

Hence, to construct a TWQ-BN model, the authors identified graph nodes by analyzing the literature on agile teamwork. They relied on experts' knowledge to identify the relationship between the variables, the arrows on the graph, quantified them, and defined the probability distributions for each node.

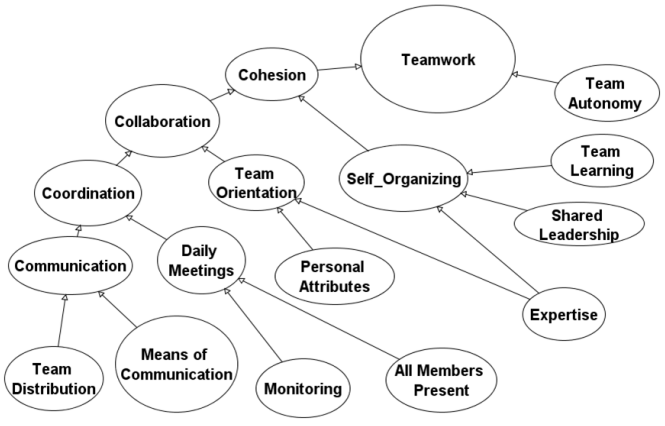


Fig. 2. Teamwork Quality based on Bayesian Network (the TWQ-BN model) [11].

TABLE I
DESCRIPTION OF TWQ-SEM MODEL VARIABLES.

Variable	Description
Communication	Frequency, formalization, and openness of the information exchange
Coordination	Common understanding when working on parallel subtasks, and agreement uncommon work-down structures, schedules,budgets, and deliverables
Cohesion	Team members' motivation to maintain the team and accept that team goals are more important than individual goals
Balance of Members Contribution	The ability to employ the team members' expertise to its full potential. Contributions should reflect the team member's specific knowledge and experience
Effort	Team members' ability and willingness t share workload and prioritize the teams' task over other obligations
Mutual Support	Team members' ability and willingness to help and support each other in carrying out their tasks

TABLE II
DESCRIPTION OF TWQ-BN MODEL VARIABLES.

Variable	Description
Communication	Information sharing across team members
Coordination	Refers to the tasks execution by team members in a synchronized and integrated manner
Cohesion	Interpersonal attraction between the team members, their commitment to the team tasks, to the team itself, and group pride spirit
Collaboration	Refers to the commitment that the team members have with each other to achieve the common goals specific knowledge and experience
Shared Leadership	The authority on the decision-making process is shared
Team Orientation	Refers to the respect that the team member shave with each other
Team Autonomy	Refers to the influence external agents have on the team tasks execution
Team Learning	The team's capability to identify changes on its environment and adjust its strategies as necessary
Monitoring	Team synchronization regarding the tasks and barriers
Team Distribution	The team's physical distribution
All Members Present	All members attended to the daily meetings
Means of Communication	The team members communicate face-to-face
Daily Meeting	Daily meeting to synchronize the team
Expertise	Team members' knowledge to develop tasks with redundancy
Self-Organizing	Team's capability to self-organize efficiently in order to face challenges and complex changes
Personal attributes	The mix of personalities

A key characteristic of the constructed BN, which influenced our study's design, see Section III-D, is that it is composed only by ranked nodes. A ranked node consists of an ordinal scale mapped into a scale monotonically ordered in the interval $[0, 1]$. The solution is based on a Normal distribution truncated between $[0, 1]$, TNormal, to represent the probability function. Thus, the probability function of a child node is a TNormal calculated as the mixture of the TNormals of its parent nodes. There are four expressions to model the mixture's mean (μ): weighted mean (*WMEAN*), weighted minimum (*WMIN*), weighted maximum (*WMAX*), and the mixture of the classic minimum and maximum functions (*MIXMINMAX*). For more information regarding the ranked node method, refer to Fenton et al. [18].

The BN uses a converging star topology [19], in which there are links directed from each teamwork variable (i.e., predictor) to a single dependent variable, the *Teamwork* node. From here on, we refer to the nodes with no parents (i.e., no arrows pointed towards them) as leaf nodes. We considered the BN graph shown in Figure 2 as a theoretical construct of agile teamwork and the variables as defined in Tabel II.

A user can employ the BN for two purposes: prognosis and diagnosis. For prognosis, the user enters data (i.e., evidence) into the leaf nodes, and a tool calculates the probability for each variable. The authors used AgenaRisk[1] to perform the inferences. The most frequent use case is the users having the goal to compute the probabilities for the *Teamwork* node given, the collected data, used as evidence in the leaf nodes. The leaf nodes might be considered metrics in some sense, as they are linked to data sources. In this case, the BN is connected to a questionnaire, having each leaf node associated with one question instead of the TWQ-SEM model, which is many-to-one. Thus, the users should use the questionnaire for data collection purposes when providing their answers. Section III-D presents how we managed to treat data for inputting them as evidence for the leaf nodes having multiple questionnaire answers per team.

However, BNs have the characteristic of treating all variables as observable [20] (i.e., the users may input data into any variable). Therefore, the users can also use the model to diagnose teamwork by inputting data into the *Teamwork* node (or any other non-leaf node) and observing the impact on the leaf nodes. Along with different types of analysis, such a task might support managerial decisions regarding how to improve teamwork.

The authors evaluated the model and a procedure to apply it in a case study, and concluded:

[1]https://www.agenarisk.com

TABLE III
MAPPING OF VARIABLES FROM TWQ-BN AND TWQ-SEM MODELS.

TWQ-BN	TWQ-SEM
Communication	Communication
Coordination	Coordination
Cohesion	Cohesion
Team Orientation	Balance of Members Contribution
Collaboration	Effort
Self-Organizing	Mutual Support

1) The model helped identify opportunities for improvement and assess the quality of teamwork;
2) The cost-benefit of using the model applying the process was positive;
3) The process was easy to learn and implement in the team routine.

C. Equivalence Between the Variables of the TWQ Models

As shown in Table III,to compare the models, we defined how to map the variables of each model according to the similarity of the definitions. Note that while the TWQ-BN model contains 17 variables, the TWQ-SEM model contains only seven. Since the TWQ-SEM model has fewer variables, we focused on mapping them to the TWQ-BN model ones.

We have a direct map for the variables *Communication*, *Coordination*, *Cohesion*. In other cases, the variables have different names but similar definitions. We mapped the variable *Balance of member contribution* for the TWQ-SEM model to *Team orientation* for the TWQ-BN model. Analyzing the TWQ-BN model variables individually, we identified that it was not possible to map *Balance of member contribution* variable to a single TWQ-BN model variable. Thus we considered the union variables *Personal attributes*, *Expertise*, and *Team Orientation* for the TWQ-BN model *Team* variable. Given that, *Team Orientation* is the child of *Personal Attributes* and *Expertise*.

The variables *Effort*, for the TWQ-SEM model and *Collaboration*, for the TWQ-BN model, were mapped as both reflect team members' willingness to accomplish the team's goals. *Effort* variable is related to sharing the workload and prioritizing the teams' tasks. *Collaboration* variable reflects the commitment of team members to achieve common goals. Finally, we mapped the variable *Mutual Support* for the TWQ-SEM model to *Self-Organizing* variable for the TWQ-BN model because both describe team members' ability to organize themselves to achieve common goals.

After mapping all the variables for both models, we observed that some variables for the TWQ-BN model had no similarity with variables for the TWQ-SEM model. *Team Autonomy* variable is identified by Eloranta et al. [21] as a critical factor to keep the team motivated, but it is not addressed for the TWQ-SEM model. *Daily meetings* variable, which is a prevalent work synchronization practice adopted by the majority of agile teams [22], is also not addressed for the TWQ-SEM model.

TABLE IV
STUDY HYPOTHESES

H	Description
H1	The results from both models are equivalent for the variable Communication
H2	The results from both models are equivalent for the variable Coordination
H3	The results from both models are equivalent for the variable Cohesion
H4	The results from both models are equivalent for the variable Balance of Member Contribution
H5	The results from both models are equivalent for the variable Effort
H6	The results from both models are equivalent for the variable Mutual Support

III. RESEARCH METHODOLOGY

This study focuses on comparing the equivalence between TWQ-BN and TWQ-SEM models. We structured the research in terms of the theoretical equivalence between the variables as discussed in Section II-C. In our study six dependent variables are considered, namely: *Communication*, *Coordination*, *Cohesion*, *Balance of Member Contribuion*, *Effort*, and *Mutual Support*. Observe that we did not consider the overall TWQ as a dependent variable because, for the TWQ-SEM model, it is a latent variable. Thus, the model does not calculate a value for this variable. Besides, we considered the inputs for each of the analyzed TWQ models as independent variables.

As a result, we defined a hypothesis for each dependent variable, each one claiming that both models' results are equivalent for the given variable. Table IV shows this study's hypotheses.

We verified the hypotheses by collecting data from two software development companies and analyzing them using the Bland-Altman method [15]. This method combines Student's T-test, confidence interval, and linear regression to analyze the level of agreement between two models. The Bland-Altman method has been extensively used in the medical domain to compare the agreement between two measurement methods and has been claimed to be more robust than the traditional statistical measures such as kappa statistic, correlation coefficient, and means and ranks comparisons [23], [24].

In what follows, Sections III-A, III-B, III-C, and III-D describe, respectively, the study's subjects, instrumentation, data collection procedures and data analysis procedures.

A. Subjects

The recruiting process started by selecting two of our research group's industry partners to collect data from their teams. From here on, we refer to them as Organization A, and Organization B. Organization A executes research, development, and technological innovation projects with industry partners. Organization B performs research, development, and technological innovation work in Data Analysis, including Data Mining, Machine Learning, and Data Visualization.

Then, the project managers were closely involved in the study by explaining its goals and how it could benefit them and that the data would be kept anonymous. Further, we

TABLE V
ROLES OF SOFTWARE PROFESSIONALS.

Role	Number of professionals
Developer	69 (42.59%
QA	38 (23.46%)
Scrum Master	8 (4.94%)
Technical Leader	25 (15.43%)
Manager	21 (12.96)

explained that the original data, which mapped the project identification with the collected data, would be destroyed after the analysis. Then, the project managers negotiated with their teams' members. We restricted participation in the study for teams that everybody agreed to participate because considering answers from only part of a team could bias our results. As a result of the recruiting process, all the invited teams agreed to participate: 24 from Organization A and one from Organization B, totaling 162 subjects.

Scrum was applied to manage all projects, and developers applied software development practices given their domain. Scrum events were held, including the Daily Scrum, Sprint Planning, Sprint Review, and Sprint Retrospective. The length of the Sprints was between two to three weeks.

The subjects worked in different roles, developers, quality analysts, technical leaders, Scrum Masters, and managers. Managers take the lead in all phases and activities, including project planning, management, monitoring, and closing. They are responsible for the communication between the customer and the Scrum Team. Scrum Masters act as facilitators and coaches for teams and facilitate the removal of impediments. The team leaders work closer to the managers and ensure that the products are delivered on time and within the specified quality. Both Scrum Masters and team leaders execute the work to deliver the product, cooperating with the developers and quality analysts. Table V shows the number of respondents in each role. Additional about relationship per team is available in the Supplementary Material. Figure 3 shows more information about the subjects.

B. Data Collection Instruments

We used questionnaires to have data about each of the studied models' variables since both models applied the data collection instrument. For gathering data for the TWQ-BN model, we adopted the questionnaire presented by Freire et al. [11]. Altogether, the questionnaire had nine questions, one for each the TWQ-BN model's input nodes.

For the TWQ-SEM model, we adopted the questionnaire presented by Lindsjørn et al. [3], but including only the questions related to the TWQ construct, resulting in 38 questions. For the TWQ-SEM model and the TWQ-BN model questionnaires, we modeled the answers with a 5-points Likert scale: 1 (strongly disagree) - 5 (strongly agree).

To compare the answers of each questionnaire individually, preserving the respondents' anonymity, we generated a unique ID for each respondent. In addition to the questionnaires mentioned above, we also collected demographic data through a questionnaire, including project ID, role, age, and experience.

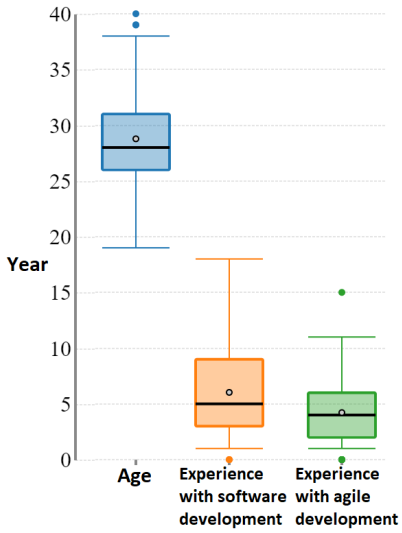


Fig. 3. Demographic data for the subjects

Further, we also included a consent form to comply with ethical principles. The form contained information about the study's goals, participants' rights, and our research team's accountabilities. The questionnaires are in an online Supplementary Material [2].

C. Data Collection Procedure

The subjects were not familiar with the concepts and terminology for the TWQ-BN model and TWQ-SEM models, so we collected data through in-person data collection sessions. We ran a session per team. The sessions were held in the rooms where the participants performed their project activities and took place right after the Sprint Retrospective meetings. We selected this moment because the respondents had just reflected on the results of the previous Sprint.

Each session was structured into two parts: leveling and data collection. We reinforced the study's goals during the leveling and motivated the participants to answer questions based on reality (not intentions). Further, we explained both models and their associated questionnaires. The leveling lasted around ten minutes for each session. Then, we had the participants answer the questionnaires on their computers independently. The data collection lasted around twenty minutes for each session.

Notice that all the teams that participated in this study were static, which means they remained the same for multiple Sprints. Given this context, we can assume that each individual has developed a certain level of consciousness that enables them to reflect on their perceived level of teamwork quality [25].

To avoid the effect of learning bias, we crossed the sequence in which the TWQ-BN and the TWQ-SEM models were answered by the participants, having half the team members answer the TWQ-BN model rather than the TWQ-SEM model, and the other half the TWQ-SEM model than the TWQ-BN model. In the cases of a team having an odd number

[2]https://zenodo.org/record/4763135#.YJ7JjqhKiUk

of members, we randomly chose which questionnaire would be answered first by one of the team members. Further, a researcher was present during the sessions to prevent the participants from exchanging information. Further, a researcher was instructed to interrupt the respondents if they felt tired, but this issue did not happen.

D. Data Analysis Procedure

We registered the collected data from the questionnaires with spreadsheets. We mapped the questionnaire's answers into inputs for the models to analyze the results. For the TWQ-SEM model, we followed the method proposed by Hoegl and Gemuenden [12] and Lindsjørn et al. [3]. First, we mapped the responses from the 5-points Likert scale into an integer scale [1, 5], having the lower values on the Likert scale mapping to the smaller values on the integer scale.

Then, we calculated a score for the TWQ-SEM model variables for each subject. For this purpose, for each variable, we took the arithmetic mean of the answers to the questions related to the given variable. For instance, the variable *Effort* has four questions associated with it. Thus, if a given subject answers $5, 5, 4, 4$, his/her score for *Effort* is 4.5. After applying this procedure for each subject, we had their scores for all the TWQ-SEM model variables.

Later, we calculated the teams' scores for TWQ-SEM variables. For this purpose, we calculated a team's scores by, for each variable, taking the arithmetic mean of the answers of its members. For instance, let us assume that a given team has five members with the scores $4.5, 5, 4, 4, 3.5$ for *Effort*. As a result, the given team's score for *Effort* is *4.2*.

It is worth mentioning that the collected data was originally in an ordinal (Likert) scale. We are aware that some statisticians disagree about using the arithmetic mean for ordinal scales [26]. However, we followed the methods provided by the original authors of the TWQ-SEM model.

For the TWQ-BN model, the questionnaire contained one question for each Bayesian network's leaf nodes. Therefore, to calculate the other variables' values, we executed the Bayesian Network using the AgenaRisk toolbecause it was the one used by Freire et al. [11].

To enter data into the Bayesian network, we mapped the answers from the questionnaire into the values for each associated random variable of the Bayesian network. For this purpose, we defined a rule to aggregate the answers from a team. The rule consisted of counting the number of answers for each possible answer (i.e., *Very Low*, *Low*, *Medium*, *High*, *Very High*). For instance, consider that for a team with five members, we had two answers *Low* and three *Medium* for the variable *Expertise*. Given this, the values for *Expertise* would be [*Very Low* = 0, *Low* = 2, *Medium* = 3, *High* = 0, *Very High* = 0]. Since, in Bayesian networks, the values of a given variable are represented as a probability function that must add up to one, AgenaRisk automatically normalizes the values. Thus, for the given example, the probability function of *Expertise* would be [*Very Low* = 0, *Low* = 0.4, *Medium* = 0.6, *High* = 0, *Very High* = 0].

After defining the procedure to come up with the results for the models, we determined the procedure to analyze the calculated data. For the TWQ-BN model, the variables are based on the ranked nodes method, as discussed in Section II-B. Therefore, each variable for the TWQ-BN model is a random variable with values between [0, 1] [18]. By contrast, the data calculated for the TWQ-SEM model is within the range [1, 5]. To allow comparing data from both models, we normalized the data from TWQ-SEM by applying Equation 1, where x_i represents the value of a given variable, *min(x)*, and *max(x)* represent the minimum and maximum values of the scale, respectively.

$$Normalized.Value(x) = \frac{x_i - min(x)}{max(x) - min(x)} \quad (1)$$

The next step was to define how to assess the equivalence between both models for each of this study's dependent variables. As discussed previously, we verified each case using the Bland-Altman method. The Bland-Altman method focuses on studying the differences, not the agreement or correlation [27]. We chose this method because it is the most recommended statistical approach to assess the agreement between two quantitative measurement methods. It is based on four steps, as shown in Figure 4. We detail how we applied such steps in what follows.

1) Step 1: Let b_{k_i} be the *ith* team value for the *kth* dependent variable for the TWQ-BN model and s_{k_i} be the *ith* team value for the *kth* dependent variable for the TWQ-SEM model. Thus, we have the sets $b_k = \{b_{k_1}, b_{k_2}...b_{k_|t|}\}$ and $s_k = \{s_{k_1}, s_{k_2}...s_{k_|t|}\}$, where t is a team for which data was collected. Given this, for each dependent variable k, we calculated the differences $d_{k_i} = b_{k_i} - s_{k_i}\ \forall i \in t$. Then, we calculated the mean (μ_{d_k}) and standard deviation ($\sigma^2_{d_k}$) of these differences. Further, we calculated the mean of the two paired values for each team: $m_{k_i} = (b_{k_i} - s_{k_i})/2\ \forall i \in t$.

2) Step 2: For each dependent variable k, we performed the Student's t-test for the differences d_k with a significance level $\alpha = 0.5$. For each dependent variable k, the null hypothesis was $d_k = 0$. For *p*-value ≥ 0.05, we failed to reject the null hypothesis; thus, we concluded that there is not sufficient evidence to support a conclusion that the models disagree for k. For *p*-value ≤ 0.05, we rejected the null hypothesis; thus, we concluded that the models disagree for k;

3) Step 3: In this step, we plot and analyze the Bland-Altman graph plots. The Bland-Altman graph plot consists of a scatter plot XY representing every difference between two paired methods against the average of the measurement. Thus, for each dependent variable k, we created a Bland-Altman graph plot with a Y-axis showing the difference d_k and the X-axis representing the mean m_k. To support our conclusions regarding the degree of agreement between the models, we drew three lines on the graph. The first one represents a "true value" (gray line in the scatter plot shown in Figure 4). Since we do not know the "true value", the mean of both measurements is the best estimate we have [28]. Thus, in Figure 4, the gray line represents μ_{d_k}. The other two lines represent the upper and lower limits of agreement, respectively, u_k and l_k (red lines in the scatter plot shown in Figure 4). Since we used a significance level $\alpha = 0.5$, $u_k = \mu_{d_k} + 1.96 * \sigma^2_{d_k}$ and $u_k = \mu_{d_k} - 1.96 * \sigma^2_{d_k}$. The

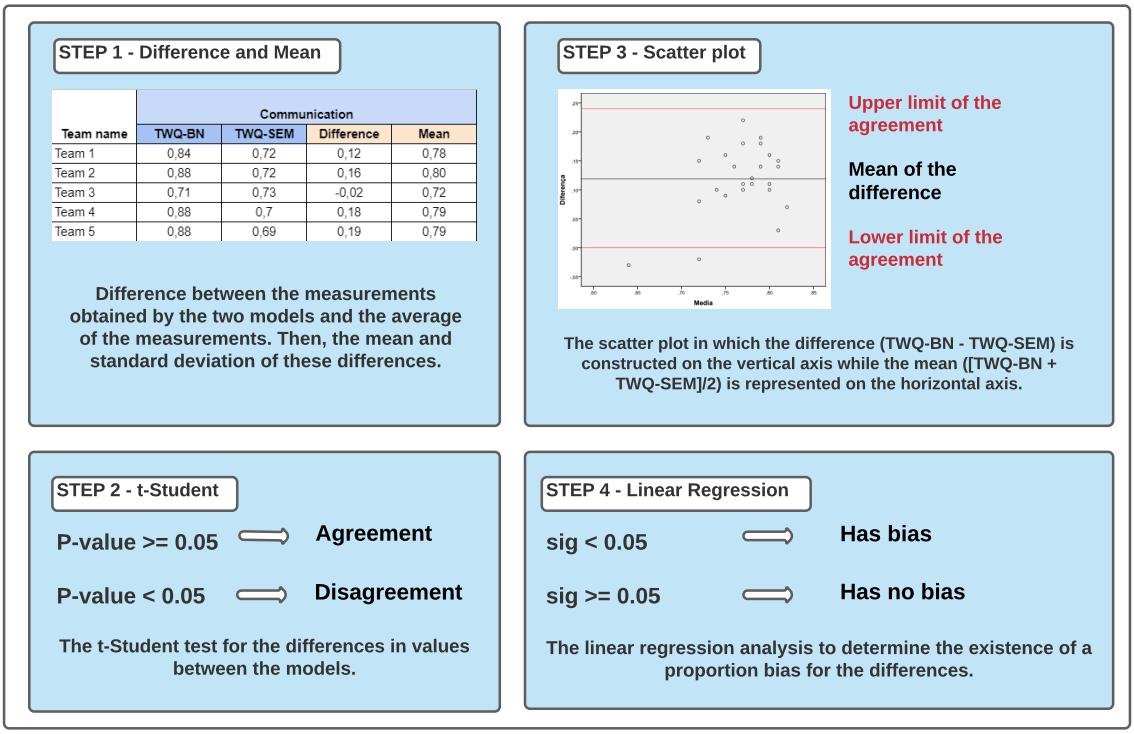


Fig. 4. Details of the employed Bland-Altman method.

interval between u_k and l_k is the 95% confidence interval. In the Bland-Altman plot, if the line representing zero difference falls outside the 95% confidence interval, there is a significant difference between the measurements, which means that one model overestimates or underestimates the other. However, if the line that represents zero difference falls within the 95% confidence interval and close to the line for μ_{d_k} and most data points also fall within the 95% confidence interval, it means that the models have a relatively high degree of agreement;

4) Step 4: For each dependent variable k, we performed a linear regression analysis to determine the existence of a proportion bias for the differences. We used d_k as the dependent variable for the regression analysis and m_k as the independent variable. For significance levels < 0.05, we concluded that the different values tend to be higher or lower than the average; in other words, this result indicates that the models disagree. For significance levels ≥ 0.05, this trend does not exist.

We executed the data analysis using SPSS, and the SPSS files, the AgenaRisk model file, the collected data, and the spreadsheets used in this study can be found online as Supplementary Material [3].

IV. Results and Discussion

This section presents and discusses the results of this study's hypotheses (Section IV-A) as well as the study's implications for research and practice (Section IV-B).

[3]https://zenodo.org/record/4763135#.YJ7JjqhKiUk

A. Are the Results from Both Models Equivalent?

Table VI presents a summary of our results for Student's t-test and linear regression and Figure 5 shows the Bland-Altman graph plots for each dependent variable. As shown in Table VI, for the Student's t-test, we only failed to reject the null hypothesis for *Effort* (marked in green). With regards to the regression analysis, the results only indicate a proportion bias for the *Cohesion* variable (marked in red).

Figure 5 shows that the zero difference line is within the 95% for *Communication*, *Coordination*, *Balance of Member Contribution*, and *Effort*. Further, we have more insights regarding the degree of agreement between the models, specifically, the bias and the percentage of values within the 95% confidence interval. In this case, the smaller the bias and percentage of values within the 95% confidence interval, the more the models agree. For *Communication*, we have a bias of 0.12 and that 8% of the values fell outside of the confidence interval. For *Coordination*, we have a bias of 0.07 and that 8% of the values fell outside of the confidence interval. For *Cohesion*, we have a bias of -0.09 and that 8% of the values fell outside of the confidence interval. For *Balance of Members Contribution*, we have a bias of 0.06 and that 4% of the values fell outside of the confidence interval. For *Effort*, we have a bias of 0.02 and that 8% of the values fell outside of the confidence interval. For *Mutual Support*, we have a bias of -0.24 and that all values fell within the confidence interval. Next, we discuss the results for each variable individually.

For the *Communication* variable we obtained a *p*-value < 0.05, indicating that it is not equivalent in both models.

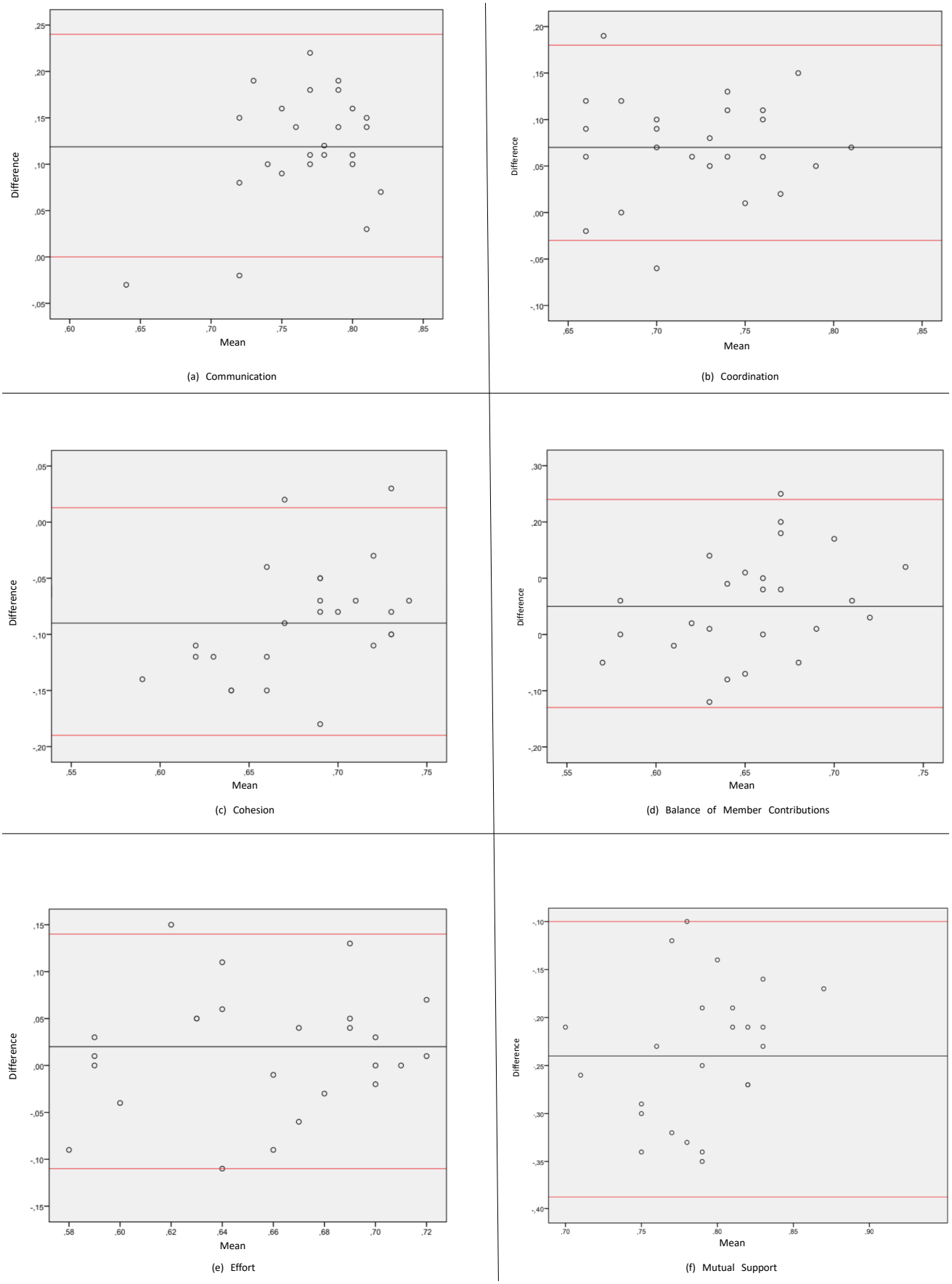


Fig. 5. Bland-Altman graph plots for the models investigated. The red lines represent 95% confidence intervals. The gray line represent the mean of the paired values, in other words, the best estimate for the "true value" [16].

TABLE VI
RESULTS FOR THE STUDENT'S T-TEST AND LINEAR REGRESSION.

Variable	Student's T-test p-value	Linear Regression Significance Level
Communication	0.000	0.050
Coordination	0.000	0.920
Cohesion	0.000	0.024
Balance of Members Contribution	0.010	0.072
Effort	0.255	0.536
Mutual Support	0.000	0.220

Furthermore, looking at Figure 5, we can see that the line for the means of the differences is far from zero (bias = 0.12). Besides, the *Communication* variable presented a significance of 0.05, indicating no proportion bias. Despite having similar definitions, the *Communication* variable showed significantly different results between the models' evaluations. Such difference might be explained because the assessment for the TWQ-SEM model does not take into account aspects related to the team's geographical distribution. Conversely, for the TWQ-BN model, the *Communication* variable was evaluated based on *Team distribution* and *Means of communication*, which refer to sharing the same work environment and face-to-face communication between team members.

Further, while TWQ-BN has two questions related to *Communication*, TWQ-SEM has ten questions. Of the ten questions, only one is related to the personal communication between members, which means that TWQ-SEM gives emphasies to communication factors not considered by TWQ-BN. As a consequence of the analysis, we rejected H1.

For the *Coordination* variable we obtained a p-value < 0.05, indicating that it is not equivalent in both models. Also, the bias is far from zero (.07). The significance for *Coordination* variable is greater 0.05, thus indicating no proportional bias. Note that for the TWQ-BN model, the *Coordination* variable has the *Communication* and *Daily Meetings* variables as its parent nodes. Further, notice that the *Communication* variable presented divergences in its results, which might explain the divergences for the *Coordination* variable. Thus, we rejected H2.

The p-value for variable *Cohesion* < 0.05, indicating that it is not equivalent in both models. Furthermore, looking at Figure 5, we can see that the average of the differences is far from zero (bias = -0.09). Finally, the significance of the *Cohesion* variable is < 0.05, indicating a proportional bias. Such divergence may be related to the spread of the divergence presented for *Self-organizing* variable, the father of *Cohesion* variable for the TWQ-BN model. Thus, we rejected H3.

For the *Effort* variable we obtained a p-value ≥ 0.05, which means that we do not have enough evidence to claim that they are not equivalent. Furthermore, looking at Figure 5, we can see that the line that represented the means of the differences is near zero (bias = 0.02)and 92% of the values fall within the confidence interval. Finally, *Effort* presented a significance ≥ 0.05, indicating that there is no proportion bias. For the TWQ-SEM model, this variable is measured by how each team member focuses on TWQ. For the TWQ-BN model, this variable is related to the *Collaboration* node, which has *Coordination* and *Team Orientation* as parent nodes. Thus, we failed to reject H4.

For the *Balance of Members Contribution* variable we obtained a p-value ¡ 0.05, indicating that it is not equivalent in both models. However, Figure 5 shows that the zero difference line is within the 95% confidence interval, the bias is small (bias = 0.06) and that 96% of the points fell within the confidence interval. Finally, it presented a significance ≥ 0.05, indicating no proportion bias. For the TWQ-BN model, this variable relates to *Team Orientation*, which has as parents *Personal Attributes* and *Expertise*. Analyzing the TWQ-SEM model questions for *Balance of Members Contribution* they cover aspects similar to *Personal Attributes* and *Expertise*. *Personal Attributes* and *Expertise* cover, respectively, the relationship between the team members and the knowledge redundancy, which is a critical factor for enabling self-managed teams. The TWQ-SEM model questions for *Balance of Members Contribution* focus on the relationship between the team, how it recognizes its members' competencies, and how each individual contributes with its specific potential. This similarity between both models' measured aspects might explain the measures agree. Thus, contradicting the results for the p-value for the Student's T-test we failed to reject H5.

For the *Mutual Support* variable we obtained a p-value < 0.05, indicating that it is not equivalent in both models. Furthermore, looking at Figure Figure 5, we can see that the line that represents the average of the differences is very far from 0. Besides, its confidence interval is below 0. Finally, *Mutual Support* presented a significance ≤ 0.05, indicating proportion bias. The *Mutual support* variable showed a discrepancy between the results of the models. This can be explained by the fact that Mutual Support (TWQ-SEM model) has been mapped for *Self-organizing* (the TWQ-BN model), and this, in turn, is composed of *Expertise*, *Shared Leadership*, and *Team Learning*. Of the seven questions used to assess *Mutual support*, two are related to *Expertise*, and five are related to *Shared Leadership*, but none address *Team Learning*, which might have influenced the discrepancy of the results. Thus, we rejected H6. As a consequence, from the six variables analyzed, we have enough evidence to reject the equivalence of four of them: *Communication*, *Coordination*, *Cohesion*, and *Mutual Support*.

B. Implications for Research and Practice

This section discusses the implications of the results for researchers and practitioners by evidencing similarities and discrepancies of the evaluation measures of two agile TWQ models' results.

The TWQ-BN and TWQ-SEM models presented similar results for only two variables from the research perspective. Thus given that the measures collected might be context-sensitive, our findings cannot be generalized at this point. Indeed, additional studies have to be conducted to explore

how to measure *Communication*, *Coordination*, *Cohesion* and *Mutual Support*.

A first step to explore the measurement of the variables of interest would be running studies and constructing a catalog of valid TWQ measures values, either based on questionaries or automatically collecting them from tools related to each TWQ model variable. One such catalog to be helpful would have to characterize the context in which the measures were successfully adopted following appropriate guidelines as defined in Petersen et al. [29]. Also, each candidate measure must be analyzed based on criteria such as internal, external, and construct validity, among others, described by Meneely et al. [30]. Finally, there is a need for guidelines or processes to help professionals to adopt them, including the use of action research [31].

Given that the TWQ-BN model and TWQ-SEM models had similar results, researchers should further investigate the significance of *Team Autonomy* as part of the agile TWQ construct since it is considered for the TWQ-BN model. However, it is not for the TWQ-SEM model. Additionally, one can assess the relationship between the concepts of TWQ and Team Climate [32].

We presented evidence establishing the repercussions of this study from a practice perspective. BNs support both prognosis and diagnosis while SEM does not. Further, while SEM focuses on demonstrating a theory, BNs assume that the main role of causal modeling is to facilitate the analysis of potential and real actions to introduce a conceptual intervention, evaluate observable changes expected and perform"what-if" analysis. Also, BNs provide detailed non-linear information on the relationship that should be easily consumed by managers and academics [33].

In terms of practical utility, besides the previously reported benefits of using BNs over SEM to support decision-making [34], the TWQ-BN model has the advantage of reducing the effort to measure TWQ over the TWQ-SEM model. The TWQ-BN model questionnaire contains nine questions, and the TWQ-SEM model contains 38 questions to measure the same variables (except for *Team Autonomy*.

The result of this study cannot be generalized to support any statement about which model has greater validity because the measures can be context-sensitive. Considering that the problem of defining measures for TWQ is an instance of defining measures for any software measurement program, it should be addressed as an effort to define valid software measures [30]. Additional studies have to be performed to define a process to help teams adopt these models in their work environment, considering the choice of questions to evaluate the variables according to the context the team is inserted.

V. Threats to Validity

In this section, we present the threats to validity following the classification presented by Wohlin et al. [35], and the scheme we applied to minimize them.

Internal validity. During the data collection sessions, the subjects answered questions about the two models, which took approximately 40 minutes. Such time may have influenced the results due to fatigue. All subjects answered simultaneously while being monitored the entire process to minimize this. Another threat is understanding the research objectives and the questions of both models. To minimize it, we took two approaches:

1) we trained the subjects to elucidate the research objectives, guarantee the understanding of the questions for the questionnaires related to the two models. Subjects answered based on their daily experience and not on their intentions;
2) we used the questionnaires formerly applied and validated by Freire et al. [11] and Lindsjørn et al. [3]. The subjects provided data about their teams, which could be biased. We minimized this bias by guaranteeing the confidentiality of the data.

External validity. We collected data from 25 teams 24 from the same organization. Therefore, the organization's culture may be influenced the results. We collected data from teams working in different domains and industry partners to minimize this threat. Furthermore, most teams had separate managers, a total of 13 managers for the 25 teams.

Another threat to external validity is data generalization because all teams applied Scrum. Although this can limit the generalization of the scrum teams' results, Scrum is present among the five most-used agile practices in the industry, and 75% of projects use Scrum according to the latest report by the State of Agile [7].

Conclusion validity. The equivalence between the measurement models using only the *p*-value may yield false positives. To minimize this, we carried out the analysis also using the 95 % confidence interval approach. Another threat is related to the sample size of the subjects. The results may have been directly influenced by the sample size and the work experience of the subjects. We interviewed 162 different team members across different roles interviewed, as mentioned before. To mitigate this threat, we chose a balanced sample of different roles and conducted training to level the knowledge of the subjects on these TWQ models.

Construct validity. We performed the mapping between both models. Although the TWQ-BN model has more variables than the TWQ-SEM model, their definitions were mapped by similarity independently by the first, second, and third authors to minimize the bias. After that, conflicts were discussed before reaching the final mapping. Furthermore, we adapted the questionnaire proposed by Lindsjørn et al. [3], combining the perception of individuals in the same team to a single perception using arithmetic means, and normalized the data. Even though knowing that taking the arithmetic mean from an ordinal scale is a controversial topic in statistics, we decided to follow the procedure applied by Lindsjørn et al. [3] for compliance purposes.

VI. Final Remarks

This study presented a comparative analysis between the results of two TWQ models in the context of ASD, the TWQ-BN, and TWQ-SEM models. Our results demonstrated a substantial difference between the measures for the *Communication*, *Coordination*,*Cohesion*, and *Mutual Support* variables.

On the other hand, the results obtained for the variables *Balance of member contribution* and *Effort* are equivalent.

Hence, we conclude that the models are not equivalent and that it is up to the teams to choose the TWQ model that best suits their context. While the TWQ-BN model demands less effort to be applied and allows for diagnosis and prognosis, the TWQ-SEM model calculates prognostic inferences based on more granular data.

The results discussed contribute to researchers and professionals. For researchers, our study offers insights to assess the variables that make up the TWQ construct, detailing the limitations and discrepancies between the TWQ-BN and the TWQ-SEM models. On the other hand, practitioners provide a more detailed understanding of how the analyzed TWQ models work, empowering the team to decide which model to use, given its context.

We plan to analyze different methods to gather more objective data generated in the software development lifecycle, thus reducing operational effort. Given that the measures can be context-sensitive, we plan to define a process to assist in implementing the models in the context of each team. Also, we plan to extend our analysis to compare other TWQ measurement models in the context of ASD.

VII. Acknowledgments

This work has been developed by Intelligent Software Engineering Group (ISE) and financed by Virtus Innovation, Research and Development Center (VIRTUS), Federal University of campina Grande, Campina Grande, PB, Brazil.

References

[1] M. F. Van Assen, "Agile-based competence management: the relation between agile manufacturing and time-based competence management," *International Journal of Agile Management Systems*, 2000. [Online]. Available: https://doi.org/10.1108/14654650010337168

[2] L. Gren, R. Torkar, and R. Feldt, "Group maturity and agility, are they connected? – a survey study," in *2015 41st Euromicro Conference on Software Engineering and Advanced Applications*, 2015, pp. 1–8. [Online]. Available: https://doi.org/10.1109/SEAA.2015.31

[3] Y. Lindsjørn, D. I. Sjøberg, T. Dingsøyr, G. R. Bergersen, and T. Dybå, "Teamwork quality and project success in software development: A survey of agile development teams," *Journal of Systems and Software*, vol. 122, pp. 274–286, 2016. [Online]. Available: https://doi.org/10.1016/j.jss.2016.09.028

[4] T. Dingsøyr, T. E. Fægri, T. Dybå, B. Haugset, and Y. Lindsjørn, "Team performance in software development: Research results versus agile principles," *IEEE Software*, vol. 33, no. 4, pp. 106–110, 2016. [Online]. Available: https://doi.org/10.1109/MS.2016.100

[5] L. Lukusa, S. Geeling, S. Lusinga, and U. Rivett, "Teamwork and project success in agile software development methods: A case study in higher education," in *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, ser. TEEM'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 885–891. [Online]. Available: https://doi.org/10.1145/3434780.3436648

[6] S. Stavru, "A critical examination of recent industrial surveys on agile method usage," *Journal of Systems and Software*, vol. 94, pp. 87–97, 2014. [Online]. Available: https://doi.org/10.1016/j.jss.2014.03.041

[7] VersionOne, "14th annual state of agile development survey results," https://bit.ly/3kRlcs3, 2020, accessed: 05-18-2021.

[8] A. Poth, M. Kottke, and A. Riel, "Evaluation of agile team work quality," in *Agile Processes in Software Engineering and Extreme Programming – Workshops*, M. Paasivaara and P. Kruchten, Eds. Cham: Springer International Publishing, 2020, pp. 101–110. [Online]. Available: https://doi.org/10.1007/978-3-030-58858-8_11

[9] S. Cruz, F. Q. da Silva, and L. F. Capretz, "Forty years of research on personality in software engineering: A mapping study," *Computers in Human Behavior*, vol. 46, pp. 94–113, 2015. [Online]. Available: https://doi.org/10.1016/j.chb.2014.12.008

[10] G. Marsicano, F. Q. da Silva, C. B. Seaman, and B. G. Adaid-Castro, "The teamwork process antecedents (tpa) questionnaire: developing and validating a comprehensive measure for assessing antecedents of teamwork process quality," *Empirical Software Engineering*, vol. 25, no. 5, pp. 3928–3976, 2020. [Online]. Available: https://doi.org/10.1007/s10664-020-09860-5

[11] A. Freire, M. Perkusich, R. Saraiva, H. Almeida, and A. Perkusich, "A bayesian networks-based approach to assess and improve the teamwork quality of agile teams," *Information and Software Technology*, vol. 100, pp. 119–132, 2018. [Online]. Available: https://doi.org/10.1016/j.infsof.2018.04.004

[12] M. Hoegl and H. G. Gemuenden, "Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence," *Organization Science*, vol. 12, no. 4, pp. 435–449, 2001. [Online]. Available: https://doi.org/10.1287/orsc.12.4.435.10635

[13] N. B. Moe, T. Dingsøyr, and T. Dybå, "A teamwork model for understanding an agile team: A case study of a scrum project," *Information and Software Technology*, vol. 52, no. 5, pp. 480–491, 2010, tAIC-PART 2008. [Online]. Available: https://doi.org/10.1016/j.infsof.2009.11.004

[14] I. Fatema and K. Sakib, "Factors influencing productivity of agile software development teamwork: A qualitative system dynamics approach," in *2017 24th Asia-Pacific Software Engineering Conference (APSEC)*. Piscatway: IEEE, 2017, pp. 737–742. [Online]. Available: https://doi.org/10.1109/APSEC.2017.95

[15] J. Martin Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986, originally published as Volume 1, Issue 8476. [Online]. Available: https://doi.org/10.1016/S0140-6736(86)90837-8

[16] M. Silva, A. Freire, M. Perkusich, D. Albuquerque, K. C. Gorgônio, H. Almeida, A. Perkusich, and E. Guimarães, "A comparative analysis of agile teamwork quality models," in *2021 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2021, pp. 1–6. [Online]. Available: 10.23919/SoftCOM52868.2021.9559062

[17] I. Ben-Gal, "Bayesian networks," in *Encyclopedia of Statistics in Quality and Reliability*, F. W. F. Fabrizio Ruggeri, Ron S. Kenett, Ed. Hoboken: John Wiley & Sons, 2008, ch. 1, pp. 1–6. [Online]. Available: https://doi.org/10.1002/9780470061572.eqr089

[18] N. E. Fenton, M. Neil, and J. G. Caballero, "Using ranked nodes to model qualitative judgments in bayesian networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 10, pp. 1420–1432, 2007. [Online]. Available: https://doi.org/10.1109/TKDE.2007.1073

[19] L. Radlinski, "A survey of bayesian net models for software development effort prediction," *International Journal of Software Engineering and Computing*, vol. 2, no. 2, pp. 95–109, 2010.

[20] R. D. Anderson and G. Vastag, "Causal modeling alternatives in operations research: Overview and application," *European Journal of Operational Research*, vol. 156, no. 1, pp. 92 – 109, 2004, eURO Excellence in Practice Award 2001. [Online]. Available: https://doi.org/10.1016/S0377-2217(02)00904-9

[21] V.-P. Eloranta, K. Koskimies, and T. Mikkonen, "Exploring scrumbut—an empirical study of scrum anti-patterns," *Information and Software Technology*, vol. 74, pp. 194 – 203, 2016. [Online]. Available: https://doi.org/10.1016/j.infsof.2015.12.003

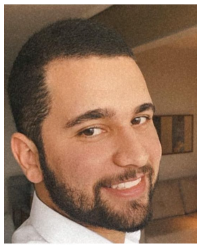
[22] V. Stray, D. I. Sjøberg, and T. Dybå, "The daily stand-up meeting: A grounded theory study," *Journal of Systems and Software*, vol. 114, pp. 101–124, 2016. [Online]. Available: https://doi.org/10.1016/j.jss.2016.01.004

[23] O. Hirsch, H. Keller, C. Albohn-Kühne, T. Krones, and N. Donner-Banzhoff, "Pitfalls in the statistical examination and interpretation of the correspondence between physician and patient satisfaction ratings and their relevance for shared decision making research," *BMC Medical Research Methodology*, vol. 11, no. 1, pp. 1–10, 2011. [Online]. Available: https://doi.org/10.1186/1471-2288-11-71

[24] R. Zaki, A. Bulgiba, R. Ismail, and N. A. Ismail, "Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review," *PloS one*, vol. 7, no. 5, p. e37908, 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0037908

[25] S. D. Vishnubhotla, E. Mendes, and L. Lundberg, "Investigating the relationship between personalities and agile team climate of

software professionals in a telecom company," *Information and Software Technology*, vol. 126, p. 106335, 2020. [Online]. Available: https://doi.org/10.1016/j.infsof.2020.106335
[26] F. Franceschini, M. Galetto, and M. Varetto, "Qualitative ordinal scales: The concept of ordinal range," *Quality Engineering*, vol. 16, no. 4, pp. 515–524, 2004. [Online]. Available: https://doi.org/10.1081/QEN-120038013
[27] D. Giavarina, "Understanding bland altman analysis," *Biochemia medica*, vol. 25, no. 2, pp. 141–151, 2015. [Online]. Available: https://doi.org/10.11613/BM.2015.015
[28] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *International Journal of Nursing Studies*, vol. 47, no. 8, pp. 931–936, 2010. [Online]. Available: https://doi.org/10.1016/j.ijnurstu.2009.10.001
[29] K. Petersen and C. Wohlin, "Context in industrial software engineering research," in *2009 3rd International Symposium on Empirical Software Engineering and Measurement*, 2009, pp. 401–404. [Online]. Available: https://doi.org/10.1109/ESEM.2009.5316010
[30] A. Meneely, B. Smith, and L. Williams, "Validating software metrics: A spectrum of philosophies," *ACM Trans. Softw. Eng. Methodol.*, vol. 21, no. 4, feb 2013. [Online]. Available: https://doi.org/10.1145/2377656.2377661
[31] V. Antinyan, M. Staron, A. Sandberg, and J. Hansson, "Validating software measures using action research a method and industrial experiences," in *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE '16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: https://doi.org/10.1145/2915970.2916001
[32] A. Açıkgöz, A. Günsel, N. Bayyurt, and C. Kuzey, "Team climate, team cognition, team intuition, and software quality: The moderating role of project complexity," *Group Decision and Negotiation*, vol. 23, no. 5, pp. 1145–1176, 2014. [Online]. Available: https://doi.org/10.1007/s10726-013-9367-1
[33] K. Verbert, R. Babuška, and B. De Schutter, "Bayesian and dempster–shafer reasoning for knowledge-based fault diagnosis–a comparative study," *Engineering Applications of Artificial Intelligence*, vol. 60, pp. 136–150, 2017. [Online]. Available: https://doi.org/10.1016/j.engappai.2017.01.011
[34] R. D. Anderson and G. Vastag, "Causal modeling alternatives in operations research: Overview and application," *European Journal of Operational Research*, vol. 156, no. 1, pp. 92–109, 2004.
[35] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, and B. Regnell, *Experimentation in Software Engineering*. Heidelberg: Springer, 2012. [Online]. Available: https://doi.org/10.1007/978-3-642-29044-2



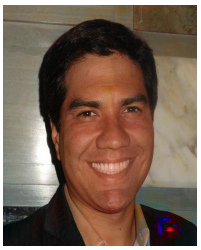
Manuel Silva received the MSc degree in computer science from Federal University of Campina Grande, Brazil, in 2021. His research focuses agile software development. He is a member of the Intelligent Software Engineering Group (ISE/Virtus).

Mirko Perkusich received his PhD degree in Computer Science in 2018. He is a Research Manager at VIRTUS, leading the Intelligent Software Engineering (ISE/VIRTUS) research group. His current research interests are in applying intelligent techniques, including recommendation systems, to solve complex software engineering problems.

Arthur Freire Arthur Freire holds a Ph.D. in Computer Science and is a Software Engineer at Amazon. His current research interests are in applying intelligent techniques for managing agile software development..

Danyllo Albuquerque received the MSc in informatics from the Federal University of Paraiba, Brazil, in 2013. Currently, he is a Ph.D. student in computer science at the Federal University of Campina Grande, Brazil. He is also a member of the Intelligent Software Engineering (ISE/Virtus) research group. His research interests are in applying intelligent techniques to solve software engineering problems.

Kyller Costa Gorgônio graduated in computer science from the Universidade Federal da Paraíba, in 1999. He received a master's degree in computer science from the Universidade Federal da Paraíba, in 2001, and the Ph.D. degree in software from the Universitat Politècnica de Catalunya, in 2010. He is currently a Professor with the Universidade Federal de Campina Grande. He has experience in computer science focusing on software engineering. His research interests include Petri nets, protocols, asynchronous communication mechanisms, coloured Petri nets, and model checking.



Hyggo Almeida is a professor at the Computer and Systems Department, Federal University of Campina Grande (UFCG) since 2006. Dr. Almeida has got his Ph.D. in Electrical Engineering and M.Sc. in Computer Science from the Federal University of Campina Grande in 2007 and 2004. He is currently the head of the Intelligent Software Engineering group and Founder and Director of Operations at Virtus Innovation, Research and Development Center(VIRTUS/UFCG). He is a researcher at Embedded Systems and Pervasive Computing Laboratory (Embedded/UFCG). He is also Executive Director of EMBRAPII Unit at CEEI-UFCG, with more than 150 RD&I projects developed in cooperation with industrial companies within Information, Communication, and Automation Technologies. His current research interest is applying intelligent techniques to solve complex software engineering problems.

Angelo Perkusich (Member IEEE) is a professor at the Electrical Engineering Department (DEE), Federal University of Campina Grande (UFCG) since 2002. Dr. Perkusich got his Ph.D. and Master's degrees in Electrical Engineering from the Federal University of Paraiba in 1987 and 1994, respectively. He was a visiting researcher at the Department of Computer Science, University of Pittsburgh, PA, USA, from 1992 to 1993. He is currently the principal investigator of research projects financed by public institutions such as FINEP (Brazilian Agency for Research and Studies), CNPq (Brazilian National Research Council), and private companies. He is the founder and Director of Virtus Innovation, Research and Development Center and Embedded and Pervasive Computing Laboratory. Research projects focus on formal methods, embedded systems, mobile pervasive and ubiquitous computing, and software engineering. He has over 30 years of teaching experience in the university and training courses for industry in the context of software for real-time systems, software engineering, embedded systems, computer networks, and formal methods. His main research areas are Embedded Systems, Software Engineering, Mobile Pervasive Computing, and Formal Methods.

Everton Guimarães is an Assistant Professor at Penn State University. He has collaborated with many partners in industry and academia, both in Brazil and abroad, over the past nine years. His most recent research investigates software architecture and source code problems and their impact on the overall software quality attributes (i.e., maintenance, evolution). He is also interested in research topics related to technical debt, architecture recovery techniques, pattern detection, and mobile computing.