Assessing the Inter-rater Reliability of the Usability Model for Software Development Processes and Practices

Diego Fontdevila (dfontdevila@untref.edu.ar)  
Universidad Nacional de Tres de Febrero

Marcela Genero  
University of Castilla-La Mancha

Nicolás Paez  
Universidad Nacional de Tres de Febrero

Alejandro Oliveros  
Universidad Nacional de Tres de Febrero

Research Article

Keywords: process and practice, usability, inter-rater reliability, scrum, TDD, BDD

Posted Date: July 7th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3104393/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Software processes and practices have a leading role in software development and in the last few decades a wide variety of processes and practices have emerged to face the challenges arising in the software industry. The success of process and practice adoption initiatives depends at least partially on the experience and satisfaction of the people who use them, thus making usability an interesting process quality attribute.

This paper describes the inter-rater reliability evaluations performed on the UMP to assess consistency among metric values produced by different evaluators.

It presents two inter-rater reliability assessment studies, the Scrum study and the TDD-BDD study.

The paper presents four inter-rater reliability statistics for the process and practices under study, a comparative analysis of their strengths and weaknesses, and an analysis of the study results and their interpretation.

The results show varying reliability results among the different UMP metrics, which seem to depend mostly on object of evaluation (Scrum, TDD, BDD), metric subjectivity and context sensitivity. The UMP was refined from the data gathered in the studies and a metric categorization is proposed to facilitate metric usage.

1 Introduction

Processes are central elements of software development, and they have changed in the past decades, from views inspired in manufacturing to more innovative approaches, like Agile and DevOps. New approaches tend to focus more on people and practice (Jacobson et al., 2007), and have usually been created by practitioners, not methodologists. Processes and practices are tools (Cockburn, 2004; Pfleeger, 1999), and there is evidence that the interactions between users and their methods are alike to their interactions with their tools (Riemenschneider et al., 2002). Given that usability characterizes artifacts that are easier and more attractive to use, usability might improve process and practice adoption, and also make those adopted processes and practices more sustainable. Therefore, applying usability principles and heuristics to software development processes and practices might help adoption initiatives and improve the experience of the people involved. For example, feedback is a usability principle applied in iterative processes, allowing teams to collect information about their product, processes and practices, in order to learn and improve.

There is little research on process and practice usability. Although several researchers have proposed process quality models (Feiler & Humphrey, 1992; Guceglioglu & Demirors, 2005; Kroeger et al., 2014), very few include usability among them.
To assess the impact opportunities addressed by this research, it is enough to regard how real-world teams and organizations strive to adopt Agile and DevOps practices and embrace their mindsets. Many agile transformation initiatives struggle to accomplish their objectives (Dikert et al., 2016) and practice adoption levels are not what might be expected given the popularity of agile methods (Kuhrmann et al., 2019; Paez et al., 2018). This produces a negative impact on process improvement initiatives and negatively affects costs and motivation. Also, many improvement initiatives are planned and conducted in top-down fashion without considering the perspective of the people affected (J. S. Brown & Duguid, 2000). Process and practice improvement through adoption is hard, even for effective organizations. Lack of clear and concrete guidance makes these challenges even more difficult.

In order to support the application of usability concepts to process and practice, the Usability Model for Software Development Process and Practice (UMP) was developed following the Design Science Research approach (Johannesson & Perjons, 2014). UMP is formed by characteristics and metrics. The model can be applied to specific processes and practices to evaluate their usability and identify related improvement opportunities. UMP’s goal is to help improve the usability of software development processes and practices, to improve the work experience of software developers and the overall effectiveness of process and practice improvement and adoption initiatives.

One of the main requirements established for UMP is that it should be reliable, meaning that evaluations by different people should produce consistent results when applied to the same process or practice. For example, if several people evaluate a given process or practice, the values produced for each metric should not present high variance because that might imply that the UMP is not able to reliably describe the usability features of the process or practice. This type of reliability is usually defined as inter-rater reliability (Hallgren, 2012). Reliability is significant in this context because the model, to be effective, must be used by different people, particularly given that software development, is in general a collective endeavor (this also applies to improvement initiatives in organizations). If different people produce very different metric values for the same process or practice, there is high risk the values cannot be effectively unified and thus impair effective interpretation of the metric values and prevent further improvement actions. Particularly, issues of low reliability might point to “instability of the measuring instrument when measurements are made between coders” (Hallgren, 2012), and thus, provide opportunity for improving the UMP.

The Design Science Research approach guides the development and validation of artifacts aimed at solving specific practical problems. The purpose of this paper is to present two validation studies conducted to assess inter-rater reliability for the UMP’s 23 metrics. We also share the challenges that emerged in applying statistics traditionally recommended in software engineering literature like those from the Kappa family, the learnings they provided and describe the alternative statistics eventually applied to resolve those challenges.

The first study was conducted by applying the UMP to Scrum, which was selected for the initial study for several reasons. First, to pursue the research line initiated with the UMP feasibility study in which two
external experts evaluated Scrum (Fontdevila et al., 2017). Secondly, being a process framework, Scrum was considered an object of evaluation larger than most practices and at the same time, simple enough for holistic evaluation. It is also the most popular agile method and some of its stated values resonate positively with usability. For example, transparency, one of Scrum’s values (Schwaber & Sutherland, 2017) matches visibility, one of the UMP characteristics.

The second study was conducted on TDD and BDD, and they were selected to complement Scrum (a process oriented example) with two practices, which tend to be more fine-grained. Also, some of the UMP’s characteristics like Feedback seemed to describe part of the appeal of these practices, while some specific metrics like Conceptual model correspondence seemed to highlight some of the challenges perceived with TDD adoption, particularly issues with the test-first approach (Beck & Andres, 2004).

The rest of this paper is structured as follows: Section 2 presents the background related work; Section 3 describes the overall research strategy applied to develop the UMP following the DSR framework and the research method applied to assess UMP’s metrics inter-rater reliability; Section 4 describes the UMP’s characteristics and metrics in detail; Section 5 and Section 6 present the Scrum study and the TDD-BDD study, respectively. Finally Section 7 outlines the conclusions and the lines that are pending for future research.

2 Background and Related work

This section presents the background and related work for this research, from process and practice usability to inter-rater reliability studies, including the sources used to create the UMP.

2.1 Process Usability Models

This section presents existing research on process and practice usability and the sources used in the construction of UMP.

As stated before, there is little research on process and practice usability, particularly process quality models. Kroeger et al. defined a process quality model from the users’ perspective, including usability as one of four quality attributes (Kroeger et al., 2014). Their model consists of a set of four process quality attributes: suitability, usability, manageability and evolvability, obtained through a grounded-theory approach from case study reports and interviews. The study also identifies four attribute sub-types for usability: accessibility, understandability, learnability and adaptability. This is the only model explicitly created from the process user’s perspective, but it is not completely focused on usability and it does not include dynamic aspects of process and practice (for example, controllability or error tolerance). Culver-Lozo and Mahrin have studied the usability of process descriptions, but not of process enactment, thus reducing users’ interactions with the process to mere information access experiences, and none of the actual experience of participating in the process (Culver-Lozo, 1995; Mahrin et al., 2008). Again, in the case of the works of Culver-Lozo and Kroeger et al., the interpretation of the accessibility sub-characteristic is focused on accessibility of information about the process, not on the actual ability of
process users of conducting/participating in the process effectively. Feiler and Humphrey mention process usability in the introduction to their paper, specifically reflecting on the need to mind process usability early on given that it is not easy to improve later since gathering feedback from process enactment takes significant time, but they do not include usability as a process quality attribute in their conceptual model (Feiler & Humphrey, 1992).

An SMS was conducted to identify pertinent literature about process and practice usability, which produced a short list of process and practice usability research. Overall, it confirmed that very little research has been conducted on software development process and practice usability. Only 16 papers were finally selected to be analyzed (from an initial search result set of 1326). From those 16 papers, 5 describe usability aspects of specific processes or practices, most of them proposed by the authors and with no apparent user base (except for formal methods). These were not considered viable sources for UMP. The rest of the research produced was focused on generic processes and practices, some of them describing the application of usability techniques to process improvement (Polgár, 2015; Polgár & Biró, 2011) and the rest describing usability as a process quality attribute. This last category (the one that matched the objectives of the UMP) included: the aforementioned process quality model including usability by Kroeger et al. (Kroeger et al., 2014; Kroeger & Davidson, 2009); studies focused on the usability of process descriptions for their users (e.g. manuals) (Culver-Lozo, 1995; Mahrin et al., 2008); a study on measuring process understandability focused on process descriptions (Melcher & Seese, 2008); two studies on usability of collaborative processes by different authors (Cardoso, 2005; den Hengst et al., 2006), in the second one the focus is a hybrid of process/tool usability; and a definition of software process quality based on statistical process control (Li et al., 2005).

Beyond the results of the SMS on process and practice usability (Fontdevila, 2021), it is noteworthy that there is no standard to evaluate process quality (as there for product quality, for example (International Organization for Standardization, 2011)), and there are only some proposals for process quality models such as those described in (Feiler & Humphrey, 1992; Guceglioglu & Demirors, 2005; Kroeger et al., 2014). There is also no consensus on what characteristics of the software processes should be evaluated or what measures to use to evaluate these characteristics (Kroeger et al., 2014). Finally, there is no process quality model focused on the evaluation and improvement of usability aspects of software development processes and practices.

UMP was initially constructed based on three sources: Kroeger et al.’s process quality model (Kroeger et al., 2014), the International Standard on Software and Systems Quality ISO 25000 (International Organization for Standardization, 2011a) and product usability classic literature (Nielsen, 1994; Norman, 1988). The work of Kroeger et al. provided a research source approaching process quality from the user’s perspective; the ISO 25000 included an industry perspective since there is no international standard on process quality; and product usability classic literature provided the initial inspiration and insights for this research, in particular the resonance between the concepts of feedback and error tolerance as described in usability and software engineering literature.

2.2 Reliability Studies
The main focus of this paper is presenting the studies conducted to assess UMP’s reliability, which we have described as its ability to produce consistent results when applied by different subjects to the same process or practices. For example, if multiple experts evaluate a specific process, the metric values should not display too high variance, otherwise, the model’s ability to describe the usability features of the process would be compromised. This field emerged in the context of psychometric and medical studies and has grown significantly and been applied to other contexts (Gwet, 2014; Hallgren, 2012). It applies whenever a set of subjects (raters) measure some characteristic of an object (which can also be a person) and there is a need to assess the quality of those measurements.

This type of reliability can be assessed with two approaches, each with their own set of statistics (Hallgren, 2012): inter-rater reliability (Gwet, 2014) and inter-rater agreement (James, 1982). Each of the approaches has advantages and disadvantages, and they differ both in their theoretical foundations and on their applicability requisites. Both approaches were used to evaluate UMP, inter-rater agreement in the Scrum study and inter-rater reliability in the TDD-BDD study.

### 2.2.1 Inter-rater reliability and Inter-rater agreement

Both inter-rater reliability and inter-rater agreement are measures of association (Kitchenham et al., 1995). The main theoretical difference between them is in how they conceptualize variance (Liao et al., 2010).

Inter-rater reliability states, following classical psychometric test theory, that observed values (X) are the sum of a true score (T) that would be the true value for the characteristic of that object if there were no measurement error, and a measurement error (E) (Hallgren, 2012). Thus:

\[
X = T + E
\]

And, assuming that true scores and errors are uncorrelated:

\[
\text{Var}(X) = \text{Var}(T) + \text{Var}(E)
\]

Inter-rater agreement, on the other hand, states that total variance is the sum of random measurement error (E) variance plus systematic variance, which is comprised of true (T) variance and variance that reflects bias (B) among raters (Liao et al., 2010). Thus:

\[
\text{Var}(X) = \text{Var}(T) + \text{Var}(B) + \text{Var}(E)
\]

The inter-rater agreement statistic is calculated following James et al. (James, 1982) as described in (Liao et al., 2010):

\[
r_{WG} = 1 - \frac{\text{Var}_{obs}}{\text{Var}_{rand}}
\]

In the formula, r is the inter-rater agreement coefficient, \(\text{Var}_{obs}\) is the observed variance in the sample, and \(\text{Var}_{rand}\) is the variance if all ratings emerged from random measurement error only. Variance in inter-rater
agreement is thus defined in terms of the relationship between the sample variance and random measurement error variance (Liao et al., 2010). In other words, when the sample variance is close to random error $r$ is close to 0 (low agreement), and when the sample variance is very small, $r$ is close to 1 (high agreement).

In inter-rater reliability, statistics correct in varying ways for chance agreement (that is, when evaluators are uncertain and they assign a metric value at random among those they consider potentially appropriate), and they are recommended for software engineering studies (Kitchenham et al., 1995).

Measurement errors can be due to “imprecision, inaccuracy, or poor scaling of the items within an instrument (i.e., issues of internal consistency); instability of the measuring instrument in measuring the same subject over time (i.e., issues of test-retest reliability); and instability of the measuring instrument when measurements are made between coders” (Hallgren, 2012). This last type of error is the one that affects inter-rater reliability and is the focus of this article. Issues with imprecision, inaccuracy and poor scaling of the items were evaluated and refined initially through the focus group as described in (Fontdevila, 2021); issues with test-retest reliability were not considered an issue in this context, given how challenging and complex the evaluating processes and practices is.

In terms of study design, inter-rater agreement supports study designs with multiple raters evaluating a single object, while inter-rater reliability studies require at least two objects of evaluation and multiple raters, although not all raters need to rate all objects (Hallgren, 2012).

[1] This was considered appropriate because the process as software product is an analogy that other researchers have already used (Feiler & Humphrey, 1992; Osterweil, 1987).

### 3 Research Strategy

UMP was created and validated following a research strategy based on the Design Science Research framework (Johannesson & Perjons, 2014). The Design Science Research framework is an innovative approach to the creation and validation of artifacts that provide solutions or support for a specific practice in the real world. DSR has two main objectives, the creation of the artifact (or artifacts) and the generation of scientific knowledge about how those artifacts affect the world around them. This knowledge provides insights to determine how and when the artifacts can be used in different contexts to provide the value they were designed for, ensuring that their applicability goes beyond the context in which it was created (Johannesson & Perjons, 2014).

Figure 1 presents the research strategy, for each Design Science Research activity it shows the tasks performed during the research process.

The following sections describe the DSR activities and the tasks performed, focusing on the reliability evaluations and providing references for more detailed descriptions of the other activities.
3.1 Explicate Problem

This activity aims at defining a practical problem and its significance in context for people sharing a specific practice.

The problem was defined in terms of the needs of organizations to improve, adopting existing processes and practices, or evolving their own, and how the limited existing research on software process and practice usability provided an opportunity.

The state of the art on software process and practice usability was established by performing a Systematic Mapping Study (SMS) following the guidelines proposed in (Kitchenham & Charters, 2007; Petersen et al., 2015).

3.2 Define Objective and Requirements

This activity aims at defining the objective and requirements for the artifact to be produced.

The objective is aligned with addressing the problem, in this case, improving the usability of processes and practices in order to improve the development experience for the people involved and the effectiveness of improvement initiatives.

The structure for UMP was defined to be as simple as possible, namely a set of usability characteristics and a set of metrics for each characteristic. The aim was to facilitate UMP usage.

The requirements describe the expected features or quality attributes of the artifact. In this case, utility was considered one of the required quality attributes (this applies to all DSR artifacts, since they are about improving a specific practice) and reliability the second. Utility is about UMP being useful to developers, consultants, coaches, improvement specialists, researchers and teachers for identifying usability related improvement opportunities in software development processes and practices. Reliability is about different people producing consistent results when evaluating the same process or practice, so that they are able to use those evaluation results effectively. UMP Reliability is the focus of this article. To operationalize these requirements, a set of UMP usage scenarios were defined and eventually guided the selection of evaluation contexts (Fontdevila, 2021).

3.3 Design and Develop Artifact

This activity aims at creating the artifact to fulfill its requirements and achieve its objective.

UMP was initially constructed based on three sources as described in Section 2.1, and an evaluation process was defined for systematically applying UMP to specific processes and practices, based on (International Organization for Standardization, 2011b).

After its initial version was defined, an iterative refinement process based on empirical evaluation studies and UMP refinements was conducted. The refinements included adding, removing and modifying
characteristics and metrics, and also the definition of three usage modes for UMP: evaluation, profile and framework (see details in (Fontdevila, 2021)).

3.4 Demonstrate Artifact

This activity aims at showing that the proposed artifact actually works in a specific context.

This was achieved by conducting a feasibility study by applying UMP to Scrum (Fontdevila et al., 2017)

3.5 Evaluate Artifact

This activity aims at evaluating the extent to which the artifact fulfills its stated requirements and addresses the practical problem that motivated its creation.

UMP was evaluated in two dimensions determined by the requirements originally defined for it: utility and reliability. Utility was initially evaluated through a case study on applying UMP to the Visual Milestone Planning method (Miranda, 2019), which is described in detail in (Fontdevila et al., 2019). Furthermore, a quasi-experiment was conducted on the application of UMP to the evaluation of a team’s BDD implementation (Fontdevila, 2021).

This paper focuses on the studies conducted to evaluate UMP’s reliability, as described in Section 5.

4 UMP description

UMP was created through a series of refinements and validations as described in Section 2. The usability characteristics were taken and adapted from the selected sources to fit processes and practices.

UMP was created from several sources through an iterative process. During the initial definition, three sources were established as described in Section 2.1:

- Kroeger et al.’s process quality model (Kroeger et al., 2014)
- Product usability classic literature (Nielsen, 1994; Norman, 1988).

UMP is comprised of 10 characteristics and 23 metrics, as shown in Fig. 2, which allow for evaluation of process and practice usability and identifying improvement opportunities and adoption related risks.

Table 1 and Table 2 show an overview of UMP characteristics and metrics, respectively.
### Table 1
UMP Characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-evident purpose</td>
<td>Ease with which users can recognize what a process or practice is for by its name.</td>
</tr>
<tr>
<td>Learnability</td>
<td>Ease with which process or practice users are able to learn how to perform its activities at a novice level of ability (Dreyfus &amp; Dreyfus, 1980).</td>
</tr>
<tr>
<td>Understandability</td>
<td>Ease with which process or practice users are able to apprehend how the underlying principles, structure and dynamics make it work to achieve the desired results.</td>
</tr>
<tr>
<td>Safety</td>
<td>Degree to which a process or practice is safe for its users, preventing errors or limiting their impact, including using the practice or process incorrectly.</td>
</tr>
<tr>
<td>Feedback</td>
<td>Degree to which the use of a process or practice produces or promotes reactions or responses to actions performed.</td>
</tr>
<tr>
<td>Visibility</td>
<td>Degree to which a process or practice helps make activities, status, obstacles and information inputs and outputs visible to people.</td>
</tr>
<tr>
<td>Controllability</td>
<td>Degree to which a process or practice allows its users to check status and make decisions that affect the outcomes during process or practice execution.</td>
</tr>
<tr>
<td>Adaptability</td>
<td>Ease with which process or practice users are able to adapt the process or practice for use in different contexts.</td>
</tr>
<tr>
<td>Attractiveness</td>
<td>Degree to which users find a process or practice attractive or appealing by its form, structure or reported results.</td>
</tr>
<tr>
<td>User satisfaction</td>
<td>Degree to which user needs are satisfied when using a process or practice.</td>
</tr>
<tr>
<td>Characteristic</td>
<td>Metric</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>Self-evident purpose</td>
<td>Appropriateness of name</td>
</tr>
<tr>
<td></td>
<td>Recognized purpose</td>
</tr>
<tr>
<td>Learnability</td>
<td>Time required to learn to perform</td>
</tr>
<tr>
<td></td>
<td>Standard introductory course duration</td>
</tr>
<tr>
<td></td>
<td>Number of new concepts</td>
</tr>
<tr>
<td>Understandability</td>
<td>Conceptual model correspondence</td>
</tr>
<tr>
<td></td>
<td>Conceptual model complexity</td>
</tr>
<tr>
<td>Safety</td>
<td>Cost of incorrect adoption</td>
</tr>
<tr>
<td></td>
<td>Reduction in cost of error</td>
</tr>
<tr>
<td>Characteristic</td>
<td>Metric</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Safety perception</td>
<td></td>
</tr>
<tr>
<td>Use of restraining</td>
<td></td>
</tr>
<tr>
<td>functions</td>
<td></td>
</tr>
<tr>
<td>Feedback</td>
<td>Timeliness of feedback</td>
</tr>
<tr>
<td>Feedback richness</td>
<td></td>
</tr>
<tr>
<td>People feedback</td>
<td></td>
</tr>
<tr>
<td>Automatic feedback</td>
<td></td>
</tr>
<tr>
<td>Visibility</td>
<td>Defines indicators</td>
</tr>
<tr>
<td>Controllability</td>
<td>Defines checkpoints</td>
</tr>
<tr>
<td>Explicit outcomes</td>
<td></td>
</tr>
<tr>
<td>Level of autonomy</td>
<td></td>
</tr>
<tr>
<td>Adaptability</td>
<td>Defines adaptation</td>
</tr>
<tr>
<td>points</td>
<td></td>
</tr>
<tr>
<td>Characteristic</td>
<td>Metric</td>
</tr>
<tr>
<td>---------------</td>
<td>--------</td>
</tr>
<tr>
<td>Ratio of roles allowed to adapt</td>
<td>Measures how many roles among the process or practice users are allowed to modify the process or practice out of the total number of roles (evaluators must specify the roles considered, if no roles are distinguishable, value should be non-applicable).</td>
</tr>
<tr>
<td>Attractiveness</td>
<td>User attractiveness rating</td>
</tr>
<tr>
<td>User satisfaction</td>
<td>User satisfaction rating</td>
</tr>
</tbody>
</table>

5 Scrum study

The Scrum study was conducted to assess UMP’s reliability by having experts apply it to Scrum. It was designed as a baseline study, to provide information on UMP’s reliability and eventually complemented by the TDD-BDD study.

5.1 Study Design and Statistic Selection

The study was designed as an inter-rater agreement study, requiring only one object of evaluation (i.e. Scrum). This simplified data collection and statistic implementation selection (Hallgren, 2012).

5.1.1 Object Selection

Scrum was selected as the object of evaluation for the following reasons:

- Granularity: Being a process framework, Scrum is larger in size than most practices, but at the same time manageable enough for holistic evaluation.
- Process and practice: Scrum is a process related object of evaluation, thus complementing practices like TDD and BDD.
- Previous research: Scrum had been used as object of evaluation during UMP’s feasibility study (Fontdevila et al., 2017), and it was considered valuable to collect further data on Scrum’s usability.
- Popularity and rate of usage: Scrum is the most popular agile framework (Version One, 2020).
- Affinity with usability: Some of Scrum’s pillars and values are well aligned with usability. For example, transparency is akin to visibility, one of UMP’s characteristics (Schwaber & Sutherland, 2017).
5.1.2 Subjects

The study subjects were selected by convenience, with initial study subjects being direct contacts of the authors. These subjects were invited and asked to recommend other candidates, using snowball sampling to increase the sample size (Mockus, 2008).

The total number of subjects in the study was 13, and the following criteria were applied for acceptance:

- At least 5 years of experience with Scrum.
- Acceptable roles were practitioner, mentor, coach, teacher, consultant, manager/supervisor and researcher/academic.

Average experience with Scrum was 10.38 years, which is considered high practical experience. Table 3 shows the distribution of roles among the study participants (each subject could select more than one role).

<table>
<thead>
<tr>
<th>Role</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practitioner</td>
<td>10</td>
</tr>
<tr>
<td>Coach</td>
<td>8</td>
</tr>
<tr>
<td>Consultant</td>
<td>8</td>
</tr>
<tr>
<td>Mentor</td>
<td>7</td>
</tr>
<tr>
<td>Teacher</td>
<td>6</td>
</tr>
<tr>
<td>Manager/supervisor</td>
<td>5</td>
</tr>
<tr>
<td>Researcher/academic</td>
<td>5</td>
</tr>
</tbody>
</table>

As shown in Table 1, most of the participants were expert practitioners, with experience using Scrum, and coaches/mentors, experienced in helping others use Scrum.

5.1.3 Statistic and Variable Selection

The statistic selected for assessing inter-rater agreement is $r_{WG}$, calculated following James et al. (James, 1982) and already described in Section 1.6. Thus, 24 variables were defined as $r_{WG_i}$, with $i$ corresponding to each of the UMP metrics.

5.1.4 Planning

The study was designed so that experts could participate through a self-administered questionnaire. This provided access to a wider range of experts, including those in remote locations. Supporting this study
design required improving, with focus on simplification, the existing materials describing the UMP, particularly the evaluation questionnaire. First, it was converted from a spreadsheet into an online form. Figure 3 shows an example question from the Scrum questionnaire (The full questionnaire is available at http://bit.ly/usabilityofscrumform3_1).

A dry-run was performed by having a fellow researcher perform an evaluation of Scrum with the new materials, gathering feedback and implementing several improvements. One of these improvements included the creation of a short (7 min) video explanation of UMP (http://bit.ly/processandpracticeusabilityvideo).

### 5.2 Study Execution

The study was started by issuing invitations to all participants, followed by reminders after a few weeks for those who had not completed their evaluations. Two participants out of 13 requested minor clarifications and all participants completed their evaluations.

### 5.3 Data Analysis

Data from the UMP evaluation questionnaire was exported to a spreadsheet, reviewed and formatted, including converting all answers to numeric scales suitable for processing.

Before even starting the analysis, the data available presented some interesting and unexpected characteristics. Three of the metrics, *Time required to learn to perform*, *Standard introductory course duration* and *Number of specific conceptual definitions*, which allowed any positive number, were found to have some answers very widely spread (e.g. *Time required to learn to perform* values ranged from 9 to 320 hours). Before performing the inter-rater agreement calculations, the values for *Time required to learn to perform* were divided in discrete sub-ranges to create an ordinal numerical scale (this adjusted scale is marked with an asterisk in Table 2). This was consistent with the advice provided in (Hallgren, 2012) to evaluate scales after they have been processed and not as raw data.

Then the inter-rater agreement coefficients were calculated for each variable as described in Section 2.2.1. Table 4 shows the inter-rater agreement coefficient for each variable and the main causes identified for values with moderate or lower agreement.

### 5.4 Results and Conclusions

Table 2 shows that only 11 out of 24 metrics have good or very good inter-rater agreement coefficients, 8 have moderate inter-rater agreement, and the rest fair or poor inter-rater agreement.

Following the guidelines proposed in (Hallgren, 2012), potential causes were identified for each of the metrics with moderate, fair or poor inter-rater agreement. The analysis was based on the metrics definitions and the qualitative comments for each metric provided by the experts during their evaluation.

The following potential causes were identified:
- Scale: ordinal scales with many values produced very poor agreement. Examples range from unbounded integer scales to single metric labeled scales, including the following metrics:
  
  - Appropriateness of name: the scale was formed by the values Deceiving, Ambiguous, Inappropriate, Partial, Appropriate and Precise. In this case, a mix of long scale and associated metric specific labels make this metric very unreliable. Although the meaning of each value was explicitly defined, the amount of information made it very difficult for evaluators to discern the difference between values and thus affected inter-rater agreement negatively.
  
  - Time required to learn to perform, Standard introductory course duration and Number of specific conceptual definitions: the positive integers range allowed evaluators to choose from a very wide spectrum, again producing poor inter-rater agreement.

- Subtle metric semantics: some metrics include evaluator assessment of subtle issues, for example whether newcomers to a process or practice usually perceive it in some specific way. Examples include:
  
  - Recognized purpose: different evaluators focused on different aspects of Scrum, some on issues with the name, other on whether people really understand the changes that are supposed to come about with Scrum, or whether they just approach it because of its attractiveness.
  
  - Conceptual model correspondence: experts evaluated Scrum at different levels of granularity. Some considered specific elements of Scrum in their comments (e.g. backlog refinement and self-organization are deemed hard) while others provide more holistic considerations. There was also a reference to the deceivingly simple nature of Scrum (Schwaber & Sutherland, 2017).
  
  - Defines adaptation points: the concept seems to have been interpreted with a wide range of meanings. For example, the core concept of adapting the process appears associated with the retrospective, as expected, and other examples mention operational parameters like sprint (iteration) length. Some evaluators even considered Scrum has no adaptation points.

- Context sensitivity: some experts commented that certain metrics, particularly those from the Understandability characteristic, seem highly dependent on context. For example, in traditional hierarchical organizations concepts like self-organization might seem very surprising while others like planning meetings might seem very normal.

In general, the results highlighted many improvement opportunities. Some, mostly those around scaling issues, were relatively easy to improve, while more conceptual issues like subtle semantics and context sensitivity were addressed but proved to be very challenging.

The improvements applied produced a new version of UMP (3.2), on which the next study was conducted.

### 5.5 Threats to Validity
The threats to validity for the Scrum study are presented in this section following the categorization by (Wohlin et al., 2012).

Threats to construct validity: Construct validity for this study might have been affected by questionnaire design. Beyond careful initial definition of characteristics and metrics, two researchers reviewed and refined the questionnaire. Although the feedback from experts during the study did not provide specific criticism on the questionnaire, some of the clarifications required were considered sources for improvements, which were applied producing a newer version of UMP (3.2), which was used in the TDD-BDD study.

Threats to internal validity: The study results are considered valid, given the simple nature of statistics applied and the fact that the recommendations provided in (Hallgren, 2012) were followed. The main reservation is the lack of adjustment for chance agreement in the statistics, but this was addressed in the TDD-BDD study described in the next section and is considered a reasonable trade-off with the lower costs in this study since it allowed having a single object of evaluation (Scrum).

Threats to external validity: The main restriction in this study regarding generalizability is the small size of the sample but given that participants were required to be experts and that the study required a 1hr dedication, this is considered a reasonable size. The distribution of participant profiles, described in Table 1, is considered reasonable with a predominance of experts in close contact with practitioners (e.g. coaches) or practitioners themselves. Also, generalizability from a single preliminary study is very limited, that is why a second study was conducted on TDD-BDD to increase generalizability and complement this study.

Threats to conclusion validity: The number of observations restricts the conclusion validity of this study, again this was addressed by conducting the complementary TDD-BDD study.

6 TDD-BDD study

The TDD-BDD study was conducted to assess UMP’s reliability by having experts apply it to TDD and BDD. It was designed as a complementary study, to provide information on UMP’s reliability and improve on some of the limitations of the Scrum study, for example, lack of correction for chance agreement.

6.1 Study Design and Statistic Selection

The study was designed as an inter-rater reliability study, according to the guidance provided in (Hallgren, 2012), and it differed from the Scrum study in that it required at least two objects of evaluation (i.e. TDD and BDD). This allowed the application of statistics which supported correction for agreement due to chance (Hallgren, 2012). Initially, statistics from the Kappa family were initially selected following the recommendations in (Kitchenham et al., 1995).
Although the statistics from the Kappa family have been widely used, early tests with Fleiss’ Kappa produced surprisingly poor inter-rater reliability scores which seemed inconsistent with low variance test samples. This motivated further research leading to the identification of severe limitations with Kappa statistics which had been proved by Byrt et al. (Byrt et al., 1993), and the confirmation that these limitations also affected the study data. Thus, two statistics were selected that were not subject to these limitations, Bennet’s S and Gwet’s Gamma (Girard, 2016; Gwet, 2014).

### 6.1.1 Object Selection

The selection of objects for the second study was aimed at expanding the variety of available research data with the following the criteria and related rationale:

- **Granularity**: considering Scrum a medium size object of evaluation (it is a process framework with several practices), finer grained objects of evaluation were desired.
- **Process and practice**: given that Scrum is a process framework, it was decided to focus on practices.
- **Previous research**: there was expert data available on UMP evaluation of BDD.
- **Popularity and rate of usage**: with Scrum being the most popular agile method (Version One, 2020), it was decided to use objects with lower rates of usage. TDD has relatively low usage rates (Paez et al., 2018) and BDD is a newer practice. Also, TDD had been identified as the hardest agile practice to learn (Ambler, 2009).

### 6.1.2 Subjects

TDD evaluations subjects were selected by convenience, with initial study subjects being direct contacts of the authors. These subjects were invited and asked to recommend other candidates, using snowball sampling to increase the sample size (Mockus, 2008).

The total number of TDD expert subjects in the study was 17, and the following criteria were applied for acceptance:

- At least 5 years of experience with TDD.
- Acceptable roles were practitioner, mentor, coach, teacher, consultant, manager/supervisor and researcher/academic.

Average experience with TDD was 9.63 years, which is considered high practical experience. Table 5 shows the distribution of roles among the TDD experts (each subject could select more than one role).
Table 5 shows that all TDD experts were practitioners, thus experienced in using TDD, and most were mentors, experienced in helping others apply TDD. This is different from the Scrum study participants, in which there were non-practitioner experts. This might be due to Scrum’s popularity or the fact that TDD is the hardest agile practice to learn according to (Ambler, 2009).

BDD experts’ data was obtained from previous studies (Fontdevila, 2021), by filtering data belonging to experts. This was done to exclude less experienced subjects and preserve the integrity of the data, focused on experts in this study.

The total number of BDD expert subjects in the study was 7. The criteria applied was the same as that for the Scrum study.

### 6.1.3 Statistics and Variable Selection

The Kappa statistics include Cohen’s Kappa, a 2x2 inter-rater reliability statistic (meaning two evaluators for two objects of evaluation), and Fleiss’ Kappa, an extension of Cohen’s Kappa for more than two evaluators, designed for nominal variables (Hallgren, 2012). Other inter-rater reliability statistics include Kendall’s Coefficient of Concordance W (Kitchenham et al., 1995) and ICC (Intra-class correlations) aimed at continuous variables (e.g. interval and ratio)(Hallgren, 2012; Kitchenham et al., 1995).

Kappa statistics have the limitation of misrepresenting the inter-rater reliability in samples of data displaying prevalence or bias. Prevalence means that some of the values in the sample data has significantly higher frequency than the others, and bias means that the marginal distribution of values between evaluators varies significantly.

The problem is caused by the statistic over-adjusting for chance agreement (agreement caused by chance, when evaluators made a random decision between available values). This, in turn, happens because the statistic relies on the values being evenly distributed (fixed-marginal), which is not always the case (Girard, 2016; Hallgren, 2012). This was not an issue in the original context in which the
statistics were defined, which was for patient treatment, in which subjects to be evaluated were selected carefully to be evenly distributed along the spectrum of variable values.

Distributions that are not fixed-marginal are called free-marginal (Hallgren, 2012). Byrt et al. proposed a free-marginal alternative to Cohen's Kappa called PABAK (Prevalence Adjusted Kappa), to correct the prevalence issue, but it only supports two evaluators (Byrt et al., 1993). Gwet mentions a Kappa<sub>BP</sub>, for Brennan & Prediger, also a free-marginal kappa-like statistic, which supports more than two evaluators (Gwet, 2014). Girard, in turn, describes Kappa<sub>BP</sub> as analogous to Bennet's S, and chooses that name stating that it identifies the original proponents (Bennett et al., 1954; Girard, 2016). Girard provides an implementation for a version of Bennet's S which is free-marginal, thus avoiding the prevalence issue, supports missing data and more than two evaluators (Girard, 2016, 2020). Gwet created a free-marginal Gamma statistic, which is very robust when dealing with prevalence and bias, also called AC1 and AC2 depending on the type of data it is applied to (Gwet, 2014).

To summarize, although Fleiss’ Kappa had initially been selected, the prevalence issue made it inapplicable to the study data. For this reason, Bennet’s S and Gwet’s Gamma were used as the inter-rater reliability statistics for the study. For the sake of completeness, and to illustrate its issues with prevalence, the values for Fleiss’ Kappa are also provided for the study data.

Given that these issues had ample impact on the study, the subtle nature of the differences between statistics and the fact that fixed-marginal samples might not be readily available in software engineering studies, these details are provided to support and guide other researchers conducting related research. Finally, following the recommendation by (Hallgren, 2012) confirmed by the experience described, the specific variations and implementations used for each statistic are reported to ensure repeatability and support correct interpretation of the results.

As in the Scrum study, one variable was defined for each UMP metric, and the three statistics were calculated for each variable. In this study there were 22 variables instead of 24, since one metric (Information tailored to audience) had been removed from UMP after the Scrum study and Ratio of roles allowed to adapt was not applicable since TDD and BDD do not define specific roles.

### 6.1.4 Planning

The TDD-BDD study was planned following the recommendations in (Hallgren, 2012) with the following characteristics:

- Design not fully-crossed, meaning that not all evaluators evaluated both practices.
- Both TDD and BDD were evaluated by multiple evaluators.

Hallgren recommends that the protocol used for evaluation be reviewed, particularly scales and value ranges (this had been done after the Scrum study), and that evaluators perform practice evaluations (this was not considered viable given that evaluations took around 1hr, but evaluators were provided example values for Continuous Integration).
The evaluation procedure was guided by a self-administered questionnaire, refined after the Scrum study. The questionnaire was tested by two researchers who performed trial evaluations.

The questionnaire form (http://bit.ly/usabilityoftddform) included:

- A link to the video introduction for UMP.
- Characteristic and metric definitions
- Example application of each metric to Continuous Integration.

The questionnaire was rewritten for this study to improve ease-of-use, particularly clarifying the object of evaluation and improving presentation and organization of characteristics and metrics. Also, as described above, an example evaluation for Continuous Integration was introduced for each metric. shows an example question from the TDD questionnaire.

The *Ratio of roles allowed to adapt* metric, was excluded because it did not apply to TDD or BDD, leaving only 22 metrics.

### 6.2 Study Execution

The TDD evaluation was conducted in the same manner as the Scrum evaluation described in Section 1.7. Subjects received email invitations and eventually reminders to complete their evaluations. All participants completed the evaluations without requiring assistance.

The BDD evaluation was conducted in a similar manner but experts were together in the same room and conducted individual evaluations at the same time.

### 6.3 Data Analysis

To begin the analysis, all data was reviewed and normalized to numerical values. Then, for each metric, a pair of evaluation records was created, one for each practice (i.e. TDD and BDD). Shows an example of the resulting data structure:

<table>
<thead>
<tr>
<th>Practice</th>
<th>Evaluator 1</th>
<th>Evaluator 2</th>
<th>Evaluator 3</th>
<th>Evaluator 4</th>
<th>Evaluator 5</th>
<th>Evaluator 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDD</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>BDD</td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

According to the guidelines provided in (Hallgren, 2012), statistics are reported including their specific version and the implementation used for calculations. In this study, the statistics were calculated using the R Agreement package (Girard, 2020), which unlike other implementations reviewed, supports not fully crossed designs, that is, with missing data because not all evaluators evaluated both practices (as shown in Table 6). The R agreement package also has the advantage of supporting nominal, ordinal and
continuous scales. The statistic variant was selected for each metric according to the metric scale: for nominal variables the standard Kappa-like statistic was used (Gwet’s Gamma, Bennet’s S and Fleiss Kappa) and for ordinal variables the linearly weighted version of the statistic was used (Girard, 2020).

The results of the inter-rater reliability assessment are shown below in Table 7.

For each metric a reliability level is assigned according to the interpretation guidelines provided in (Altman, 1991). Table 8 presents the details for each level, including the count of metrics in each level.

### 6.4 Results and Conclusions

For 10 out of 22 metrics assessed the study produced positive results (moderate to very good). These results are not comparable to those produced by the Scrum study, among other reasons because the statistics used in this study apply corrections for agreement due to chance, thus being more demanding than James’ \( r_{WG} \).

Each metric with poor or fair results was analyzed together with the available data to identify potential causes for such low reliability scores, producing the following potential causes:

- **Intrinsic subjectivity**: for some metrics like Recognized purpose, User attractiveness rating and User satisfaction rating are evidently subjective and dependent on the personal experience and perspective of the evaluators, and this might explain the dispersion in the sample.

- **Subtle metric semantics**: in the case of some metrics like Appropriateness of name the nuances of meaning might have affected the results, since some evaluators appear to have considered literally only the name, others how well the name relates to the purpose (it is in the context of the Self-evident purpose characteristic), and yet others considered the perceptions of users in general. It is interesting to note that in the Scrum study this metric also had a low reliability statistic result, and it had been attributed to a complicated and extensive scale (which was improved before this study, and thus might not have been the only cause).

The same reasoning seems to be applicable to Time required to learn to perform, although a reference to the Dreyfus model was made explicit, it is not likely that evaluators have a unified perception of what it means to perform at a basic level of ability.

For the Cost of Incorrect Adoption metric, for some evaluators this was related to the product, for others to the developer’s experience, for others with the difficulty of detecting adoption errors, and yet for others with risk of frustration and practice abandonment.

Finally, something similar might have happened with the Defines adaptation points metric, in that a wide spectrum of considerations were presented in the evaluators’ comments.

- **Scale**: The results do not show any obvious remaining scale issue with the improvements applied after the Scrum study. For example, Number of specific conceptual definitions, which has a scale of
positive integer values shows a good Gamma value. Nonetheless, there is not sufficient information to confirm that there are no scale issues.

- **Context sensitivity**: For some metrics, like People feedback, evaluator provided clear hints that point to context sensitivity.

In general, the study shows significant differences in metric reliability. As expected, binomial metrics (e.g. yes/no) show higher reliability (this is in part due to the fact that the number of values bears on the estimated agreement due to chance calculated by the statistics). On the other hand, subjective and context sensitive metrics show lower reliability values.

Another interesting pattern is that certain metrics have high agreement but only for some of the practices or processes under study. For example, Scrum presents high agreement in all its metrics for the Feedback characteristic, and TDD for its Timeliness of Feedback metric (with unanimous agreement on immediacy). This might point to evaluators having special affinity with some UMP characteristics or metrics (for example, Feedback is one of the most popular UMP characteristic among the participants of different studies conducted for this research), or simply that the practice or process itself has a strong affinity for those metrics (e.g. feedback is one of Extreme Programming’s core values (Beck & Andres, 2004)).

Finally, in both studies there is evidence that metric reliability varies with the object under study (Scrum, TDD, BDD), and this seems to point towards true variance, that is, variance in what evaluators perceive about the object beyond the features of the UMP and the questionnaires.

### 6.5 Threats to validity

The threats to validity for the TDD-BDD study are presented in this section following the categorization by (Wohlin et al., 2012).

**Threats to construct validity**: Construct validity for this study might have been affected by questionnaire design. A focus group study was conducted to improve the clarity of UMP characteristics and metrics definitions (Fontdevila, 2021). Also, internal reviews were conducted and the questionnaire was refined accordingly. Finally, after the Scrum study, evaluator feedback was incorporated, a new UMP version (3.2) was produced and the questionnaire was improved significantly and then reviewed and tested internally by two researchers.

**Threats to internal validity**: Given the relatively simple nature of the statistics applied and that the study design followed the recommendations in (Hallgren, 2012), the study results are considered valid. In particular, specific statistics were selected that compensate for chance agreement and that deal effectively with prevalence and bias issues present in the sample.

**Threats to external validity**: The main restrictions to generalizability are the small sample size and the fact that it cannot be considered representative although it was defined following specific criteria (Kitchenham & Pfleeger, 2008). Given that the study required experts and that evaluations took
considerable time (around 1hr), this is considered reasonable. Experts were carefully reviewed to ensure they met the inclusion criteria and the distribution of participant’s profiles was appropriate for the study objectives, since most participants had roles (practitioner and coach) that put them close to actual practice by users.

This study complements the Scrum study, and although the metrics cannot be composed for integrated evaluation, the availability of data on three objects of evaluation (Scrum, TDD, BDD) provided more insights than those achieved with the Scrum study.

Threats to conclusion validity: The number of observations limits the conclusion validity of this study, and although it complements the Scrum study, there is not enough evidence to promote drastic modifications to UMP (e.g. removing all metrics with poor reliability ratings). Still, the results of these studies were among the main inputs for the metric selection recommendations established in UMP, see (Fontdevila, 2021).

7 Conclusions and future work

This paper presents the UMP and the inter-rater reliability studies conducted to assess the reliability of its metrics. Both the Scrum study and the TDD-BDD study produced valuable results, showing varying reliability results among the different metrics, which seem to depend mostly on object of evaluation (Scrum, TDD, BDD), metric subjectivity and context sensitivity. Of particular interest is the affinity between some metrics and specific objects of evaluation. For example, all Feedback metrics produced very high reliability values for Scrum, and the Timeliness of Feedback metric produced unanimous response from the evaluators stating it was immediate. This is consistent with explicitly expressed XP values (Beck & Andres, 2004). These findings have prompted specific recommendations for UMP metric selection, separating metrics in three categories: core, recommended and complementary, based on their reliability and other considerations, like feedback from an expert focus group conducted to gather feedback and identify improvement opportunities. This metric categorization also enhances UMP ease of use and promotes reduced evaluation time and complexity. The categorization and focus group study are described in detail in (Fontdevila, 2021). They also prompted the deletion of a metric and changes in scale for others.

The research process also provided an opportunity for exercising recommendations on statistic selection and issues with free-marginal statistics were explored in depth, including dealing with chance agreement, prevalence, and bias. This experience might help other researchers by providing concrete examples of the ramifications of statistics’ selection in the research process.

A key aspect of this research is that it was conducted with experts who are industry professionals, and thus match the candidate user base for the UMP. These experts provided insights based on their experience working in and mentoring software development teams.
UMP was designed to support process and practice adoption and evolution by providing insights into adoption risks and improvement opportunities. Future publications will focus on UMP’s utility evaluation, that is, its ability to serve its users by fulfilling its objectives, and providing easy access tools for practitioners to use the UMP. For practitioner access a website is planned to provide materials, report research results in formats that are appropriate for direct industry application, and promote further data collection.

Declarations

Acknowledgements

The research presented in this paper was developed in the context of the following projects: ADAGIO (Consejería de Educación, Cultura y Deportes de la JCCM, España, SBPLY/21/180501/000061) and AETHER-UCLM (MICINN, Spain, PID2020-112540RB-C42), and the Adopción de Prácticas project at Universidad Nacional de Tres de Febrero, Caseros, Argentina.

Authors’ Contribution

Diego Fontdevila and Marcela Genero. wrote the main manuscript, Diego Fontdevila, Nicolás Paez and Alejandro Oliveros designed the studies and reviewed material, Diego Fontdevila and Nicolás Paez conducted the studies. All authors reviewed the manuscript.

Funding

The research presented in this paper received funding from the following projects: ADAGIO (Consejería de Educación, Cultura y Deportes de la JCCM, España, SBPLY/21/180501/000061) and AETHER-UCLM (MICINN, Spain, PID2020-112540RB-C42), and the Adopción de Prácticas project at Universidad Nacional de Tres de Febrero, Caseros, Argentina.

Conflict of Interest

The authors identify no conflict of interest in this research.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References


**Tables**

Tables 4 and 7-8 are available in the Supplementary Files section

**Figures**
Figure 1

Research strategy overview
Figure 2

UMP Structure
6.2 Metric: Feedback timeliness *

Measures the timeliness of the feedback as perceived by the actor with respect to the action performed and the consequent actions that need to be performed

- Immediate
- Prompt
- Delayed
- Non existent

Figure 3

Scrum questionnaire example question

Example: "Feedback timeliness" of Continuous Integration

If the practice were Continuous Integration, and you considered that CI feedback is timely by design (it is continuous, after all), you might evaluate “Feedback timeliness” as “Prompt” and write the following observation “The CI build feedback is expected to be fast, and the related 10 "build practice supports this"

5.1.1 - How would you rate the timeliness of feedback in TDD? *

- Immediate
- Prompt
- Delayed
- Non existent

Figure 4

Example question from the TDD questionnaire
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.docx