Applying the Psychometric Theory to Questionnaire-Based Appraisals for Software Process Improvement

Garcia, Ivan., Pacheco, Carla & Andrade, Gabriel
Postgraduate Department
Technological University of the Mixtec Region
Huajuapan de Leon, Oaxaca (Mexico) http://www.utm.mx
{ivan@mixteco.utm.mx; leninca@mixteco.utm.mx; gandrade@mixteco.utm.mx}

Abstract—Software Process Assessments is a key factor within organizations to determine their current capability/maturity level and to adopt a Software Process Improvement initiative. Their chance of success using a standard model is determined by a reliable assessment of their current processes and in determining which processes need to be improved. However, the current existing automated tools for process assessments, can not verify the authenticity of answers, and are therefore limited in their reliability level depending only on the employees’ responses. This paper presents our research in psychometric theory applied to questionnaire-based appraisals; to determine the feasibility of combining them together in order to develop a reliable assessment mechanism providing more reliable evidence about the organizations’ current maturity/capability level. We have found that current existing tools implicitly include some elements of psychometric theory but more can be included. Thus, this paper presents a first attempt to formally integrate the psychometric theory into the questionnaire-based appraisals.

Keywords—Appraisals, software process improvement, questionnaire-based appraisals, psychometric theory

I. INTRODUCTION

Over the last years, more organizations have become concerned about Software Process Improvement (SPI). Consequently they have decided to adopt international initiatives related to SPI, such as Capability Maturity Model Integration for Development v1.2 (CMMI-DEV) [23] and ISO/IEC 15504:2004 [1]. In Mexico, SPI has become more important with the emergence of Small and Medium Enterprises (SMEs) dedicated to Software Development representing over 80% of all enterprises engaged in this activity (see Table I) [22].

<table>
<thead>
<tr>
<th>Size</th>
<th>Number of employees</th>
<th>Total number of organizations</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro</td>
<td>1 to 19</td>
<td>619</td>
<td>41%</td>
</tr>
<tr>
<td>Small</td>
<td>11 to 50</td>
<td>629</td>
<td>42%</td>
</tr>
<tr>
<td>Medium</td>
<td>51 to 100</td>
<td>130</td>
<td>9%</td>
</tr>
<tr>
<td>Large</td>
<td>More than 100</td>
<td>114</td>
<td>8%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1492</td>
<td>100%</td>
</tr>
</tbody>
</table>

With the aim of enabling those enterprises to compete with larger organizations in Mexico, most of them have shown interest in SPI models. These advances are verified with information from the same study. According to the results, about 20% of the organizations exceed capability level 2 of SPICE appraisal model [12].

However, the total balance of the same study shows that the average maturity level of all organizations is less than 1, as shown in Figure 1. This means that the most of the organizations do not perform their processes causing among other things to not meet deadlines of projects, that the products delivered do not meet the expectations of clients or projects that exceed budget constraints. According to the study there are bases for improving processes of these companies, however they must invest time and resources in its implementation in the organizations. Unfortunately many organizations do not consider the adoption of any process model in their plans, especially by the idea that the implementation of these models is very expensive and time consuming. However, researches such as those made in [27][21][17][9][24] show that there are viable alternatives for SMEs wishing to adopt a process model.

The first step to apply a software process model consists on assess the current organization’s processes. Hence choosing and adequate appraisal method becomes a crucial step to determine its current status in an accurate and reliable way. Nowadays, there are automated tools available which allow organizations to assess their current processes using questionnaire-based appraisals (QBA) [14]. Nevertheless all of these tools can be effective only according to the information provided by the organizations in their assessments. Several factors can affect the appraisal including employee’s skills, the questionnaire’s degree of difficulty and different answering styles like randomly choosing answers. Moreover, it can be possible to falsify answers knowing specific activities of the model being evaluated even if the organization does not implement those processes only to obtain a better degree in the organization’s current capability/maturity level.

Figure 1. Maturity level by region
II. QUESTIONNAIRE-BASED APPRAISALS OVERVIEW

Developing the questionnaire is an important step in the assessment process, and much depends on the easiness in collecting data and speed on retrieving and analyzing information. Questionnaires can be classified into open and closed questions. An open-question provides more information than a closed one. The complexity of analyzing data provided by open questions, however, is higher than those in closed-questions [28]. Moreover, a closed-question provides less information but its results can be more easily analyzed and are obtained faster than with the open one. Consequently most of the research made on developing questionnaires has focused on closed questions.

In order to propose a new approach on developing appraisal questionnaires, a review was performed of the questionnaires available in the literature.

The first questionnaire to be reviewed was the SEI Maturity Questionnaire. This questionnaire can be applied in several appraisal methods like CBA-IPI [6] or CMM v1.1 [18] and it does not provide a full assessment of an organization’s processes instead it was designed to be used as an appraisal tool for other methods like the above. However this questionnaire only provides general information about the process without analysis on not performed practices. Another disadvantage is that this questionnaire is limited in the number of responses that can be selected: Yes, No, Does not Apply and Don’t Know. In fact, there are only two options - Yes and No- because Does not Apply and Don’t Know are used to validate the application of the questionnaire. Using the maturity questionnaire limits the information to two extreme ends: Yes, if the practice is performed and No if the practice is not performed. Therefore, it does not leave room for intermediate points. There are, for example, no options to pick up cases where practices are performed but rarely documented or when they are not documented at all. This type of question cannot be addressed with the options provided in the Maturity Questionnaire.

Another study reviewed was the software improvement model proposed by the ISPI. This model was used by [3] and [2] in their research. For the appraisal stage, they proposed a questionnaire structure using five types of responses: Always when the practice is documented and performed between 100% – 75% of the time, More often when the practice is documented and performed between 74% – 50% of the time, Sometimes when the practice is not documented and is performed between 49% – 25% of the time, Rarely when the practice could be documented or not and is performed between 25% - 1% of the time, and Never when the practice is performed. These will enable us to know the extent to which each practice is performed.

In order to perform an appraisal, the questionnaire uses closed questions and limits the number of possible answers to seven. These are organized as follows:

- **Five level-perform-answers**: Always, Usually, Sometimes, Rarely if ever, and Never. These will enable us to know the extent to which each practice is performed.
- **Two validity-answers**: Don’t Know and Not Apply. These will be used to appraise the validation of the questions, to validate the correctness of the question, and to check the syntax of questions.
- **Additional information spaces (Comments)**: To extract supplementary background information. It is mandatory to write some comments when checking any of the validity-answers.

This questionnaire has also been applied on Small and Medium Enterprises (SMEs) for Software Process Management and for Software Acquisition Management having satisfactory results [11].

Based on the previously reviewed literature we can conclude a lot of research has been made on improving appraisal questionnaires reliability by increasing the number of possible answers. However there is still a lot of work to do to improve the reliability of each question, to verify the relationship between questions and individual weight of each question based on its difficult level.

III. A BRIEF APPROACH TO PSYCHOMETRIC THEORY: THE LIKERT SCALE

The Likert scale is a psychometric scaling method developed by Rensis Likert. According to Likert scale a questionnaire’s measurement is the sum of responses on several Likert items[1]. We note several characteristics or features that define a Likert scale [26]:

1) The scale contains several items.
2) Response levels are arranged horizontally.
3) Response levels are anchored with consecutive integers.
4) Response levels are also anchored with verbal labels which connote more-or-less evenly-spaced gradations.
5) Verbal labels are bivalent and symmetrical about a neutral middle, and
6) In Likert’s usage, the scale always measures attitude in terms of level of agreement/disagreement to a target statement.

Criterion 5 usually means there are an odd number of response levels. Typically the number is 5, though sometimes 7, 9, or 11 levels are used. After the questionnaire is completed, each item may be analyzed separately or in some cases item responses may be

---

1 We denote by Likert item or simply item to each one of the questions being assessed in the questionnaire.
summed to create a score for a group of items. Hence Likert scales are often called summative scales.

As we can see in Figure 2 it’s a common practice in appraisal questionnaires to implement a Likert scale like the appraisal questionnaire used by SysProVal [10] which meets the following Likert-scale criteria:

1) Sometimes is the neutral point in the appraisal questionnaire. The neutral point is related to the degree to which an activity is performed in the organization.

2) Response levels are arranged horizontally from highest (Always) to the lowest (Never).

3) The scale contains several items. Only five items (Always, Usually, Sometimes, Rarely if ever and Never) are related to the Likert scale.

4) Item responses with verbal labels which connote more- or less evenly-spaced gradations.

5) Verbal labels are bivalent and symmetrical about a neutral middle (The neutral point).

Figure 2. An Appraisal Questionnaire using a Likert Scale

We can observe that currently existing questionnaires have already applied fundamentals of psychometric theory. But still a lot of features can be improved in appraisal questionnaires as we show with our findings in the next section.

IV. FUNDAMENTALS OF PSYCHOMETRIC THEORY APPLIED TO APPRAISAL QUESTIONNAIRES

Based on the literature reviewed, we can see a breakthrough in the processes involved to develop appraisal questionnaires focused to improve granularity of responses. The aim of this is to increase the answer’s distribution used to determine capability/maturity levels. However we think there is still a lot of work to improve further to increase accuracy and reliability of appraisal. This improvement can be achieved applying statistical techniques developed for over 50 years by psychometric theory. We believe this is possible because both the processes of organizations as personality processes share many features in its conceptual framework: both are internal states, both are subject to social desirability factors and both can be distorted when questioned about them. Psychometric techniques were developed to control these problems of measurement.

However, it is not sufficient that both processes share common factors in its conceptual framework for the proper application of psychometric principles. It is also necessary determine that these principles are applicable in developing assessment questionnaires and provide significant improvements in assessments, which can not have been achieved without applying these principles. Below is a list of factors considered as capable of improvement in existing questionnaires, based on developments made by psychometric theory in these fields, related weaknesses in current evaluation questionnaires and feasible suggestions for improvement applied using psychometric theory:

- **Differential Item Weighting.** The method used so far for the evaluation of questionnaires is to assign equal weight to all the questions, meaning that each question contributes equally to the final score of a test like in the SEI maturity questionnaire [29] (see Figure 3). This way of assigning weights however, does not reflect the real contribution of each activity to the total process as seen in the way of assigning ratings in existing appraisal questionnaires, which assigns same percentage of the level of the extent to which the practice is carried out for each question, however, some practices may be more critical to an organizations success than others. A common practice in other areas is to assign weights according to the experience of the questionnaire developers. To improve the quality of evaluations, methods have been developed as described in [7] and [15] to tests design criteria and subsequent analysis, and techniques have been proposed for assigning weights to the items as proposed in [8] based on statistical techniques, which guarantees the objectivity of the

![Figure 3. SEI Maturity Questionnaire format](image-url)
assignments of weights. Especially in an area where is difficult to quantify the value of activities and establish appropriate metrics as Software Process assessment, the contribution offered by psychometric theories may be very useful. This is especially true when observing that in the questionnaires analyzed all questions provide the same value to the overall result of the assessments, which are based on calculating only the total percentage of activities regardless of their difficult level or a possible relationship between them.

**Difficult level.-** Difficult level of an item on a questionnaire is related to the proportion of assessed that guesses the right answer, higher values means that the difficult value decreases. For evaluation purposes, items that have a value of 0.0 or 1.0 are not useful or descriptive. Items that have a value of 0.5 provided the best differentiation between those that are right answers and those that fail [13]. Difficult value of the questions is used to determine the overall difficulty of the entire test. Still another useful implication of the difficult level says that it may not be better, a test of 10 items with a difficulty value of 0.5 in each question than a test of a single item with a difficult value of 0.5, since in both cases 50% of the sample would pass the test while 50% would not. This implication allows an important observation: That it is possible to develop short and specific test in which all items have a similar level of difficulty or take a much longer test whose overall difficulty level remains the same. This would allow us to develop shorter tests when they want to evaluate a specific process which provide us enough information and larger and highly differentiated tests in terms of difficult levels when we need to assess maturity levels, thus adapting more efficient tests while providing the same results for the purpose of our evaluation.

**Possibility of guessing answers.-** There are appraisal questionnaires that include items in which it is not required that the person knows the correct answer of the items. This it is the type of items used in the tests of multiple options. Tornimbeni [25] emphasizes that in this type of test guessing probabilities plays an important role in the final scores. This probability is inversely related to the number of choices for each question, thus in a dichotomic item (with true-false, correct incorrect, yes- no options) the probability of obtaining a correct answer by guessing is about 50%. In a test with several options the probability of finding correct answers obtained at random falls according to the number of options available. According to [4] reliability of tests is greater as choices increases for each question in a limited rank.

**Obtained results interpretation.-** It is important to consider in questionnaires the minimum degree required to be given by approved tests. The possibility of obtaining correct answers at random plays an important role on determining the final score. For example, if in a test each question has five alternatives, the probability of obtaining a correct answer by guessing for each answer will be of 0.20, if the test consists of 100 questions, in average total of 0.20*100=20 questions can at random be guessed right. This value fixes a minimum limit to consider to given by an approved a test. In addition, the final score of a test is obtained generally from a sum of the correct answers in the test. Nevertheless, this evaluation by itself provides little information on the abilities of a particular individual in the area of evaluation of the test. According to [16] the original scores are not sufficient to give an interpretation of the results which are limited by the following reasons: (1) Lack off absolute zero (the zero in scores do not indicate absence of skills because an exhaustive or representative sampling of it was never made of their “operational indicators”). and (2) they do not have constant units of measurement like other units such as length (measured in meters) or weight (measured in kilograms). The commonly used strategy to give some meaning to the original scores consists in comparing individual results with a reference group’s scores in the same test.

Moreover, it is important to consider other key factors in the development phase of appraisal questionnaires and when applying them in process assessments, as is described in [5] and [26]:

- Appraisal Questionnaires can include more than one type of question. Thus appraisals will be less monotonous.
- It is important to avoid all kind of regularity in the questionnaire’s answers distribution. A good practice is to place them randomly in each question.
- The more items included in the test, the higher the reliability will be.
- The answering process must be easy to understand.
- The questionnaire’s instructions must be clear, complete and accurate and must include all of the following:
  - Available time to finish the appraisal.
  - How to answer the questions (filling up instructions).
  - Instructions must include some easy examples of the questions included.

From the previous review we can observe that existing appraisal tools apply psychometric principles, most of
them focused on the analysis of results and mistakes made in the organizations. Although this is a breakthrough there still exist many factors that can be improved as shown in Tables II and III.

**TABLE II. COMPARATIVE TABLE BASED ON INCLUDED PSYCHOMETRIC FEATURES ON APPRAISAL QUESTIONNAIRES.**

<table>
<thead>
<tr>
<th>Appraisal Questionnaires</th>
<th>SEI Maturity Questionnaire</th>
<th>CMMI Interpretive Guidance Project [4]</th>
<th>Two-Phase Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert scale-based assessment method</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Differential Item Weighting</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Test based on activity’s difficult level</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Difficult level of questions</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Possibility of guessing answers</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Obtained results interpretation</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

**TABLE III. COMPARATIVE TABLE BASED ON PSYCHOMETRIC FEATURES INCLUDED ON AUTOMATED APPRAISAL TOOLS.**

<table>
<thead>
<tr>
<th>Automated Appraisal Tools</th>
<th>CMM- Quest</th>
<th>Spice I- 2-1</th>
<th>SysProVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert scale based assessment method</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Differential Item Weighting</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Test based on activity’s difficult level</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Difficult level of questions</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Possibility of guessing answers</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Obtained results interpretation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

V. DEVELOPING THE APPRAISAL QUESTIONNAIRE

In order to apply the psychometric principles analyzed in Section 4, an Appraisal Questionnaire was developed to assess the activities of the Requirements Management Process Area of CMMI for Development [24] maturity model. The first step proposed was to reduce the possibility of guessing answers. As seen in Figure 4, existing questionnaires with a Likert scale-based format are easily manipulated because the answers’ distribution remains in the same place and in the same kind of answers. For example in questionnaires A and B, an answer Yes always provides the higher score throughout the entire test. To reduce the possibility of guessing answers and to increase the type of questions included, three types of questions were applied in the appraisal questionnaire developed:

- Type A include questions with multiple options and only one correct answer as shown in Figure 5. These questions include activities related to each of the Specific Practices of Requirements Management and others Process Areas. Also, the order of the answers can be changed to avoid existing patterns in the distribution of answers throughout the whole test.

- Type B includes questions with multiple answers. These questions are related to Subpractices and Work Products of each Specific Practice. To score a right answer you must select all of the correct answers depending of the number of Subpractices or Work Products as shown in Figure 6. To avoid guessing answers, wrong answers decrease each question’s score.

- Type C includes questions with only one answer. These questions include knowledge of total Specific and Generic Practices and Goals. In these questions a numeric answer is expected (see Figure 7).

![Figure 4. Correct answers in Likert scale based questionnaires](image)

![Figure 5. Type A Answers](image)

![Figure 6. Type B Answers](image)
The second step includes Differential Item Weighting and Results Interpretation. The proposed questionnaire includes questions about Practices and Work Products to assess the model’s implementation in the organization, and questions about Requirements Management goals to assess the model’s understanding. Model implementation plays a more important role in capability assessments (CMMI’s Level 1), while model understanding is related to maturity assessments (CMMI’s Level 2 and above). Moreover, the appraisal includes several scores related to each Specific Practice activities and an overall score obtained with the scores of each Specific Practice. This way we can provide a general score which enables organizations to assess the current levels of their processes, and Specific Practice scores to determine their best practices and which areas that need to be improved.

VI. CONCLUSIONS AND FUTURE WORK

From the above analysis, we can see that existing appraisal tools already apply statistical techniques to determine the final assessment, and questionnaires take into account some factors considered important by the theory in the development of psychometric tests. But there still has not been developed a questionnaire specifically using psychometric principles, with the aim of improving the assessment process. This may represent major advances for SMEs in implementing a SPI model requiring less time for assessments and having knowledge of their real situation and what they should improve. Future work includes focusing on a specific level of difficulty and developing tools to perform automated assessments with the proposed techniques.

VII. ACKNOWLEDGEMENT

This work is part of a research to develop an appraisal questionnaire of SPI activities based on psychometric theory in the Postgraduate Department, Technological University of the Mixtec Region, Huajuapan de León, Oaxaca (Mexico) http://www.utm.mx. The main goal of this research, is to develop an assessment methodology that can be used with questionnaires-based appraisals, for any reference model applying Psychometric Theory. This kind of methodology would increase the reliability of assessments and could be included in software tools to automate the process.

REFERENCES


