Setbacks of the Development of a Concept Inventory for Scrum: Contributions from Item Response Theory

Walter Aoiama Nagai[1], Rui M. Lima[2], Diana Mesquita[3]

[1] Institute of Technological Sciences, Federal University of Itajubá, Itabira Campus, Brazil
[2] Department of Production and Systems, Centro ALGORITMI, School of Engineering, University of Minho, Guimarães, Portugal
[3] Universidade Católica Portuguesa, Faculty of Education and Psychology, Porto, Portugal

Email: walternagai@unifei.edu.br, rml@dps.uminho.pt, dmesquita@ucp.pt

DOI: https://doi.org/10.5281/zenodo.5070578

Abstract

Scrum is the more common framework for agile project management. Agile project management requires frequent feedbacks and delivered items in projects with dynamic requirements and changes. Training learners in Scrum permits building agility in solving problems and teamwork competencies. Measuring training effectiveness is essential to identify students' learning lacks or misconceptions to improve the training outcomes. To assess the development of competences, it is possible to use concept Inventories, which are an essential educational tool to observe students' learning gain between two moments, before and after training. Additionally, the Item Response Theory may be applied to concept inventory items to identify latent characteristics as guessing, difficulty, and discriminant values. Guessing is related to an arbitrary answer to one question and gets the correct answer with common learner knowledge. Difficulty characteristic is related to student knowledge level to one question. Discriminant characteristic considers that learners with high score get accurate answers to the questions. Thus, this work aims to present some of the main setbacks of developing a concept inventory for Scrum, supported by the Item Response Theory. In this way, other researchers may understand how to develop a concept inventory and some of the main obstacles they may have to overcome or avoid. The Item Response Theory offers some indexes and criteria values to each latent characteristic to improve the concept inventory questions. Therefore, this work focuses on the process of conceptualizing, building, applying, and improving a Scrum Concept Inventory in a training situation with engineering students.

Keywords: Scrum, Agile Project Management, Training, Concept Inventory, Item Response Theory.

1 Introduction

Nowadays, engineers face challenges that require a solid foundation in engineering competences such as teamwork, project management, interdisciplinary problem-solving, and oral/written communication (Mesquita et al., 2015). According to Project Management Institute (2013), Project Management is an area of knowledge that mobilizes management concepts, tools, and methods for planning, executing, and closing projects in an efficient way.

The realization of a project varies from determinable and probable to indeterminable and uncertain. A project is considered determinable if characterized by clear, successful procedures and based on similar past projects, such as cars, electrical appliances, or houses. When a project requires a new or innovative design, the people involved can carry out exploratory, collaborative actions and create new solutions, making the project indeterminable and highly uncertain. Examples of people involved who face high uncertainty jobs include software systems engineers, product designers, doctors, teachers, lawyers, and engineers (Project Management Institute & Agile Alliance, 2018).

Traditional predictive approaches applicable to determinable projects attempt to determine the most advanced requirements and control changes through a change request process. In indeterminable projects, it is necessary to explore and carry out actions in short cycles so that the people involved adapt quickly based on evaluation and feedback. An agile approach has dynamic requirements during the project and frequent deliveries of items done. In this approach, Scrum is currently one of the most common projects management frameworks, focusing on managing projects with frequent changes driven by the client's needs and desires. Briefly, according to Sliger (2011), Scrum is an agile method of quickly, iterative and incremental delivery of products that uses frequent feedback and collaborative decision making.

Scrum training allows developing agile collaborative and teamwork competences in solving problems and continually improving products. Adding, training is a process to design, deliver, and implement a learning program for learners about a specific subject or concept. Still, it is necessary to measure the learning before and after the process. According to Lindell et al. (2007), the Concept Inventory (CI) is an instrument to measure learning in education or training situations. Design CI to assess learners' conceptual knowledge or misconceptions as multiple-choice questions (MCQ) to test learners' understanding of concepts. A prominent example, the Force Concept Inventory (FCI), designed by Hestenes et al. (1992), started developing research-based distracter-driven multiple-choice instruments.

This work aims to show details of the process design of a new concept inventory for Scrum and the main obstacles found in the process, identified mainly by analyzing the answers using the Item Response theory to identify latent characteristics - difficulty, discriminant, and guessing.

2 Scrum Concept Inventory

Concept Inventories are a promising tool test for measuring learning gains in specific areas of the curriculum. Tests necessarily measure the type of development in students that a learning gain test also measures. Sands et al. (2018) divided the questions into crucial concepts regarding a subject. Each of them has a correct answer and some incorrect answers or distractions. Identifying misconceptions or mistakes is essential to characterize a student's understanding, becoming a central point to build a valid concept inventory with the right questions and appropriate distractors.

Make the concept inventory's application in two different moments, one moment before the instruction of the concepts, also called pre-test or pre-instruction or pre-training, and another moment after, named post-test, or post-instruction or post-training (Madsen et al., 2017). This allows comparing the two moments' scores to assess the effectiveness of the training performed by an instructor. Concept inventory aims to evaluate the understanding and the implication of concepts differently from the final exams that test various subjects.

Using Scrum Guide designed by Schwaber & Sutherland (2017) as the first source, the research team developed the Scrum Concept Inventory (SCI) with 20 questions in multiple-choice format. Each question had one or more right answers and wrong answers, also known as distractors. The Scrum Guide's choice to create the Scrum concept inventory was motivated because it is the Scrum creators' primary material. The whole community always suggested a continuous improvement Scrum Guide focusing the topic's importance and relevance of Scrum items and events. The SCI has 20 questions divided into five parts: (i) Scrum framework with four questions; (ii) Scrum Team with two questions; (iii) Scrum Team roles with four questions, (iv) Scrum events with seven questions; and (v) Scrum artifacts with three questions. The authors of the study designed all the questions.

Figure 1 describes the process of the design, application, and collection results phases of the Scrum Concept Inventory (SCI) in this study. Before the concept inventory application, the test was designed using the Scrum Guide concepts. The concept inventory application phase was developed in an online training using the Zoom video conference tool. Finally, participants' responses were collected in pre-instruction and post-instruction, in a digital form (Google forms) to posterior analysis.

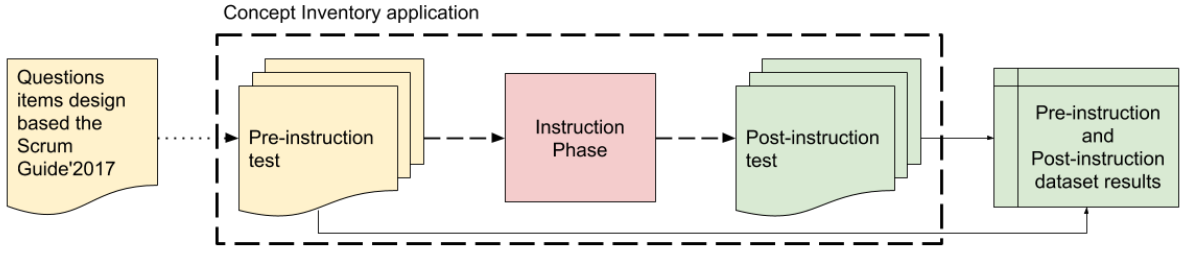

Figure 1. Phases of design, application, and collect results.

The authors created a Scrum Training with an expected time duration between 2.5 and 3 hours to apply the SCI. As commented before, this training's primary material was the Scrum Guide, and participants download it after the training. The Scrum Training used Google solutions like Sheets and Slides to simulate the Scrum events and teamwork communication, respectively. The training theme was about building a city inspired by the Lego4Scrum training (Krivitsky, 2019). The Scrum Training was delivered in four higher education institutes in Brazil and Portugal in 2020, with 51 participants' total. The participants were undergraduate and master's students with different education levels (all with a Bachelor's), ages, gender, and profiles. All participants considered in this study responded to the pre-instruction and post-instruction tests. Their results were collected using a Forms solution, and after the post-instruction, the participants receive their score performance.

3 Item Response Theory basics

A test is a prevalent way to assess learners' learning after the training or teaching, being the obtained score a way to represent the learning result. The Classic Test Theory (CTT) analyses the learners' scores and determines the best or worst results in the same test. According to Rabelo (2013), two learners could have the same score, but their ability levels could be different answering the questions. CTT does not consider the assessed latent features like guessing or question discriminant, for example. The Item Response Theory (IRT) considers the item's test as elements to scores' effects of the assessed' abilities with latent features, not only test score. In IRT, the score result is related to the demonstrated ability level of the assessed.

An IRT model considers that a learner's probability of getting the correct answer is related to his ability level. A high ability should have a high chance, and a low capacity a low likelihood (Sijtsma & Junker, 2006). The IRT permits calculating the learner proficiency or competency according to the test's abilities to compare different learners. The IRT model with three parameters (3PL) is mainly used to estimate the learner's probability in the test's items. The IRT with the 3PL model, as defined in Rabelo (2013):

$$P(X_{ji} = 1 | j) = c_i + \frac{(1 - c_i)}{1 + e^{-Da_i(\theta_j - b_i)}}$$

X_{ji} is the j answer of i item that equals 1 when the learner answer is correct, otherwise 0. The main three parameters of the 3PL model are the a_i, b_i, and c_i. The a_i is the discriminant parameter of the i item that considers that learners with high scores get the correct answers in the easy items. The b_i is the difficult parameter of the i item related to the ability level that considers the necessary learners' ability to get the correct answer to the test item. The c_i represents the learner's guessing feature, which means learners could risk an arbitrary response to the test item and get the correct answer. The θ_j represents the j learner's ability level. The e represents the exponential math function, and D is a scale factor. In IRT, the parameter a_i has a positive value greater than zero. The parameter b_i has values between negative and positive values. The c_i parameter has a variation between zero and one, representing 0% to 100%, respectively.

There are other models with one parameter or two parameters, where difficult and discriminant are the latent characteristics evaluated. In this study, we considered the three parameters model because it had the latent guessing characteristic too.

4 Analysis

With the SCI and the participants' responses, we analyzed the dataset results to compare the performances between two stages, pre-training, and post-training. All the analyses were developed using the R language and some packages like mirt (Chalmers, 2012), ltm (Rizopoulos, 2006), and the R Studio. Before the IRT application, the data from one participant was removed because her/his response was entirely correct in the post-training test. If one participant has a wrong answer to all questions, it would be removed from the dataset. It is similar when there are missing values to any question of the concept inventory. This pre-processing is necessary to avoid bias in the IRT algorithm. Therefore, this work considered the post-training dataset from 50 participants to apply the IRT algorithm.

Figure 2 shows the participants' performance between the two stages, pre-training, and post-training. Two points are outliers of the pre-training results. There are no outliers in the post-training.

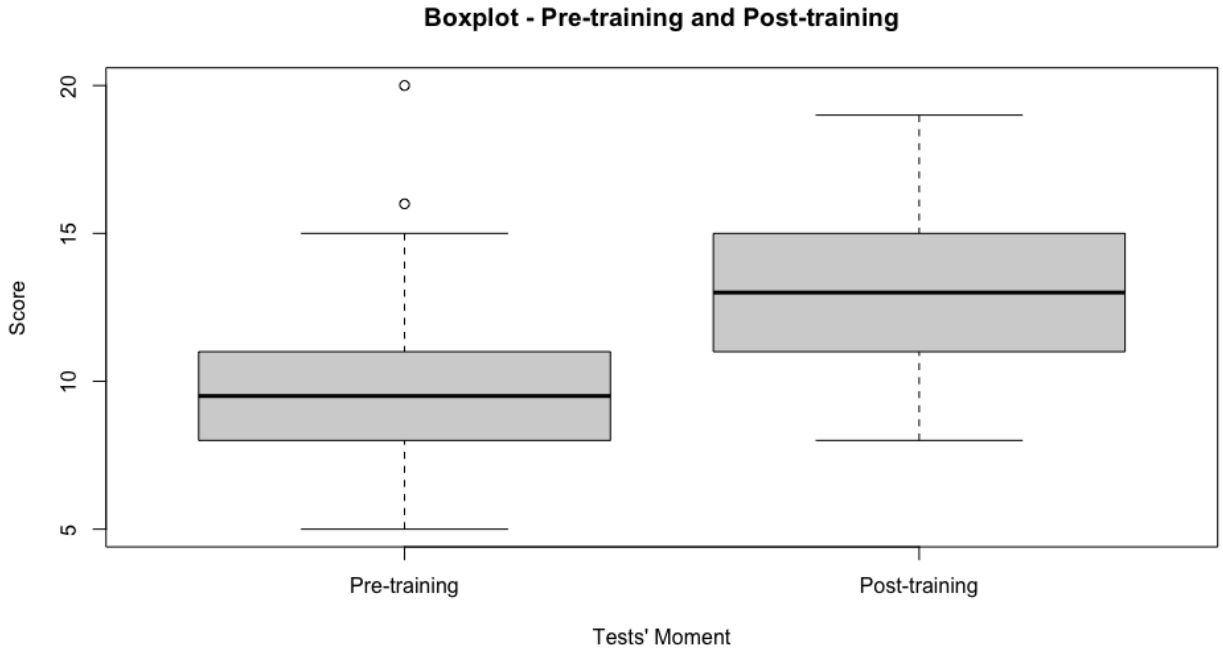

Figure 2. Pre-training and post-training scores' boxplot.

Table 2 shows the descriptive statistics of pre-training and post-training scores for each stage. It was considering the means $\mu_{pre\text{-}training}$ and $\mu_{post\text{-}training}$, was applied paired *t*-student to verify if hypothesis h_0: $\mu_{pre\text{-}training}=\mu_{post\text{-}training}$ is accepted or not. The p-value of paired t-student is least than 0.01, which means that hypothesis h_0 is rejected.

Table 2 Descriptive statistics.

	Min. Score	Max. Score	Median	Mean ± Standard Deviation
Pre-training	5	20	9.5	9.9 ± 3.01
Post-training	8	19	13	13.3 ± 2.48

Table 3 shows the results from the analysis of each SCI item with the mirt package of Chalmers (2012). The dataset is organized by the probability value P(θ) in descending order. The last column shows the probability of getting the correct answer to the SCI questions. According to the IRT-3PL model described before, Q01, Q06, and Q19 have high probability values because all participants responded correctly. On the other hand, questions Q12, Q07, and Q16 had the lowest probability values.

Table 3. IRT Model with three parameters: guessing, difficult, and discriminant, sorted by the probability **P(θ)**.

Position	Item	Guessing	Difficult	Discrim.	P(X_{ji}=1)	Position	Item	Guessing	Difficult	Discrim.	P(X_{ji}=1)
1	Q01	1.17E-09	-1.819	39.484	1.000	11	Q04	8.33E-08	-0.936	0.905	0.700
2	Q06	7.28E-01	0.437	-88.973	1.000	12	Q14	6.34E-01	0.689	115.177	0.634
3	Q19	8.06E-01	0.663	-45.433	1.000	13	Q17	1.67E-16	-0.315	1.742	0.634
4	Q15	597E-01	-0.038	41.814	0.931	14	Q10	9.96E-17	-0.251	0.993	0.562
5	Q05	1.51E-04	4.549	-0.572	0.931	15	Q03	5.03E-01	0.868	58.400	0.503
6	Q18	5.19E-01	26.157	-0.042	0.880	16	Q08	3.74E-29	0.129	1.236	0.460
7	Q13	8.03E-01	1.437	0.719	0.855	17	Q20	3.36E-01	0.343	131.834	0.336
8	Q11	8.51E-01	1.037	68.393	0.851	18	Q12	8.16E-02	1.391	42.088	0.082
9	Q09	8.01E-01	1.019	156.704	0.801	19	Q07	2.75E-02	1.383	74.179	0.028
10	Q02	1.14E-02	-4.114	0.308	0.782	20	Q16	2.87E-20	1.984	2.338	0.010

Analyzing Table 3 data and the probability associated with each item, Q01 (first position) is the item that requires a lower ability to choose the correct answer. On the other hand, Q16 (last position) is the question that requires a higher ability to choose the correct answer. As defined before, the probability P(θ) is an equation that considers three characteristics, guessing, difficult, and discriminant. All values commented are underlined in Table 3.

Figure 3 shows the ability θ and logistic curve probability P(θ) of each SCI item based on the IRT-3PL model.

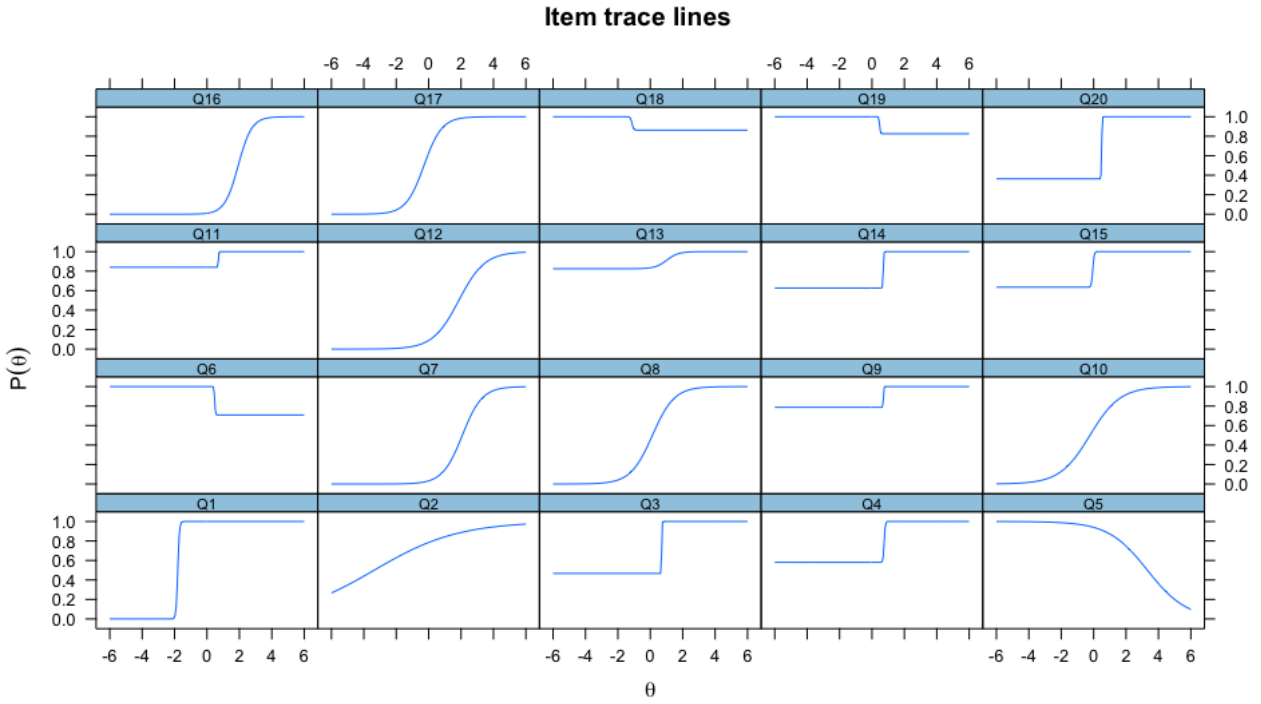

Figure 3. IRT Model with three parameters' curves between ability and probability P.

Concerning the SCI reliable instrument's internal consistency, considering the post-training only, the Cronbach's alpha was evaluated with a 0.545, which means the items are poorly correlated on the test, or there are not enough questions on the test (Taherdoost, 2016). The SCI Cronbach's alpha value should be above 0.7 to be acceptable, but this frequently occurs in initial applications of tests.

Table 4 shows the performance scores for each group separately, considering the pre-test and post-test moments and the number of participants.

Table 4. University performance scores.

Group	**# Participants**	**Pre-test Score** Mean ± Standard Deviation	**Post-test Score** Mean ± Standard Deviation	**Raw Gain**	**Effect-size**
A	14	9.71 ± 2.58	14.07 ± 2.7	4.36	1.65
B	9	9.11 ± 2.93	12 ± 2.5	2.89	1.06
C	19	10.47 ± 3.53	13.11 ± 2.47	2.64	0.87
D	8	9.75 ± 2.71	13.88 ± 1.73	4.13	1.81

The pre-test and post-test scores represent the participants' pre-training and post-training scores' means and standard deviations. The Raw Gain column in Table 4 is the difference between pre-test and post-test mean scores. The Effect-size column is related to a quantitative measure of the experimental effect's magnitude, which means the more significant the effect sizes, the stronger the relationship between two variables.

The effect-size *d* is described by Fritz et al. (2012) as:

$$d_{effectsize} = \frac{raw\_gain}{\sqrt{\frac{\sigma_{pre}^2 + \sigma_{post}^2}{2}}}$$

The goal of the effect size is to provide a measure *d* of the size of the effect from the pre-training and post-training moments. Therefore, the *d* measure determines the efficacy of an educational practice relative to a comparison group. According to McGrath et al. (2015), *d* values were more significant than 0.8, which means more than 79% of participants in the post-training test had learning gains comparing the pre-training test.

5 Discussion

The concept inventory is a helpful tool to measure learning between two moments in a training situation, but the design should discriminate valid questions to participants. The IRT was used in the questions considering three latent characteristics and the participant's scores in training. According to participants' responses to assess the SCI, we were able to verify the quality of each question and a proficiency model of training participants (Figure 4).

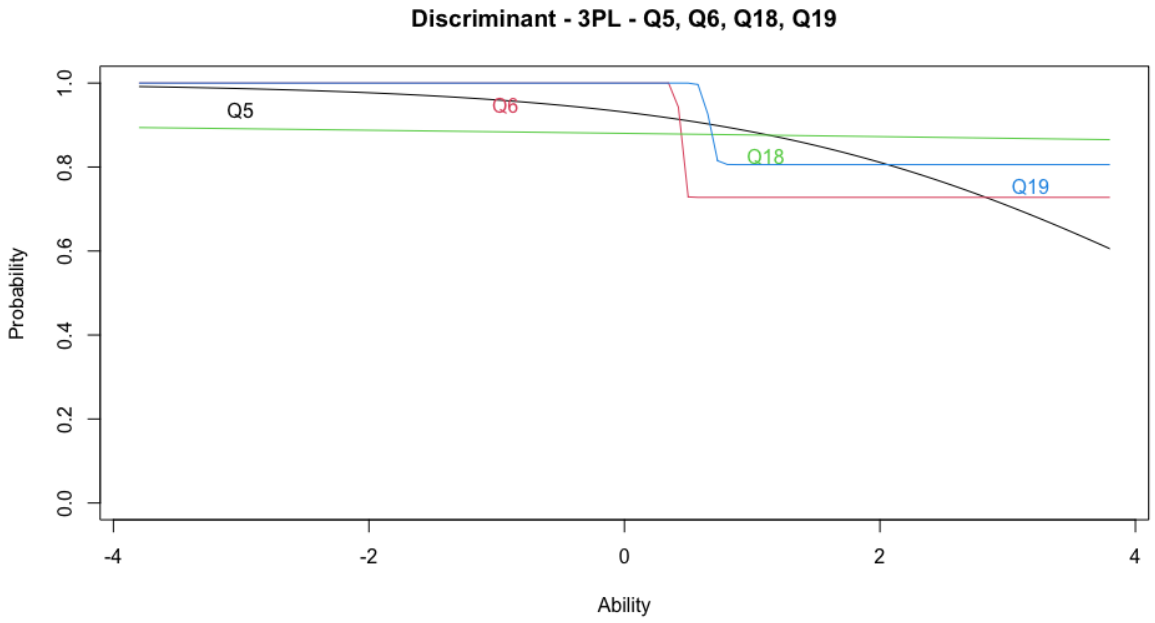

Figure 4. Questions Q5, Q6, Q18, and Q19 curves.

Considering the interpretation of the curves of items Q5, Q6, Q18, and Q19 in details shown in Figure 4, that had negative values, the IRT model indicated textual or misunderstanding problems in the four questions' descriptions. These four items should be discarded or rewritten because their discriminants have negative values, according to Table 3, and the curve's behavior decreases with higher participants' abilities, which is not desirable.

Items Q3, Q4, Q9, Q11, Q13, Q14, Q15, and Q20, have high guessing values and straight-up stair curves. The discriminant's higher values caused the straight-up stair curves. Guessing characteristics with high values indicates that learners with low ability probably choose the correct answer in these questions.

An IRT model's perfect curve occurs when the P(θ) equals 0.5 (representing 50%) for an ability parameter of zero. A curve like the IRT ideal curve is the one for item Q8. Item Q8 has a lower guessing value near to zero that represents the guessing chance to learners with low ability to choose the item's correct answer. The Q8 item difficult is almost zero value that represents the ideal item difficult to the IRT model. The Q8 discrimination value item is related to the curve's slope that is a positive value that causes the tilt direction to be upwards.

Otherwise, a negative discriminant value causes the direction of the slope to be down. According to the Q8 item three values, the P(θ) is equal to 0.46, representing a 46% probability value.

Concerning effect size d, all students' groups in Table 4 had values greater than 0.8, which means more than 79% of participants in the post-training test had learning gains comparing the pre-training test.

6 Conclusion

This study used the Scrum Guide to design the SCI questions, a reference source to Scrum worldwide. We considered that all concepts described in Scrum Guide are essential in the Scrum training.

Madsen et al. (2017) and Lindell et al. (2007) described some methods to assess learning using concept inventories and how they were designed for each topic or area. In Goldman et al. (2008), the Delphi process was used to identify important and difficult concepts about some disciplines in Computer Science, permitting a collection of information and reach consensus within a group of experts. The experts share observations in a structured way, preventing a few panelists from having excessive influence. The experts remain anonymous during the process so that they are influenced by the logic of the arguments rather than other experts' reputations.

As an education questionnaire, SCI must have good values of reliability and validity. As described previously in the Cronbach alpha value, the SCI had low reliability, but this is the first application of concept inventory. According to Kimberlin & Winterstein (2008), validity requires that an instrument is reliable, but an instrument can be reliable without being valid. One type of validity to SCI is to validate its questions content with experts. However, the other strategy chosen by authors was to validate the SCI content with the participants directly using the item response theory because latent trait models have provided an alternative framework for understanding measurement and alternative strategies for judging the quality of a measuring instrument (Kimberlin & Winterstein, 2008).

The next step of this study should be to review the SCI questions and validate their content with a panel of experts. Moreover, it will be necessary to develop new applications of the SCI with other participants to measure its quality using the ITR and the learning gains between pre-training and post-training situations.

Acknowledgments

We would like to acknowledge the support of the Brazilian government, the staff of the Federal University of Itajubá - Itabira Campus, and the staff of the University of Minho. We also thank the collaboration of professors and students from Brazilian and Portuguese courses for participating in the Scrum training held at their institutions through virtual meetings at Zoom. This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope UIDCEC003192019.

7 References

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the *R* Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18. https://doi.org/10.1037/a0024338

Goldman, K., Gross, P., Heeren, C., Herman, G., Kaczmarczyk, L., Loui, M. C., & Zilles, C. (2008). Identifying important and difficult concepts in introductory computing courses using a delphi process: Selective compression of unicode arrays in java. *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education - SIGCSE '08*, 256. https://doi.org/10.1145/1352135.1352226

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158. https://doi.org/10.1119/1.2343497

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364

Krivitsky, A. (2019). *Lego4Scrum 3.0: A complete guide to #lego4scrum—A great way to teach the Scrum framework and Agile thinking*. LeanPub Publishing. https://leanpub.com/lego4scrum

Lindell, R. S., Peak, E., & Foster, T. M. (2007). Are They All Created Equal? A Comparison of Different Concept Inventory Development Methodologies. *AIP Conference Proceedings*, *883*, 14–17. https://doi.org/10.1063/1.2508680

Madsen, A., McKagan, S. B., & Sayre, E. C. (2017). Best Practices for Administering Concept Inventories. *The Physics Teacher*, *55*(9), 530–536. https://doi.org/10.1119/1.5011826

McGrath, C., Guerin, B., Harte, E., Frearson, M., & Manville, C. (2015). *Learning gain in higher education*. RAND Corporation. https://doi.org/10.7249/RR996

Mesquita, D., Lima, R. M., Flores, M. A., Marinho-Araujo, C., & Rabelo, M. (2015). Industrial Engineering and Management Curriculum Profile: Developing a Framework of Competences. *International Journal of Industrial Engineering and Management (IJIEM)*, *6*(3), 121–131.

Project Management Institute (Ed.). (2013). *A guide to the project management body of knowledge (PMBOK guide)* (Fifth edition). Project Management Institute, Inc.

Project Management Institute & Agile Alliance. (2018). *Guia Ágil*. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1814552

Rabelo, M. (2013). *Avaliação educacional: Fundamentos, metodologia e aplicações no contexto brasileiro.* (1st ed.). Sociedade Brasileira de Matemática (SBM).

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5). https://doi.org/10.18637/jss.v017.i05

Sands, D., Parker, M., Hedgeland, H., Jordan, S., & Galloway, R. (2018). Using concept inventories to measure understanding. *Higher Education Pedagogies*, *3*(1), 173–182. https://doi.org/10.1080/23752696.2018.1433546

Schwaber, K., & Sutherland, J. (2017). *The Scrum Guide*. https://www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf. https://www.scrumguides.org/scrum-guide.html

Sliger, Mi. (2011). *Agile project management with Scrum*. PMI® Global Congress 2011, North America, Dallas, TX. https://www.pmi.org/learning/library/agile-project-management-scrum-6269

Taherdoost, H. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3205040