Perceptive Agile Measurement: New Instruments for Quantitative Studies in the Pursuit of the Social-Psychological Effect of Agile Practices

Chaehan So and Wolfgang Scholl

Department of Organizational and Social Psychology
Institute of Psychology, Humboldt University
Rudower Chaussee 18, 12489 Berlin, Germany
chaehan.so@gmail.com, schollwo@cms.hu-berlin.de
http://tinyurl.com/agilestudy

Summary. Rising interest on social-psychological effects of agile practices necessitate the development of appropriate measurement instruments for future quantitative studies. This study has constructed such instruments for eight agile practices, namely iteration planning, iterative development, continuous integration and testing, stand-up meetings, customer access, customer acceptance tests, retrospectives and co-location.

The methodological approach followed the scale construction process elaborated in psychological research. We applied both qualitative methods for item generation, and quantitative methods for the analysis of reliability and factor structure (principal factor analysis) to evaluate critical psychometric dimensions.

Results in both qualitative and quantitative analyses indicated high psychometric quality of all newly constructed scales. The resulting measurement instruments are available in questionnaire form and ready to be used in future scientific research for quantitative analyses of social-psychological effects of agile practices.

Keywords: agile practices, measurement instruments, iteration planning, iterative approach, continuous integration, test-driven development, stand-up meetings, co-location, retrospectives, customer acceptance tests, customer access.

1 Introduction

Research has investigated agile methods since the late 1990s, while the underlying roots date back to the 1980s [1]. Growing interest in psychological aspects within the agile software domain has since revealed that there is more to agile methods than the technical and process issues, namely a vast area of human and

P. Abrahamsson, M. Marchesi, and F. Maurer (Eds.): XP 2009, LNBIP 31, pp. 83–93, 2009.

social aspects. A number of studies have emerged investigating various social aspects in the complex relationship between agile practices and team interaction [2,3].

All these studies have applied qualitative research methods and found positive results of specific agile practices. Future research must therefore corroborate these qualitative findings using quantitative methods. In addition, the highly influential non-scientific consultant and practitioner literature on agile methods is abundant with claims about positive effects on social phenomena. The question remains whether these claims can be substantiated in quantitative studies. In pursuit of answering such research questions through quantitative analyses, this study aimed to develop appropriate measurement instruments.

2 Theoretical Framework

For quantitative studies on social-psychological effects, it is crucial that the measurement of agile practices is not effected by technical tools, but from the angle of *perception*. In psychology, perception is defined as the outcome of human information processing involving encoding, storage, retention, information retrieval, and judgment. The theoretical underpinnings of this approach derive from the fundamental psychological mechanism that individuals' behavior and choices are more influenced by their perceptions of situations and contexts than by the objective factual situation [4]. In many cases of real-life work contexts, discrepancies can occur between the technical reality (e.g. implementation quality of agile practices) and the corresponding individuals' perceptions. These discrepancies are caused by the fact that perceptions are prone to distortions through various cognitive and emotional bias effects, e.g. *selective attention* or *confirmation bias* (i.e. the tendency to ignore facts contradictory to prior judgement) [5].

Our selection of agile practices is by no means an attempt to present an exhaustive coverage of agile methodology on the whole, but to establish a representative set of agile practices commonly used in the field. Beyond coverage of agile values and principles as delineated in the *Agile Manifesto* [6], this set was intended to provide a solid basis for the process of scale construction and validation of agile practices as described in the methods section.

Consequently, we chose to pursue the following agile practices (their core aspects denoted in parentheses): *Iteration planning* (participation of all team members), *iterative development* (short iterations, time-boxing, working software), *continuous integration & testing* (continuous integration, test-driven development), *stand-up meetings* (short, regular, focused), *customer acceptance tests* (frequent, requirements verification by the customer), *customer access* (ease of contact to the customer, useful feedback), *retrospectives* (identification and implementation of improvement points), *co-location* (degree of physical proximity).

3 Method

The framework to develop measurement instruments for agile practices in this study is derived from the methodology of *scale construction* that has been developed and refined by psychological research for the last 100 years.

Psychological scale construction follows several qualitative methods for item generation and quantitative methods for the validation of established scales [7]. If the proposed new instruments are going to be used for future meaningful analyses in contexts of social interaction, they must undergo the scrutiny of reliability analysis and validation. In this study, we chose *internal consistency* as the main measure for reliability and analyzed *convergent validity* and *discriminant validity* (often summarized as *construct validity*) by explorative factor analyses.

3.1 Sample

In this study, team members (N = 227) of 55 software development teams applying agile practices across industries were tested. The sample consisted of project and product release teams, spread over 15 countries in America (Brazil, USA), Europe (Austria, Belgium, Finland, France, Germany, Italy, UK, Switzerland) and Asia (Bangladesh, India, Korea, New Zealand, Russia). Participants were male (n = 204) and female (n = 23). Work experience in agile software development projects was majorly distributed in the small range (62.0% with up to 1 year), considerable less in the middle range (21.9% with 1-2 years) and high range (16.1% with 3 years or more).

Acquisition of the participants followed five sources: Personal contacts of the author from eight years of professional experience in the software domain (mainly agile methods), contacts through researchers in the agile and organizational psychology domain, online newsgroups for agile methods, acquisition through an online business network (Xing), and an article published in the IT journal OBJEKTspektrum [8] which holds the highest subscription rate across German-speaking countries.

3.2 Procedure

The field study was conducted through a web-based questionnaire. Participants were assured of total anonymity by a formal data privacy policy signed by the research institute.

The agile practices' part of the main questionnaire was developed in three subsequent phases, followed by the final study:

Item Generation. The main qualitative measures we applied during the item generation and validation processes consisted of several structured and unstructured interviews with experts in the field through email, phone and face-to-face conversations. The starting point of the questionnaire development was the work of William Krebs who created the *Shodan Adherence Survey* [9] which was further elaborated by Layman et. al [10]. The scales of this survey were composed

of multiple items; these items were not evaluated individually but only by a single-item for the whole scale. Hence, the corresponding psychometric quality indicators were not available. Consequently, we decided to use the shodan survey items solely as a start discussion basis for the solicited experts to generate a new item pool basis. We applied qualitative procedures of item validation involving that new items were added and existing items were discarded, modified, extended or shortened, based on the expert feedback. Finally, three of the 84 items[1] of the Shodan Adherence Survey were retained in our final questionnaire version (see appendix) comprising 48 items.

Pretest 1. The first prelimenary study was conducted 19-26 June 2008 with 37 participants from 13 teams. Participants were mainly recruited from the network of the author's personal contacts. Feedback on the survey items, composed of questions, misunderstandings and clarifications, were discussed and the result integrated into the subsequent revision.

Pretest 2. The second prelimenary study was conducted 15-24 October 2008 with 44 participants. Participants were recruited through announcement in six major online newsgroups for agile software development.

Final Study. The final study was conducted between 26 October 2008 and 31 January 2009. From the pool of 260 formally invited persons, a total of 87.3% (N=227) in 55 software development teams completed the main questionnaire containing 43 items about agile practices. The co-location scale (5 items) was answered in an additional short questionnaire on project data by the technical project managers or scrum masters.

The data collection started with 107 individuals in 21 teams. After the first week, more than 700 practioners of agile methods, enlisted in the online business network Xing (www.xing.com), were contacted with information about this research study and a request to participate. From this channel, a total of 120 additional participants could be invited, and another 33 participants from further inquiries within the personal business network.

3.3 Optimizing Variance Explained

In pretest 1, a 5-point Likert scale format ranging from 1 (not at all) to 5 (totally) was used. To increase *variance explained* in the final study according to Lozano [11], the number of response categories was extended by two additional points to a 7-point Likert scale. In consideration of participant feedback received from pretest 1, the response format was transformed from an extent to a frequency scale ranging from 1 (never) to 7 (always), with the exception of the co-location scale which maintained the 5-point ordinal scale ranging from -2 (different time zones) to +2 (same room).

[1] Item cit8 retained in original version and items acctest2, standup2 in adapted versions.

3.4 Content Validity

All items generated from the initial version underwent thorough review by six experts in agile methodology and by active practitioners in software engineering throughout the first three research stages item generation, pretest 1 and pretest 2. The experts were asked for each agile practice scale to comment on a number of aspects including '*do the question correctly reflect the main characteristics of this agile practice?*', '*would you delete any question?*', '*how would you modify a question to improve its validity?*', '*would you add any question?*'. In addition, we solicited open feedback by means of a comment functionality in the used survey tool. This feedback was collected and systematically compared with prior feedback, merged and integrated into the next questionnaire version. This whole process was repeated several times during each stage until feedback from all involved experts converged to satisfaction with content validity.

3.5 Reliability

A scale is regarded to be reliable if there is little variance that is specific to particular items [12]. The most wide-spread measure used in psychological research for reliability is *internal consistency* which is equivalent to the average of all possible combinations of *split-half reliability* (i.e. splitting the scale by every possible combination of items according to the Spearman-Brown formula). The most commonly used statistic for internal consistency is *Cronbach* α. Acceptable values of α coefficients are generally regarded in the range of 0.75 and above; yet it is important to emphasize that the factor structure must be considered as well because the differential diagnostic value decreases the more dimensions the scale incorporates.

3.6 Factor Structure

When testing several psychological constructs simultaneously, a psychological test must ensure that these constructs can be measured as distinctly separate from each other. The according standard validation method during scale construction is *explorative factor analysis*. The goal of this method is to test for a clear *factor structure* which is defined by *convergent validity* (items for one construct load on the same factor and with high factor loadings determined by the marking factor) and *discriminant validity* (items for different constructs load on different factors and without double loadings).

In the field, agile practices are mostly applied at the same time; we thus expected a certain correlation of the respective scales and accounted for it by applying the oblique rotation technique *direct quartimin*[2] which allows maximum possible correlation in the solution as described by Gorsuch [13].

For the extraction technique, we chose to apply *principal factor analysis* (PFA) in favor of *principal components analysis* (PCA) because PFA, as argued by Tabachnick and Fidell [14, p.633–636], is designed for studies which hypothesize specific underlying constructs behind the empirical data.

[2] Direct quartimin corresponds to direct oblimin with a gamma value of zero.

In order to verify whether the right number of factors (8) was extracted, we used several criteria: First, the *scree* criterion indicated extraction until the 7th factor; yet we extracted an 8th factor because of a corresponding eigenvalue of 1.45 which clearly fulfilled the *Kaiser* criterion[3]. The most important of all applied criteria was that factor extraction should essentially be guided by *interpretability* of factors. In our case, all factors extracted corresponded precisely to agile practices modeled in our data. One aggregate scale (continuous integration & testing) spread its two subscales (continuous integration and test-driven development) over two different factors and thus revealed to be a two-dimensional construct. The aggregation of these two subdimensions was justified by the high reliability of the aggregated scale (0.88).

4 Results

The following section presents the results of the statistical analysis methods applied, namely reliability analysis and explorative factor analysis employing principal factor extraction and direct oblimin rotation.

4.1 Reliability Analysis

First, we analyzed internal consistency coefficients on the individual level. In order to improve our level of confidence, we tested internal consistency additionally on the group level by using aggregated individual item scores for the calculation of alpha coefficients[4].

Results (table 1) show consistently high Cronbach α coefficients for all scales (ranging between 0.78 and 0.93 on the individual level). These high reliability values could be replicated on the group level, showing even slightly higher values, ranging between 0.82 and 0.95.

4.2 Principal Factor Analysis

The analysis of the factor structure following the principal factor analysis approach showed a distinctly clear factor structure for all scales[5] (table 2).

The factor correlation structure revealed relatively low correlations among components (24 of 28 correlations below .31, 4 correlations in the range between .33 and .39). The highest correlation (.39) was between the scales retrospectives and customer acceptance tests which both share occurrence after iteration end.

[3] The Kaiser criterion specifies to extract all factors with eigenvalues above 1.

[4] Exception: The co-location scale was exclusively answered by project managers resp. scrum masters, hence no aggregated group value could be calculated.

[5] Co-location scale excluded from principal factor analysis because this scale was evaluated solely by project managers resp. scrum masters.

Table 1. Reliability Analysis on Individual and Group Level

Scale	# items	Cronbach α Indiv. Level	Cronbach α Group Level
Iteration Planning	7	0,79	0,85
Iterative Development	7	0,79	0,83
Cont. Integr. & Testing	9	0,88	0,93
Co-Location	5	0,78	n/a
Stand-up Meetings	5	0,79	0,82
Customer Access	4	0,93	0,93
Customer Acceptance Tests	5	0,87	0,91
Retrospectives	6	0,91	0,95

Table 2. Pattern Matrix of Principal Factor Analysis

	Factor							
	1	2	3	4	5	6	7	8
retrosp2	.87							
retrosp3	.86							
retrosp1	.77							
retrosp4	.75							
retrosp5	.68							
freqretr	.56							
cit5		.88						
cit9		.85						
cit6		.84						
cit8		.81						
cit7		.64						
cit4		.58						
access3			.92					
access1			.84					
access2			.84					
access4			.80					
acctest4				.84				
acctest1				.83				
acctest2				.77				
acctest3				.72				
freqcat				.40				
plan2					.65			
plan3					.65			
plan1					.59			
plan6					.47			
plan4					.46			
plan5					.37			
plan7								
standup2						.74		
standup1						.69		
standup3						.67		
standup5						.59		
standup4						.55		
cit3							.84	
cit2							.57	
cit1							.39	
iterat5								.64
iterat7								.54
iterat1								.52
iterat2								.44
iterat6								.40
iterat4								.37
iterat3								.36

Note: Factor loadings below 0.3 (<9% of variance explained) are omitted in this table

5 Discussion

The reliability analysis has shown that all scales possess distinctly high internal consistency to be categorized as highly reliable according to most psychological methodologists [12]. Moreover, since Cronbach α coefficients are consistently in the range of 0.8 and above, the scales can also be used for *causal analysis*. This is true not only for the analysis of causal structures with single or multiple regression, but also for the analytically more sophisticated *structural equation models*. Furthermore, we can assume that the high sample size of over 200 individuals yields relatively stable correlation parameters through a subject to item ratio of above 5:1, considerably higher than the recommendation of 2:1 by Kline [15].

The factor structure using principal factor analysis showed a clear discrimination between all scales and subscales spread over eight factors, also referred to as *simple structure*. The revealed simple structure can be regarded as a result of subsequent scale modifications based on the factor analyses after the preliminary studies. For example, items of the scale iteration planning were modified to reach unidimensionality of the scale after encountering high multicollinearity in the first pretest.

Future research should test the constructed scales of this study in other samples since results of reliability and factor analyses are always somewhat prone to sample dependency. These studies should apply *confirmative factor analysis* in order to validate the factor structure. The challenge to be taken consists of the requirements of high sample size for the mathematical method applied (structural equation modeling), and the main model fit criterion (Chi^2) which must be minimized for optimal model fit but has a positive linear dependency on sample size. The solution path of choice for this dilemma consists in reducing the number of estimation parameters to obtain more stable estimations. Item parceling techniques appear to be promising approaches in this direction.

In light of the aforementioned considerations, the decisive question emerges: How *replicable* are the results of this study? To answer this question, high reliablity and a clear factor structure are good indicators of replicability, but we must also look at *generalizability* – in this latter aspect, it seems safe to say that this study fulfills high requirements due to high sample size and because the sample was recruited across 15 countries, and furthermore varies from small companies to international corporations and across a wide spectrum of industries. This variability clearly distinguishes this study from previous research which mainly focused analysis on a single team or on several teams within the same company; we can thus expect a comparatively higher generalizability of this study's outcome. Yet it will be indispensable to test the new scales in future quantitative studies with confirmative factor analysis to reach an optimum level of confidence.

References

1. Abrahamsson, P., Warsta, J., Siponen, M.T., Ronkainen, J.: New directions on agile methods: a comparative analysis. In: ICSE 2003: Proceedings of the 25th International Conference on Software Engineering, Washington, DC, USA, pp. 244–254. IEEE Computer Society, Los Alamitos (2003)
2. Robinson, H., Sharp, H.: The social side of technical practices. In: Baumeister, H., Marchesi, M., Holcombe, W.M.L., Holcombe, M. (eds.) XP 2005. LNCS, vol. 3556, pp. 100–108. Springer, Heidelberg (2005)
3. Whitworth, E., Biddle, R.: The social nature of agile teams. In: Proceedings AGILE 2007, Washington, DC, USA, pp. 26–36. IEEE Computer Society, Los Alamitos (2007)
4. Thomas, J., Clark, S., Gioia, D.: Strategic sensemaking and organizational performance: Linkages among scanning, interpretation, action, and outcomes. Academy of Management Journal 36(2), 239–270 (1993)
5. Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology 2(2), 175–220 (1998)
6. Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R., Mellor, S., Schwaber, K., Sutherland, J., Thomas, D.: Manifesto for agile software development (Online) (August 2001)
7. Giles, D.C.: Advanced Research Methods in Psychology, Routledge, East Sussex, UK, New York, USA (2002)
8. So, C.: Teamwork in agilen Softwareteams (Teamwork in agile software development teams). OBJEKTspektrum Schwerpunkt: Kosten und Nutzen von Vorgehensmodellen (OBJEKTspektrum Focus: Cost and Benefit of Software Development Processes) (1), 10 (2009)
9. Williams, L., Layman, L., Krebs, W.: Extreme programming evaluation framework for object-oriented languages version 1.4. Technical Report TR-2004-18, North Carolina State University, Department of Computer Science, Raleigh, NC (June 2004)
10. Layman, L., Williams, L., Cunningham, L.: Motivations and measurements in an agile case study. Journal of Systems Architecture: the EUROMICRO Journal 52(11), 654–667 (2006)
11. Lozano, L.M., García-Cueto, E., Muñiz, J.: Effect of the number of response categories on the reliability and validity of rating scales. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 4(2), 73–79 (2008)
12. Cortina, J.M.: What is coefficient alpha? an examination of theory and applications. Journal of Applied Psychology 78(1), 98–104 (1993)
13. Gorsuch, R.L.: Factor Analysis. Lawrence Erlbaum Associates, Hillsdale (1983)
14. Tabachnick, B.G., Fidell, L.S.: Using Multivariate Statistics, 5th edn. Allyn & Bacon, Needham Heights (2006)
15. Kline, P.: An Easy Guide to Factor Analysis. Routledge, London (1993)

Appendix:
Perceptive Agile Measurement (PAM) Scales

Item Name	Item Wording	Corrected Item-Total Correlation	Cronbach α if Item Deleted
Scale: Iteration Planning			
plan1	All members of the technical team actively participated during iteration planning meetings.	0.58	0.75
plan2	All technical team members took part in defining the effort estimates for requirements of the current iteration.	0.62	0.75
plan3	When effort estimates differed, the technical team members discussed their underlying assumption.	0.59	0.75
plan4	All concerns from team members about reaching the iteration goals were considered.	0.57	0.76
plan5	The effort estimates for the iteration scope items were modified only by the technical team members.	0.41	0.78
plan6	Each developer signed up for tasks on a completely voluntary basis.	0.55	0.76
plan7	The customer picked the priority of the requirements in the iteration plan.	0.40	0.79
Scale: Iterative Development			
iterat1	We implemented our code in short iterations.	0.53	0.76
iterat2	The team rather reduced the scope than delayed the deadline.	0.47	0.77
iterat3	When the scope could not be implemented due to constraints, the team held active discussions on re-prioritization with the customer on what to finish within the iteration.	0.45	0.78
iterat4	We kept the iteration deadlines.	0.50	0.77
iterat5	At the end of an iteration, we delivered a potentially shippable product.	0.62	0.74
iterat6	The software delivered at iteration end always met quality requirements of production code.	0.54	0.76
iterat7	Working software was the primary measure for project progress.	0.54	0.76
Scale: Continuous Integration & Testing			
cit1	The team integrated continuously.	0.50	0.88
cit2	Developers had the most recent version of code available.	0.40	0.88
cit3	Code was checked in quickly to avoid code synchronization/integration hassles...	0.39	0.88
cit4	The implemented code was written to pass the test case.	0.61	0.87
cit5	New code was written with unit tests covering its main functionality.	0.79	0.85
cit6	Automated unit tests sufficiently covered all critical parts of the production code.	0.78	0.85
cit7	For detecting bugs, test reports from automated unit tests were systematically used to capture the bugs.	0.64	0.87
cit8	All unit tests were run and passed when a task was finished and before checking in and integrating.	0.73	0.86
cit9	There were enough unit tests and automated system tests to allow developers to safely change any code.	0.80	0.85

Item Name	Item Wording	Corrected Item-Total Correlation	Cronbach α if Item Deleted
Scale: Stand-Up Meetings			
standup1	Stand up meetings were extremely short (max. 15 minutes).	0.55	0.80
standup2	Stand up meetings were to the point, focusing only on what had been done and needed to be done on that day.	0.73	0.74
standup3	All relevant technical issues or organizational impediments came up in the stand up meetings.	0.64	0.77
standup4	Stand up meetings provided the quickest way to notify other team members about problems.	0.57	0.79
standup5	When people reported problems in the stand up meetigs, team members offered to help instantly.	0.58	0.79
Scale: Customer Access			
access1	The customer was reachable.	0.84	0.90
access2	The developers could contact the customer directly or through a customer contact person without any bureaucratical hurdles.	0.82	0.91
access3	The developers had responses from the customer in a timely manner.	0.88	0.89
access4	The feedback from the customer was clear and clarified his requirements or open issues to the developers.	0.80	0.92
Scale: Customer Acceptance Tests			
freqcat	How often did you apply customer acceptance tests?	0.48	0.88
acctest1	A requirement was not regarded as finished until its acceptance tests (with the customer) had passed.	0.75	0.82
acctest2	Customer acceptance tests were used as the ultimate way to verify system functionality and customer requirements.	0.77	0.82
acctest3	The customer provided a comprehensive set of test criteria for customer acceptance.	0.67	0.84
acctest4	The customer focused primarily on customer acceptance tests to determine what had been accomplished at the end of an iteration.	0.79	0.81
Scale: Retrospectives			
freqretr	How often did you apply retrospectives?	0.64	0.90
retrosp1	All team members actively participated in gathering lessons learned in the retrospectives.	0.75	0.89
retrosp2	The retrospectives helped us become aware of what we did well in the past iteration/s.	0.82	0.88
retrosp3	The retrospectives helped us become aware of what we should improve in the upcoming iteration/s.	0.84	0.88
retrosp4	In the retrospectives (or shortly afterwards), we systematically assigned all important points for improvement to responsible individuals.	0.72	0.89
retrosp5	Our team followed up intensively on the progress of each improvement point elaborated in a retrospective.	0.70	0.90
Scale: Co-Location			
coloc1	Developers were located majorly in ...	.52	.76
coloc2	All members of the technical team (including QA engineers, db admins) were located in ...	.67	.71
coloc3	Requirements engineers were located with developers in ...	.66	.71
coloc4	The project/release manager worked with the developers in ...	.50	.76
coloc5	The customer was located with the developers in ...	.49	.77