Measuring Fidelity to Extreme Programming: A Psychometric Approach

Abstract

This study assesses the Shodan survey as an instrument for measuring an individual's or a team's adherence to the extreme programming (XP) methodology. Specifically, we hypothesize that the adherence to the XP methodology is not a uni-dimensional construct as presented by the Shodan survey but a multidimensional one reflecting dimensions that are theoretically grounded in the XP literature. Using data from software engineers in the University of Sheffield's Software Engineering Observatory, two different models were thus tested and compared using confirmatory factor analysis: a uni-dimensional model and a four-dimensional model. We also present an exploratory analysis of how these four dimensions affect students' grades. The results indicate that the four-dimensional model fits the data better than the uni-dimensional one. Nevertheless, the analysis also uncovered flaws with the Shodan survey in terms of the reliability of the different dimensions. The exploratory analysis revealed that some of the XP dimensions had linear or curvilinear relationship with grades. Through validating the four-dimensional model of the Shodan survey this study highlights how psychometric techniques can be used to develop software engineering metrics of fidelity to agile or other software engineering methods.

Keywords: *psychometrics, confirmatory factor analysis, Extreme Programming (XP), fidelity, adherence, Shodan survey*

1. Introduction

When undertaking studies of software engineering methods it is important to assess the fidelity to which the teams follow the method. Often developers do not apply the method correctly or consistently over a long period of time and this may have detrimental effects on performance. On the other hand, excessively enforcing adoption can result in frustration among developers. These issues are more pronounced in development approaches that span the main software development lifecycle, often called methodologies. To aid our understanding of how teams adopt particular methods and whether the level of adoption affects the outcomes of the team we need to develop instruments for quantifying fidelity to software engineering methodologies. Such instruments would allow (a) organizations to be able to assess their own teams, and (b) researchers in empirical software engineering to measure and statistically control adoption of the methodology in consistent and comparative ways. Without these instruments we are unable to make hard conclusions about the relative merits of a method; poor performing teams could simply be those that did not follow it comprehensively, and equally the best performing teams could be the ones that constructively adapt the methodology to meet the circumstances or discard elements of it that they find do not work. This paper will focus on the development of an instrument for measuring the

adherence to one methodology, Extreme Programming (XP). More specifically, we build on the adherence metrics section of the XP evaluation framework of Williams et al. (2004a, 2004b), the most comprehensive method available. The basis of these metrics is the Shodan questionnaire (see Krebs 2002), and we will employ a psychometric approach to evaluate them and develop a four-dimensional rather than uni-dimensional measurement model.

1.1. Approaches to XP adherence

Many studies of XP seem to focus only on specific aspects of it that are directly related to the research question at hand. We found 27 papers that contained a study of XP. Of these, three did not discuss the way in which XP was applied other than by referencing the standard method (Mannaro et al. 2004; Noll and Atkinson 2003; Young et al. 2005). Three further papers used a qualitative method that briefly described the application (Jokela and Abrahamsson 2004; Merisalo-Rantanen et al. 2005; Moser et al. 2007), ten papers discussed some but not all of the practices used with little or no information on fidelity (Bahli and Zeid 2005; Koskela and Abrahamsson 2004; Mackenzie and Monk 2004; Martin et al. 2004a, 2004b; Müller and Tichy 2001; Newkirk and Martin 2000; Robinson and Sharp 2004, 2005b; Tessem 2003) and three had significant information about fidelity (Cao et al. 2004; Chong 2005; Fruhling et al. 2005; Robinson and Sharp 2005a; Sharp and Robinson 2004). In terms of quantitative studies, two papers used metrics to assess seven of the XP practices (Abrahamsson 2003; Abrahamsson and Koskela 2004).

Two papers assessed fidelity of the complete set of XP practices. Gittins and Hope (2001) used a qualitative approach to explore each of the practices of XP systematically. In their approach, the authors described how each of the practices had been adopted by the teams. Sfetsos et al. (2006) used a combination of quantitative and qualitative methods. They used a semi-structured interview and questionnaire to systematically collect data from several companies. The questionnaire has a single question for each practice that had a three point scale to indicate that it was partially, fully or not used. The interviews were used to enhance this data to explain why it was adopted. In both studies the results were presented in terms of how each of the practices had been adopted with no attempt to describe the overall fidelity to XP.

Finally, Williams et al. (2004a, 2004b) developed the XP evaluation framework. Some examples of the application of this framework include case studies by Layman et al. (2004) and Layman (2004). In this framework a questionnaire instrument, often referred to as the Shodan survey, was developed to form quantitative assessments of adherence to XP. This is the instrument that we will be examining in this paper.

1.2. The Shodan survey

The XP evaluation framework presented by Williams et al. (2004a, 2004b; see also Krebs 2002) contains a number of components: context factors, adherence metrics and outcome

measures. Of these, the adherence metrics are used to measure fidelity to XP, through objective measures, interview techniques and the Shodan survey. In these metrics, a set of weights is proposed to combine the scores from the questions to give a single value that corresponds to fidelity. These weights were obtained though analysis of the relatedness of the practices as presented by Beck and Andres (2004).

Questions from the Shodan survey are grouped in six categories: foundations, customer planning, teaming, craftsmanship, introspection and perspectives. For the purposes of this paper we only focus on the first four. The first category, foundations, concentrates on testing and pair programming. Customer planning concerns the planning game, customer access, short releases and stand-up meetings, whereas the teaming category addresses issues related to collective code ownership, coding standards and continuous integration. The fourth category, craftsmanship, is concerned with sustainable pace, simple design and the use of metaphor. Introspection is not discussed in this paper because it is based on the assumption that teams have worked together on more than one project, which the teams in our study had not. While it is often the case that the team composition changes from project to project, this is not an inherent characteristic of agile methodologies. The Shodan survey's final category, perspectives, assesses which practices participants felt were the most threatening or promising. Again, this is not central to an evaluation of XP.

There are a number of advantages to the Shodan survey and to using a questionnaire-based methodology in general to measure adherence to XP:

(a) It provides overall adherence scores instead of relying on individual practices.

(b) It is quantitative and thus can be used in both quantitative and qualitative studies (whilst qualitative approaches are not useful to quantitative studies).

(c) It is easily applicable in organizations.

(d) Results from different organizations or research contexts should be comparable.

Nonetheless, there are a number of issues that need to be addressed if this methodology is to be developed into a robust measurement instrument. First, while Williams et al. (2004a; 2004b) group their questionnaire items into theoretical categories, there is an underlying assumption in the way they apply the method that adherence to XP can be measured as one dimension. It is important to evaluate whether the uni-dimensional model is adequate for capturing adherence or whether the categories used in the survey questions could reflect different dimensions of adherence to XP. Second, the instrument has not been validated. Thus as it stands we do not know whether the instrument measures what its architects claim it measures. Finally, Williams et al. apply weights to the various survey items according to the importance of the various XP practices, but there is no evidence that these weights correctly capture the relative contribution of each to the overall approach. It is

possible that weighting and aggregating the results of the survey questions in this way creates an invalid measure or underutilizes important information.

Given it may well be, then, that a theoretically driven four-dimensional model could more accurately measure adherence to XP, we now present a test of this using psychometric methods. Having identified dimensions of adherence we then validate and assess whether individual items need to be weighted differently to the original Shodan model.

1.3. Alternative psychometric approach

Psychometrics refers to the study of psychological measurement and it is concerned with the development and analysis of psychological tests, questionnaires and related instruments, in order to develop valid and reliable measures. One of the most frequently used techniques for testing measurement models, which will be employed in this paper, is latent or factor analysis. Factor analysis techniques are typically used to evaluate whether or not a common factor or latent variable underlies a set of survey items or observed variables (Bartholomew and Knott 1999). Specifically, these techniques aim to describe variability in the observed variables in terms of a smaller set of latent factors. Thus it is possible to assess whether one dimension can reflect the variability in the Shodan items or whether a multidimensional measure would be more appropriate.

Besides the need for data reduction and the evaluation of competing measurement models, factor analysis techniques can provide evidence for construct validity. The underlying logic behind using factor analysis to assess construct validity is that the latent construct is the reason that individuals respond to a set of items in specific ways. If the measurement model is valid then adherence is an underlying property of the software engineering process, which will result in individuals responding to the survey items in specific patterns. Through those patterns we can evaluate whether or not our latent construct or constructs are valid. In the case of the Shodan survey there are two main measurement models for which the construct validity should be evaluated: the uni-dimensional instrument and the four-dimensional instrument based on the four core categories in the survey.

Furthermore, the loadings obtained from a factor analysis signify the contribution of each of the items to the corresponding dimension or dimensions. In effect the factor loadings or standardized coefficients for the items are weights for calculating the latent factors. When the aim of the analysis is to maximize the variance explained by the underlying factors, exploratory factor analysis is better suited for obtaining the factor loadings. However, since a prior theoretical model already exists, as derived from the categorization of the questionnaire items, it is likely to be more useful to test that model rather than create another classification or set of dimensions. In addition, through confirmatory factor analysis each of the items loads only to one factor, whilst in exploratory factor analysis each item has typically a high loading in one of the identified dimensions and weaker loadings in others. Thus, for the purposes of

identifying the appropriate loadings for measuring fidelity to XP, a confirmatory rather than an exploratory approach is preferable.

In our study we will first test the uni-dimensional and four-dimensional models using confirmatory factor analyses, then derive different weights for each of the questions from that analysis, and finally compare the two models. These three steps are framed in the following hypothesis:

H1: A four-dimensional measurement model will have better fit than a uni-dimensional model for measuring adherence to the XP methodology.

In addition to evaluating a measurement model using these techniques this study will also present an exploratory analysis of the four dimensions in terms of how they relate to students' performance as assessed through their grades. The obvious assumption of the utility of an instrument that measures adherence to XP is that teams following XP practices most closely will be better performers than those who adopt the same practices to a lesser extent (Stephens and Rosenberg 2003). For instance, case studies of software development showed that (a) XP teams deliver above average quality and productivity and (b) that adoption of XP results in improvements in quality and productivity when compared to performance of the same team prior to adoption (Layman et al. 2004, 2006). Previous results have also shown that student teams using XP deliver marginally higher quality results than those using a plan-driven approach (Macias 2004) and that there is a positive relationship between quality and the number of practices used (Syed-Abdullah 2005).

Taking this evidence from past research suggests that valid measures should be able to explain more variability in students' grades than less valid measures. This leads to the second hypothesis of this paper:

H2: The four-dimensional measurement model will explain more variability in students' grades than the uni-dimensional model.

2. Methods

The Shodan questionnaire was administered to software engineering students working in teams that were following the XP methodology (Thomson and Holcombe 2009).

2.1. Measures

The questionnaire administered consisted of 13 questions on the topics that are presented in Table 1 below.

[Insert Table 1 about here]

In the original Shodan survey there were two additional questions that were omitted here. The first was about the planning game and was part of the customer planning category. The second was about the use of a metaphor and was part of the craftsmanship category. Both questions required a binary answer rather than a continuous scale. There were a large number

of missing cases for these two variables and since they were considered to overlap with other questions, we excluded them from our analysis. An 11 point scale from 0–100% was used for the included questions.

For exploring the second hypothesis, the dependent variable was the grades awarded to each of the students at the end of the year for their software engineering project. These were independently awarded to students and thus overall team performance did not determine the individual grades. The grade was based on the average of two independent assessments by two different assessors using the same guidelines and criteria, and thus the results reflect the students' experience, knowledge and expertise in a consistent way. If the two assessments differed by more than five points out of 100, the assessors would discuss the discrepancy in order to arrive at a consensus. Although these grades are not necessarily the best way to assess individual performance in software engineering we would expect them to correlate highly with other software engineering metrics.

2.2. Sample

Participants consisted of first year, second year and Master's students, who had varying experience and expertise with the XP methodology and programming in general. The first year students were working on projects assigned by the University, while the second year and Master's students were working on real software engineering projects that were assigned from clients to be used in their respective business.

Sample variability in terms of students' experience is useful when testing measurement models because it should permit greater generalizability of the findings and ensures a good spread in the data. Indeed, for the purposes of this study having greater variability in terms of experience in applying XP is a virtue rather than a limitation of the sample.

The questionnaire was given to 289 students, and completed by 187. Of those, only 138 completed all of the questions, which formed the sample used for the confirmatory factor analysis. To assess whether the missing values were systematic an ANOVA test was performed on the 13 items of the 187 returned questionnaires (138 students who had no missing cases, and the remaining 49 who had some missing values) and on students' performance. There were no significant differences for any of the 13 items tested or the students' performance scores. The missing values were therefore not systematic and thus cases with missing values were removed from the analysis.

For the exploratory analysis of students' grades, the Shodan scores were estimated for the 187 students and the missing cases were excluded pair-wise. A total of eight students did not answer all the questions that comprised one of the four latent factors and they were removed from the analysis. Out of the remaining sample there were 29 cases with missing performance data reducing the sample to a total of 150 valid cases.

To examine whether the sample size would be adequate for the predictive validity tests we estimated the effect size that would be required for the various models in order to achieve a power of 0.8 and a significance level of 0.05. The tests were performed using Cohen's (1988) tests. For the uni-dimensional model the required effect size was f^2 =.053, which suggests that variance explained by the model should be more than 5% (R^2=0.05). For the four-dimensional model the effect size was estimated at f^2 =.082, thus requiring at least 7.6% (R^2=0.076) of the variance to be explained in order to achieve the required power. Hence, the sample can be considered adequate for both analyses.

2.3. Procedure

Questionnaires were administered to all first year, second year and Master's students at the Department of Computer Science at the University of Sheffield. This was done for two consecutive years for the first and second year students, and three consecutive years for the Master's students. All questionnaires were administered at the end of the students' projects to ensure that the students' responses reflected the adopted XP practices throughout the project. Students were advised that their participation in the study was voluntary and they could withdraw from the study at any point. The performance data were collated after the students were awarded their grades at the end of the academic year. To ensure anonymity, the data were collated using an ID number assigned to each participant.

3. Results

3.1. Confirmatory factor analysis

The uni-dimensional and four-dimensional models were tested through a confirmatory factor analysis approach using the Mplus v4.21 package. Maximum likelihood estimation was used to analyse the covariances for the 13 questionnaire items. The covariance matrix is shown in Table 2 below.

[Insert Table 2 about here]

For the uni-dimensional model, the unstandardized coefficient estimates were all significant except for two of the items (items 8 and 13), which were not statistically significant ($p<.05$). The model coefficients can be found in Table 3 below. The Cronbach's alpha reliability coefficient (Cronbach 1951; Nunnaly and Bernstein 1994) was satisfactory for the uni-dimension, with α=0.83. A comparison between the unstandardized coefficients (factor loadings) and the weights of the original survey (see Table 1) highlights that the original weights did not present the best solution in terms of the covariance of the items.

[Insert Table 3 about here]

The coefficients obtained indicate that the first three questions appear to dominate the analysis with the rest of the items having small coefficients. Nevertheless, the fact that the first three questions are primarily testing questions does not imply that testing was overemphasized in the courses from which we drew the sample. A strong relationship

between the testing questions should still manifest regardless of how high or low teams scored. Figure 1 below shows the uni-dimensional model and the standardized coefficients.

[Insert Figure 1 about here]

Overall fit of the models were assessed using the χ^2 test, the Comparative Fit Index (CFI), the Tucker–Lewis Fit index (TFI), the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square residual (SRMS) indices. The χ^2 test evaluates whether the predicted and observed covariance matrices are different. A good fit is inferred when the two matrices are not different; a significant difference implies a poor fit. However χ^2 tests should be interpreted with caution. The χ^2 estimates can increase when there is a large sample size or high correlations among the variables resulting in significant test even when there is a good model fit. Using the CFI and TFI measures, a good model is indicated by higher values. Above 0.90 is generally considered an indication of an acceptable model and above 0.95 indicates an excellent model. For RMSEA, a good fit is indicated by a small value. In general, models with less than 0.10 are considered acceptable, whilst models with less than 0.05 are considered to have a very good fit. The SRMS index is the standardized difference between the observed and predicted covariance. A zero value would indicate a perfect fit, whilst the higher the value the poorer the fit of the model. A value less than 0.08 is considered a good fit. In all tests the uni-dimensional model shows a weak fit. Table 4 below shows various fit statistics for the uni-dimensional and four-dimensional models.

[Insert Table 4 about here]

For the four-dimensional model all of the unstandardized coefficient estimates were statistically significant (see Table 5). Overall, the coefficients indicate a considerable improvement over those for the single-factor model. In terms of reliability, Cronbach's alphas for the hypothesized model were α=0.83 for foundation, α=0.48 for customer planning, α=0.64 for teaming, and α=0.41 for craftsmanship. Although the foundations and teaming factors achieved a high reliability coefficient, customer planning and the craftsmanship are quite low.

[Insert Table 5 about here]

The factor loadings (unstandardized coefficients) are again different from the original Shodan weights (see Table 1) but are also different from those obtained from the analysis of the uni-dimensional model. Nevertheless, the weights for the first five questions are high and similar in both the uni- and four-dimensional models. This suggests that the latent factor in the first model reflected more of the foundations of XP but failed to adequately capture the other three categories. Figure 2 shows the four-dimensional model with the standardized coefficients.

[Insert Figure 2 about here]

For assessing the overall goodness of fit for the model, the χ^2 test is significant, showing that the observed covariance matrices are significantly different from the predicted covariance matrices. This indicates a not so good fit for the model. As explained above, though, the χ^2 test can often be unreliable and the overall fit should be evaluated using a number of indices. Using CFI, TFI, RMSEA, and SRMS (see Table 4), the model shows a very good fit.

Finally, a comparison between the uni-dimensional and the four-dimensional models was performed using the chi-square differences between the two models. The chi-square differences between two nested models follows a chi-square distribution with degrees of freedom equal to the difference of the degrees of freedom of the two models. If the two models are significantly different, the model with the smaller chi-square is significantly better than the first. The comparison indicated that the four-dimensional model was significantly better than the uni-dimensional model ($\Delta\chi^2$ =73.2, Δdf =6, p<.001). Comparisons of all other fit indices corroborate this result. These results provide strong support for the first hypothesis (*H1*) indicating that the four-dimensional model can better explain the variability in the Shodan questionnaire items, whilst at the same time providing evidence about its construct validity.

3.2. Exploratory analysis of adherence and students' grades

Linear regression models were used to assess the exploratory analysis for the two models (*H2*). The models were tested with R 2.8.0 (R Development Core Team 2008). In all models the dependent variable was students' performance, as assessed by their lecturers on a 0–100 scale.

For the uni-dimensional model the independent variable was the Shodan adherence score, calculated using its factor loadings (unstandardized coefficients). This model failed to achieve any significant results – F(1,148)=0.001, p>.05 – and accounted for only 0.001% of the variance. For the four-dimensional model, the independent variables were the four latent variables identified. These were estimated by weighting each of the items using the unstandardized coefficients obtained from the measurement model above.

Contrary to the uni-dimensional model, the four-dimensional model was very significant – F(4,145)=11.87, p<.001 – and explained 24.7% of the variance. Out of the four predictors, foundations, teaming and craftsmanship were significant and only customer planning was not. However, the variable foundations had a negative rather than a positive effect on performance. Further analysis through scatterplots of the data revealed that there could be curvilinear effects. These were also tested using second-order polynomial linear regression. For reasons of consistency and fairness we tested for the quadratic effects in both the uni-dimensional and four-dimensional models. In order to reduce the possibility for

collinearity between linear and quadratic predictors, the latter were centred on their means before they were raised to the second power. The results from all models are presented in Table 6 below.

[Insert Table 6 about here]

For the uni-dimensional quadratic model the results indicate that there is a modest effect (R^2=.07), which is significant (F(2,147)=5.53, p<.01). Nevertheless, an ANOVA comparison between the linear and quadratic models reveals that their difference is statistically significant (ΔF=11.06, ΔR^2=.07, p<.05). The t-tests for the predictor indicate that only the quadratic term and not the linear was significant, which had a negative β coefficient. The negative coefficient here indicates an inverted U-shaped relationship with the dependent variable.

For the four-dimensional model the results were further improved after addition of quadratic predictors for foundations and craftsmanship. With the exception of customer planning, all linear and curvilinear terms added to the regression equation were significant. Overall the model explained 30.9% of the variance in performance and was statistically significant (F(6,143)=10.66, p<.001). The model also provided significant improvement over the linear model (F=6.46, ΔR^2=.062, p<.01). For the two predictors that had significant quadratic effects, foundations had a negative coefficient whilst craftsmanship's was positive, indicating an inverted U type of relationship for the former and a U-shaped relationship for the latter (see Figure 3).

[Insert Figure 3 about here]

A comparison between the quadratic uni-dimensional and the quadratic four-dimensional models indicated that the latter was significantly better (F=12.38, p<.001) explaining 23.9% more variance than the quadratic linear model. Overall, the results indicate that the four-dimensional measurement model can explain more variance in students' grades and thus provide strong evidence for the second hypothesis of this paper.

4. Discussion

This study examined the measurement model of the Shodan survey for quantifying the degree of adherence to the XP methodology. First, a uni-dimensional model does not explain the variability in the Shodan items and thus aggregating them in one variable does not constitute a valid measure of adherence to XP. Second, the four-dimensional model provides a far more accurate way of measuring adherence to the XP methodology. Third, the loadings used in the original Shodan survey are very different from those obtained empirically in our analysis.

Although the analysis showed better support for the four-dimensional model, the results should be approached with caution. The Cronbach's alpha scores were not that strong for three of the four factors, with the exception of the foundations factor. This indicates that

there is relatively low internal consistency in the items making up these three factors. Thus, using the Shodan survey at its current state could be problematic: on the one hand a four-factor solution is more appropriate, but on the other hand the items have weak internal consistency and thus do not adequately describe three of the four factors.

The results from the exploratory analysis showed that the four-dimensional measurement model could explain more variation in students' grades than the uni-dimensional model. It is interesting that quadratic effects exist for both the original Shodan as well as for the measurement model identified here. For the original uni-dimensional model, the results suggest that overall there is a small curvilinear effect of adherence to the XP methodology on grades. Specifically, it was found that there is a positive effect up to medium levels of adherence and a negative effect for individuals scoring higher on the adherence scale. The nature of the relationships in the four-dimensional model indicated two curvilinear effects and one linear effect. The linear effect was between teaming and grades whilst the curvilinear were between (a) foundations and (b) craftsmanship and grades. For the foundations the relationship was similar to the one identified for the original Shodan scale. In contrast, the effect of craftsmanship was in the opposite direction, indicating that applying craftsmanship to a small extent can have a detrimental effect and only when the practices are more closely adopted do they tend to have a positive effect on grades. Customer planning did not appear to have any effect on grades.

Taken together these results show that the XP methodology is in fact multidimensional and thus it should be treated as such both theoretically and empirically in terms of measurement. For researchers, the psychometric analysis of the Shodan survey provides a step forward in creating instruments that can accurately capture the degree to which XP has been adopted. This paves the way for potential research into explaining individual or team performance as well as product quality. Furthermore, we believe that psychometric techniques are invaluable for studies that are developing instruments for quantifying the degree of adherence of software engineering methods.

For practitioners using Shodan to assess the adoption of XP by individuals or teams in their organization, using a four-dimensional instrument provides greater flexibility and better results in terms of the adoption. Such metrics as these are invaluable for identifying potential problems in the development process as well as any effects that their adoption may have on the quality of the product.

4.1. Future research

There are a number of ways that this study could be taken forward. First, a bigger sample from industry would help to establish how general the results are. A sample including some teams that have been working together for a long period of time and on more than one project would also allow the inclusion of questions on introspection (the fifth possible

dimension) that were excluded here. Second, the XP evaluation framework (Williams et al. 2004a) relies on data collection from other techniques (e.g. interviews) and thus we cannot see the complete picture with the questions used here. Effective communication, for example, which is an important element of the XP methodology but not a practice, is not measured by the questionnaire. Future research could refine the questionnaire to include such elements. In addition, the issue of the small reliability coefficients can be addressed through the inclusion of more items per latent factor or a different conceptualization of the dimensions of adherence to XP. Ideally, future research should focus on a measurement model and questionnaire items specifically developed with the assumptions of measurement underlying factor analysis and related techniques.

Third, whether these measures truly operationalize the theorized dimensions should be given consideration. Following the different factor loadings it can be argued that there is incongruence between the labels of the hypothesized dimensions and the questionnaire items. Foundations, for instance, was reflected in the first three questions which refer to testing. Similarly, customer planning seems to have more to do with planning in general rather than the customer, and the items of the third factor, teaming, appear to relate more to team code management. Finally, craftsmanship was primarily reflected in the question about sustainable pace, and simple design had only a small loading. Future research should consider whether renaming these factors would give a better perspective of the factors representing adoption of XP practices or whether the original dimensions should be operationalized with different items that more accurately capture their essence.

Fourth, future research should expand to other methodologies by identifying possible dimensions that span methodologies, ideally with the aim to create a generic way of assessing fidelity that can be applied to all methods. This will enable more refinement in the comparisons of approaches, but also allows for the fact that practices that are not highlighted as important in some methodologies may still be used when following others. For instance, although teamwork and communication are core elements of agile methods, they are not unimportant in traditional waterfall methods. Equally, XP teams could adopt practices from other methods that aid their performance. Using the techniques applied here would make possible the development of a generic method using the best elements and practices from various different methodologies.

Finally, there is a need to evaluate the predictive validity of this instrument (or an updated version of it with more reliable items), particularly using proper software engineering performance metrics or expert evaluations of fidelity to XP. In addition, although we tend to think about adherence at the team level, we were only able to assess relationships between agile methods and individual grades. Future research should thus address this issue through collecting data from a bigger population. Subject to within-group reliability (James et al.

1984), analysis could be done by either aggregating individual scores to the team level or by using multilevel confirmatory factor analysis (Muthén 1994). From this perspective, data from diverse teams from different organizations would be invaluable in developing a model that would be applicable to XP teams operating within different organizational contexts.

4.2. Conclusions

It is apparent from the results of this study that the four-dimensional measurement model of the Shodan survey is a better conceptualization of adherence to XP. However, it is recommended that more research is needed both in process conformance as well as in terms of adherence questionnaires. The psychometric approach adopted here can potentially be used to devise more complex, accurate and interesting measures. Such measures can enable quantitative research with more statistical rather than experimental control over the adopted processes. The results also uncovered nonlinear relationships between adoption of XP and performance that have not been conceptualized or theorized before.

Acknowledgements

We will leave out for now, as may aid identification of authors or their affiliation.

References

Abrahamsson P (2003) Extreme programming: First results from a controlled case study. Euromicro Conference 2003. Proceedings. 29th: 259–266.

Abrahamsson P, Koskela J (2004) Extreme programming: A survey of empirical data from a controlled case study. Proceedings 2004 International Symposium on Empirical Software Engineering: 73–82.

Bahli B, Zeid ESA (2005) The role of knowledge creation in adopting extreme programming model: An empirical study. ITI 3rd International Conference on Information and Communications Technology 2005: Enabling Technologies for the New Knowledge Society.

Bartholomew DJ, Knott M (1999) Latent variable models and factor analysis. Oxford University Press, New York.

Beck K, Andres C (2004) Extreme programming explained: Embrace change (2nd Edn). Addison-Wesley Professional, Boston.

Cao L, Mohan K, Xu P, Ramesh B (2004) How extreme does extreme programming have to be? Adapting XP practices to large-scale projects. Proceedings of the 37th Annual Hawaii International Conference on System Sciences 2004.

Chong J (2005) Social behaviors on XP and non-XP teams: A comparative study. In Proc. Agile United Conference.

Cohen J (1988) Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates, New York.

Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16: 297–334.

Fruhling A, Tyser L, De Vreede GJ (2005) Experiences with extreme programming in telehealth: Developing and implementing a biosecurity health care application. HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on System Sciences.

Gittings R, Hope S (2001) A study of human solutions in extreme programming. Proc. 13th Workshop of the Psychology of Programming Group: 41–51.

James LR, Demaree RG, Wolf G (1984) Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology, 69: 85–98.

Jokela T, Abrahamsson P (2004) Usability assessment of an extreme programming project: Close co-operation with the customer does not equal to good usability. Product Focused Software Process Improvement: 397–407.

Koskela J, Abrahamsson P (2004) On-site customer in an XP project: Empirical results from a case study. Software Process Improvement, 3281: 1–11.

Krebs W (2002) Turning the knobs: A coaching pattern for XP through agile metrics. Presented at Extreme Programming/Agile Universe, Chicago IL: 60–69.

Layman L (2004) Empirical investigation of the impact of extreme programming practices on software projects. OOPSLA '04: Companion to the 19th annual ACM SIGPLAN conference on Object-oriented programming systems, languages, and applications. Vancouver, Canada, ACM.

Layman L, Williams L, Cunningham L (2004) Exploring extreme programming in context: an industrial case study. Agile Development Conference 2004.

Layman L, Williams L, Cunningham L (2006) Motivations and measurements in an agile case study. Journal of Systems Architecture 52(11): 654–667.

Macias F (2004) Empirical assessment of extreme programming. Unpublished PhD thesis, University Of Sheffield.

Mackenzie A, Monk S (2004) From cards to code: How extreme programming re-embodies programming as a collective practice. Computer Supported Cooperative Work (CSCW), 13: 91–117.

Mannaro K, Melis M, Marchesi M (2004) Empirical analysis on the satisfaction of it employees comparing XP practices with other software development methodologies. Extreme Programming and Agile Processes in Software Engineering, Lecture Notes in Computer Science, 3092/2004: 166-174, Springer Berlin / Heidelberg.

Martin A, Biddle R, Noble J (2004a) When XP met outsourcing. Extreme Programming and Agile Processes in Software Engineering, Lecture Notes in Computer Science, 3092/2004: 51-59, Springer Berlin / Heidelberg.

Martin A, Biddle R, Noble J (2004b) The XP customer role in practice: three studies. Agile Development Conference 2004.

Merisalo-Rantanen H, Tuure T, Matti R (2005) Is extreme programming just old wine in new bottles: A comparison of two cases. Journal of Database Management, 16: 41–61.

Moser R, Scotto M, Sillitti A, Succi G (2007) Does XP deliver quality and maintainable code? Agile Processes in Software Engineering and Extreme Programming, Lecture Notes in Computer Science, 4536/2007: 105-114, Springer Berlin / Heidelberg.

Muller M, Tichy W (2001) Case study: Extreme programming in a university environment. ICSE '01: Proceedings of the 23rd International Conference on Software Engineering. Toronto, Canada, IEEE Computer Society.

Muthén BO (1994) Multilevel covariance structure analysis. Sociological Methods & Research, 22: 376–398.

Newkirk J, Martin R (2000) Extreme programming in practice. OOPSLA '00: Addendum to the 2000 proceedings of the conference on Object-oriented programming, systems, languages, and applications (Addendum). Minneapolis, ACM.

Noll J, Atkinson D (2003) Comparing extreme programming to traditional development for student projects: A case study. Extreme Programming and Agile Processes in Software Engineering, Lecture Notes in Computer Science, 2675/2003: 1013, Springer Berlin / Heidelberg.

Nunnaly JC, Bernstein IH (1994) Psychometric theory (3rd edn). McGraw-Hill, New York.

R Development Core Team (2008) R: A language and environment for statistical computing. Series in Psychology. R Foundation for Statistical Computing, Vienna, Austria.

Robinson H, Sharp H (2004) The characteristics of XP teams. Extreme Programming and Agile Processes in Software Engineering, Lecture Notes in Computer Science, 3092/2004: 139-147, Springer Berlin / Heidelberg.

Robinson H, Sharp H (2005a) Organisational culture and XP: Three case studies. Proceedings Agile Conference 2005.

Robinson H, Sharp H (2005b) The social side of technical practices. Extreme Programming and Agile Processes in Software Engineering, Lecture Notes in Computer Science, 3556/2005: 100-108, Springer Berlin / Heidelberg.Sfetsos P, Angelis L, Stalmenos I (2006) Investigating the extreme programming system: An empirical study. Empirical Software Engineering, 11: 269–301.

Sharp H, Robinson H (2004) An ethnographic study of XP practice. Empirical Software Engineering 9: 353–375.

Stephens M, Rosenberg D (2003) Extreme programming refactored: The case against XP. APress, Berkeley.

Syed-Abdullah S (2005) Empirical study on extreme programming. Unpublished PhD thesis, University of Sheffield.

Tessem BR (2003) Experiences in learning XP practices: A qualitative study. Extreme Programming and Agile Processes in Software Engineering, Lecture Notes in Computer Science, 2675/2003:1012, Springer Berlin / Heidelberg.

Thomson C, Holcombe M (2009) The Sheffield software engineering observatory archive: Six years of empirical data collected from 73 complete projects. CS-09-01. Department of Computer Science, University of Sheffield.

Williams L, Krebs W, Layman L (2004a) Extreme programming evaluation framework for object-oriented languages – version 1.4. Technical report, NCSU.

Williams L, Krebs W, Layman L (2004b). Extreme programming evaluation framework for object-oriented languages – version 1.3. Technical report, NCSU.

Young M, Edwards H, McDonald S, Thompson B (2005) Personality characteristics in an XP team: A repertory grid study. HSSE '05: Proceedings of the 2005 workshop on Human and social factors of software engineering. St. Louis MO, ACM.

Table 1: Questionnaire topics

	Questions	**Original Weights**
Item 1	Automated Unit Tests	.40
Item 2	Customer Acceptance Tests	.20
Item 3	Test-First Design	.20
Item 4	Pair Programming	.80
Item 5	Refactoring	.70
Item 6	Release/Planning Game	.32
Item 7	Short Releases	.40
Item 8	Stand-Up Meeting	.05
Item 9	Continuous Integration	.60
Item 10	Coding Standards	.30
Item 11	Collective Code Ownership	.50
Item 12	Sustainable Pace	.30
Item 13	Simple Design	.55

Table 2: Descriptive statistics, Correlation and Covariance matrix for the 13 items

Item	Question	Mean	St. D	1	2	3	4	5	6	7	8	9	10	11	12	1
1	Automated Unit Tests	3.28	2.81	7.92	6.04	5.86	3.16	2.53	3.30	3.63	0.04	3.09	1.90	1.72	2.86	0.4
2	Customer Acceptance Tests	3.32	3.33	*0.64***	11.09	5.67	3.64	2.62	3.88	4.02	0.04	1.52	1.58	0.49	2.40	0.8
3	Test-First Design	2.86	2.76	*0.76***	*0.62***	7.61	3.55	2.52	3.44	4.01	1.39	2.93	1.15	1.70	2.58	0.2
4	Pair Programming	5.64	2.61	*0.43***	*0.42***	*0.49***	6.82	1.92	2.70	3.13	0.90	2.70	0.94	2.02	2.15	-0.2
5	Refactoring	3.98	2.26	*0.40***	*0.35***	*0.41***	*0.33***	5.09	2.62	2.44	0.81	2.20	1.42	0.56	1.49	0.2
6	Release/Planning Game	4.09	2.88	*0.41***	*0.40***	*0.43***	*0.36***	*0.40***	8.28	3.11	2.31	3.05	2.40	0.88	3.10	1.0
7	Short Releases	3.79	3.02	*0.43***	*0.40***	*0.48***	*0.40***	*0.36***	0.36***	9.12	1.02	3.34	1.21	1.79	2.56	0.5
8	Stand-Up Meeting	3.68	3.12	*0.00*	*0.00*	*0.16*	*0.11*	*0.12*	0.26**	*0.11*	9.71	1.65	0.97	1.08	1.47	1.3
9	Continuous Integration	5.04	2.94	*0.37***	*0.16*	*0.36***	*0.35***	*0.33***	*0.36***	*0.38***	*0.18**	8.66	4.05	3.35	2.56	0.9
10	Coding Standards	5.99	3.02	*0.22***	*0.16*	*0.14*	*0.12*	*0.21**	*0.28***	*0.13*	*0.10*	*0.46****	9.09	2.02	2.93	1.6
11	Collective Code Ownership	5.28	2.73	*0.22***	*0.05*	*0.23***	*0.28***	*0.09*	*0.11*	*0.22**	*0.13*	*0.42****	*0.25***	7.44	2.54	0.8
12	Sustainable Pace	5.31	2.36	*0.43***	*0.31***	*0.40***	*0.35***	*0.28***	*0.46***	*0.36***	*0.20**	*0.37****	*0.41****	*0.39****	5.59	1.3
13	Simple Design	6.16	2.21	*0.07*	*0.12*	*0.03*	*-0.04*	*0.04*	*0.16*	*0.08*	*0.20**	*0.14*	*0.24***	*0.13*	*0.26***	4.8

- * <p.05, ** p<.01, *** p<.001
- The upper right triangle are the covariances, the lower left triangle (in italics) are the correlations

Table 3: Model coefficients for the uni-dimensional model

	Unstandardized Coefficients	SE	Z	Standardized Coefficients
Automated Unit Tests	1	---	---	0.823
Customer Acceptance Tests	0.997	0.115	8.657*	0.693
Test-First Design	0.992	0.091	10.940*	0.833
Pair Programming	0.665	0.093	7.117*	0.59
Refactoring	0.509	0.082	6.181*	0.522
Release/Planning Game	0.720	0.103	6.963*	0.579
Short Releases	0.766	0.108	7.080*	0.587
Stand-Up Meeting	0.220	0.121	1.827	0.164
Continuous Integration	0.640	0.108	5.932*	0.503
Coding Standards	0.407	0.115	3.546*	0.312
Collective Code Ownership	0.374	0.104	3.608*	0.318
Sustainable Pace	0.575	0.085	6.738*	0.563
Simple Design	0.128	0.086	1.492	0.134

* P <.001

Table 4: Fit assessment for models of adherence to XP

Statistic	Uni-dimensional Model	Four-dimensional Model
χ^2	171.181**	97.984*
χ^2 df	65	59
χ^2 baseline	618.600**	618.600**
χ^2 baseline df	78	78
CFI	.804	.928
TLI	.764	.905
RMSEA	.109**	.069
RMSEA CF	.089–.129	.044–.093
SRMS	.091	.065

* p<.01, ** p<.001

Table 5: Model coefficients for the four-dimensional model

	Unstandardized Coefficients	SE	Z	Standardized Coefficients
Foundations				
Automated Unit Tests	1	0	--	0.850
Customer Acceptance Tests	1.011	0.107	9.465**	0.726
Test-First Design	1.001	0.084	11.937**	0.869
Pair Programming	0.620	0.089	6.929**	0.568
Refactoring	0.470	0.079	5.950**	0.499
Customer Planning				
Release/Planning Game	1	0	--	0.633
Short Releases	0.998	0.172	5.813**	0.602
Stand-Up Meeting	0.384	0.162	2.374*	0.224
Teaming				
Continuous Integration	1	0	--	0.765
Coding Standards	0.777	0.147	5.300*	0.580
Collective Code Ownership	0.639	0.129	4.939**	0.527
Craftsmanship				
Sustainable Pace	1	0	0	0.895
Simple Design	0.308	0.129	2.391*	0.295

* p<.05, ** p<.001

Table 6: Regression models for predictive validity

Weighted models	Predictors	β	Std. Error	t	*F* (df1,df2)	R^2	Δ*F*	ΔR^2
Uni-Dimensional Linear	Intercept	63.37	1.83	34.52***	0.001(1,148)	.000		
	Adherence	1.96	1.17	1.68				
Uni-Dimensional Quadratic	Intercept	64.67	1.82	35.56***	5.53(2,147)**	.07	11.06[a]* *	.07
	Adherence	0.28	0.68	0.41				
	Adherence^2	-1.73	0.52	-3.33**				
Four-Dimensional Linear	Intercept	55.72	2.08	26.79***	11.87(4,145)***	.247		
	Foundations	-2.00	.44	-4.51***				
	Customer Planning	.25	.48	.51				
	Teaming	1.83	.41	4.44***				
	Craftsmanship	1.27	.61	2.07*				
Four-Dimensional Quadratic	Intercept	53.16	2.35	22.62***	10.66(6,143)***	.309	6.46[b]**	.062
	Foundations	-1.76	.43	-4.06***				
	Foundations^2	-.61	.20	-3.02**			12.38[c]***	.239
	Customer Planning	.25	.46	.53				
	Teaming	1.95	.41	4.81***				
	Craftsmanship	1.79	.61	2.94**				
	Craftsmanship^2	.91	.34	2.65**				

* P<.05, ** p<.01, ***p<.001

a. Comparison between 1D linear and 1D quadratic

b. Comparison between 4D linear and 4D quadratic

c. Comparison between 1D quadratic and 4D quadratic

Figure 1: Uni-dimensional model of adherence to XP

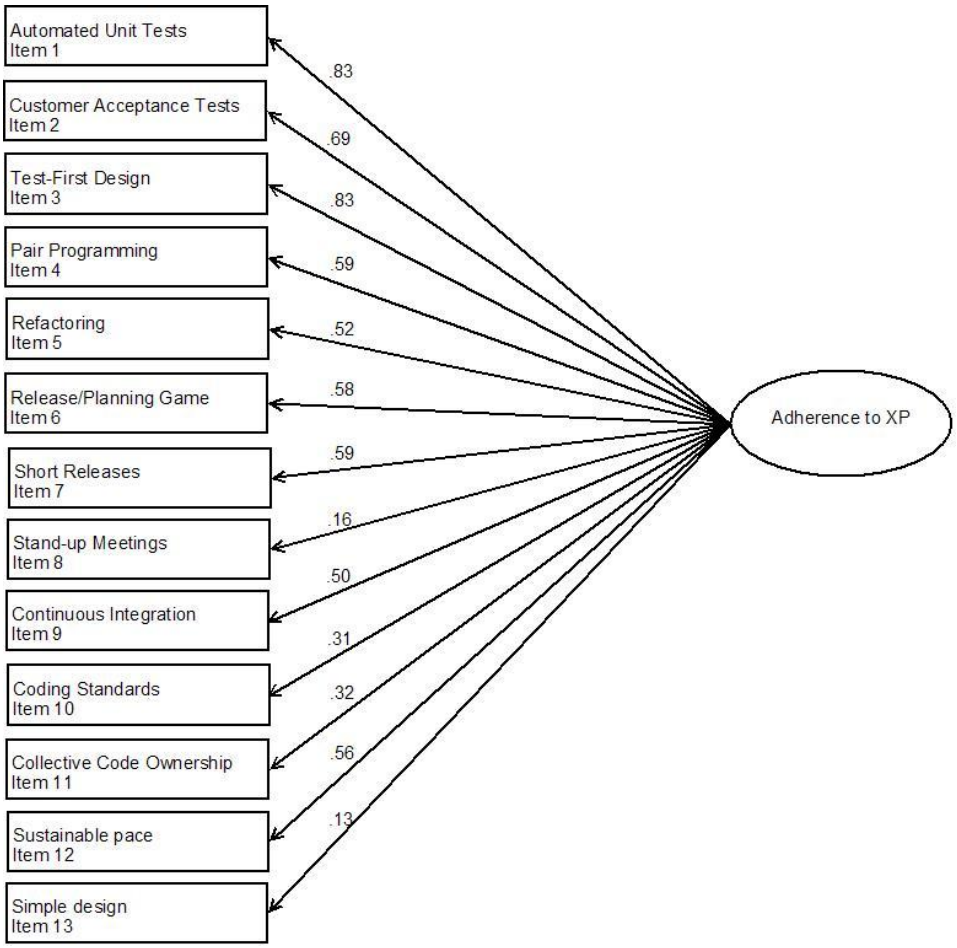

Figure 2: Hypothesized model (4 dimensions) of adherence to XP

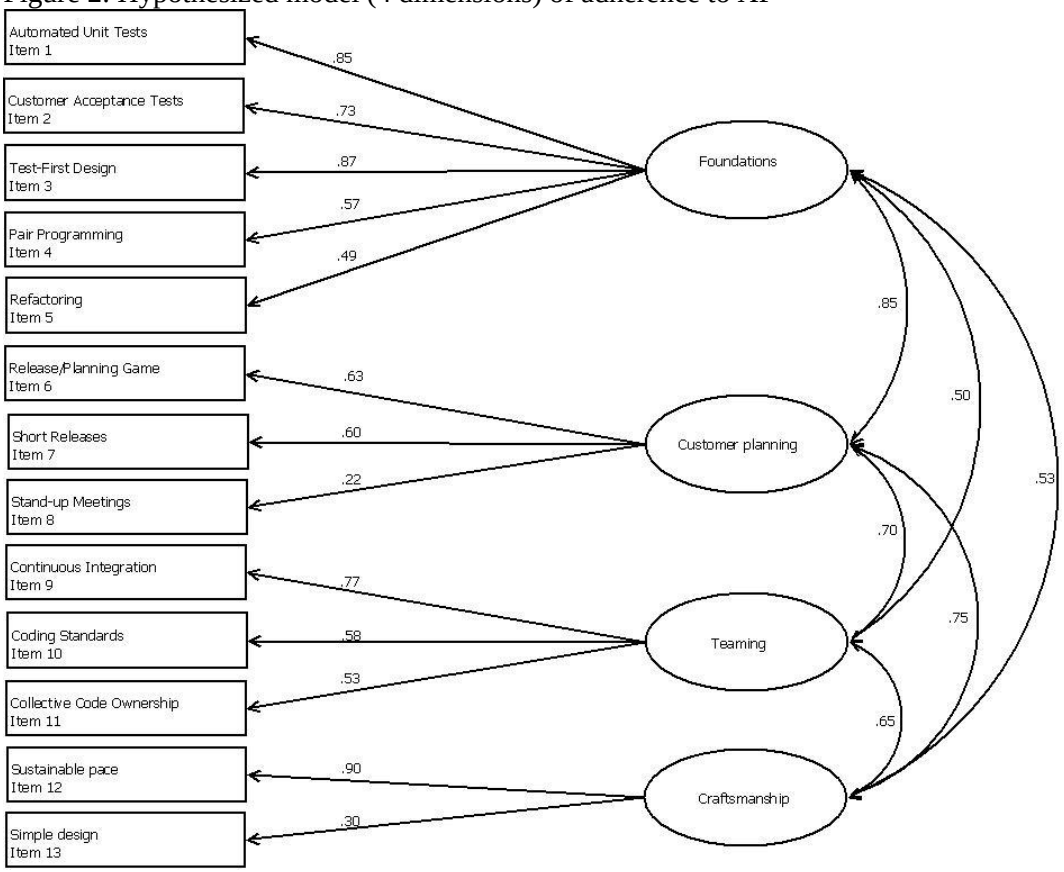

Figure 3: Plots of regression terms. Plotted data are the partial residuals and dotted lines are the pointwise standard errors