Measuring Software Product Quality: A Survey of ISO/IEC 9126

Ho-Won Jung and Seung-Gweon Kim, *Korea University*

Chang-Shin Chung, *Telecommunications Technology Association*

Rapid IT growth and the proliferation of PCs have resulted in numerous software products. For 2003, the Gartner Group estimates that worldwide user spending on software exceeded $730 billion, and IDC estimates that spending in the packaged software market topped $176 billion.[1,2] Although the market is rapidly growing, users are often dissatisfied with software quality. In fact, widespread dissatisfaction in the US drove recent legislative activity governing software quality and suppliers' responsibilities.[3] Similarly, user satisfaction is often considered a critical outcome of quality management, and studies show it as having a positive impact on organizational cost, profit, and sales growth.[4]

To address the issues of software product quality, the Joint Technical Committee 1 of the International Organization for Standardization and International Electrotechnical Commission published a set of software product quality standards known as ISO/IEC 9126. (See the sidebar for related standards.) These standards specify software product quality's characteristics and subcharacteristics and their metrics. However, some in the software engineering community have expressed concerns about a lack of evidence to support such standards. According to Shari Lawrence Pfleeger and her colleagues, "Standards have codified approaches whose effectiveness has not been rigorously and scientifically demonstrated. Rather, we have too often relied on anecdote, 'gut feeling,' the opinions of experts, or even flawed research rather than on careful, rigorous software engineering experimentation."[5]

The international standard ISO/IEC 9126 defines a quality model for software products. Based on a user survey, this study of the standard helps clarify quality attributes and provides guidance for revising the standard.

Stakeholders use evaluations of software product quality as a basis for many important decisions, including improving product quality, making large-scale acquisitions, and monitoring contracts. Given this crucial business role and the community's doubts about standards, we found it essential to empirically investigate whether the ISO/IEC 9126 categorization is correct and reliable in evaluating user satisfaction with the judgment of a packaged software product's quality.

ISO/IEC 9126

Quality, according to ISO 8402, is "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs." ISO/IEC 9126-1 defines a quality model that comprises six characteristics and 27 subcharacteristics of software product quality (see Table 1). Because the quality model is generic, you can apply it to any software product by tailoring to a specific purpose.

ISO/IEC 9126 also defines one or more metrics to measure each of its subcharacteristics. For example, the quality level of a software product's functionality can be represented by measured values of its five subcharacteristics. Together, those subcharacteristics include the properties that the standard posits to constitute functionality. Following an appropriate aggregation method, you can combine the five subcharacteristics' measured (or survey) values into a single value on a composite index of functionality.

However, if any of functionality's subcharacteristics measures doesn't correlate to functionality, the aggregated value would fail to fully represent functionality's quality as ISO/IEC 9126 defines it. In such a case, the unrelated subcharacteristic should be moved into a more appropriate characteristic, so that each characteristic includes homogeneous subcharacteristics.[6]

Background and method

Because you can call each of ISO/IEC 9126's six characteristics a dimension, or factor,[7,8] in statistical terms, the standard's structure is a multidimensional concept of software product quality. (A similar multidimensional concept would be job satisfaction. For example, you can be more satisfied or less satisfied with various combinations of your job, supervisor, pay, and/or workplace.) Our study aims to evaluate empirically ISO/IEC 9126's dimensionality, or the classification of characteristics, and internal-consistency reliability. For this purpose, we conducted a survey in which users evaluated their satisfaction concerning the quality of a packaged software product according to the criteria of ISO/IEC 9126's subcharacteristics.

Related Standards

Working Group 6 of Subcommittee 7 (Software and Systems Engineering Standardization) under Joint Technical Committee 1 of the International Organization for Standardization and the International Electrotechnical Commission is developing standards and technical reports for software product evaluation and metrics. WG 6 has developed ISO/IEC 9126-1 (quality model), 9126-2 (external metrics), 9126-3 (internal metrics), and 9126-4 (quality in use metrics). WG 6 has also developed the ISO/IEC 14598 series of technical reports, which provide guidance for evaluating software product quality on the basis of ISO/IEC 9126.

JTC1 publishes six types of documents. Each type follows all or part of the following development stages: preliminary, proposal, preparatory, committee, approval, and publication.[1] ISO/IEC 9126-1 is the international standard and 9126-2, -3, and -4 are technical reports.

Reference

1. ISO/IEC JTC1 Directives, *Procedures for the Technical Work of ISO/IEC JTC1, ISO/IEC JTC1*, ISO, 1999; www.jtc1.org/directives/toc.htm.

In ISO/IEC 9126, "satisfaction" implies "the capability of the software product to satisfy users in a specified context of use." Satisfaction in that sense refers to the user's response to interaction with the product. It includes judgments about product use rather than about properties of the software itself. Of course, in addition to ISO/IEC 9126's definition, user satisfaction also can include facets such as the quality of service, cost, or developer's reputation. However, we limited our survey questions' wording to cover only the subcharacteristics as defined by ISO/IEC 9126.

Table 1

Characteristics and subcharacteristics in ISO/IEC 9126

Characteristic	Subcharacteristics
Functionality	Suitability, accuracy, interoperability, security, functionality compliance*
Reliability	Maturity,* fault tolerance,* recoverability,* reliability compliance*
Usability	Understandability, learnability, operability, attractiveness, usability compliance*
Efficiency	Time behavior, resource utilization, efficiency compliance*
Maintainability	Analyzability, changeability, stability, testability, maintainability compliance*
Portability	Adaptability, installability, replaceability, coexistence, portability compliance*

**Denotes the omitted subcharacteristics in our survey*

Using too few rating-scale categories won't capture the questions' full discriminatory powers. Using too many might be beyond the respondents' discriminatory powers.

The survey

A marketing department identified 200 users of its company's packaged software product—a query and reporting tool for business databases. Of the 200 people we contacted through email or telephone, 75 (38 percent) responded to the questionnaire. The respondents included 48 end users, 25 developers who had modified the tool to embed it in their applications, and two others—all of whom we refer to as *users*. We were unable to find any significant differences in end-user and developer responses.

Rating scale

If you use too few rating-scale categories, the response won't capture the questions' full discriminatory power. On the other hand, using too many categories might be beyond the respondents' limited discriminatory powers. A Monte Carlo study of the number of scale points' effects on reliability showed that reliability estimates increased as the number of scale points increased from two to five, but the estimates decreased as more categories were added.[9] So, we used a five-category rating scale, which ranged from very satisfied to very dissatisfied, and included a "Don't know" option.

Using this scale, we measured the respondents' judgments of the product's capability to satisfy the ISO/IEC 9126 quality subcharacteristics. For example, one question asked, "How do you evaluate the capability of the software product to provide an appropriate set of functions for specified tasks and user objectives?" We then converted those ratings into numerical values from 1 (very dissatisfied) to 5 (very satisfied). This is a common procedure in software engineering and management information systems.

It's not uncommon for a survey to have missing values as well as "Don't know" replies. Our study considered "Don't know" as a missing value. We used a statistical technique called *imputation* to estimate the missing values.[10] Therefore, we could keep all 75 responses in the data analysis.

Data collection

We performed two pretests of the questionnaire to improve the clarity of the questions concerning the ISO/IEC 9126 subcharacteristics. We conducted the first pretest with 10 graduate students majoring in management information systems. From those results, we revised the questionnaire and then tested it using five industry end users again. The second pretest results led to minor revisions.

Because many pretest respondents had difficulty relating the ISO/IEC 9126 reliability subcharacteristics to their experience with the product, we removed the reliability questions. In addition, most respondents had the same trouble with the compliance subcharacteristics, defined for each characteristic as "the software product adheres to standards, conventions, or regulations in laws and similar prescriptions relating to each characteristic." So, we removed the compliance subcharacteristic for each of the remaining five characteristics. In the end, our questionnaire covered 18 subcharacteristics of the five characteristics. During the survey, we encouraged feedback from respondents regarding the questionnaire's contents.

Dimensionality's role

Principal component analysis (PCA) is a statistical method used to investigate an underlying concept's dimensionality.[8] Dimensionality is a concept related to the correlation between each of the measured subcharacteristics and derived dimensions. A dimension consists of its high-correlated subcharacteristics. Thus, subcharacteristics within the same dimension are highly correlated. Correlation coefficients can range from –1 (a perfect negative correlation) to 1 (a perfect positive correlation). The greater the absolute correlation, the stronger the relationship. Note that subcharacteristics within the same dimension have the same sign of correlation coefficient.

Investigating dimensionality has two important aspects. First, before you can use sums or means of measured subcharacteristics' values, you must ensure that the all of the subcharacteristics that you're summing are measuring the same thing. Second, estimating internal-consistency reliability should use subcharacteristics' values within the same dimension. If not, the estimate is deflated.

The correlation coefficients are criteria for determining the dimensionality classification's quality. Even though there's no absolute criterion, a value of more than 0.6 meets a guideline of "mediocre."[8]

Results

Table 2, a simple representation of our PCA

Table 2

Correlation coefficients between subcharacteristics and dimensions (>0.6)

	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
	Analyzability (0.616)	Understandability (0.769)	Time behavior (0.805)	Suitability (0.818)	Security (0.856)
	Changeability (0.653)	Learnability (0.827)	Resource utilization (0.766)	Accuracy (0.648)	
	Stability (0.814)	Operability (0.848)		Interoperability (0.796)	
	Adaptability (0.669)	Attractiveness (0.616)			
Cronbach's alpha	0.817	0.850	0.758	0.750	–
Percent of variance	18.67	17.42	11.90	11.82	7.80
Cumulative percent of variance	18.67	36.09	47.99	59.81	67.61

results with the varimax option, shows five dimensions and their subcharacteristics, with their correlation coefficients in parentheses. The table shows only subcharacteristics that have a correlation coefficient of greater than 0.6, along with their dimensions. Thus, of the study's 18 subcharacteristics, the table shows only 14. The remaining four subcharacteristics (testability, installability, replaceability, and coexistence) did not relate to any dimension.

Summary

We can summarize Table 2's results as follows:

Dimension 1 comprises four subcharacteristics: analyzability, changeability, stability, and adaptability. This subcharacteristics categorization isn't consistent with the ISO/IEC 9126 definition. This result implies that the four subcharacteristics actually measure the same intrinsic concept, which is a mixed concept of "maintainability and portability."

Dimension 2 comprises four subcharacteristics of usability in the ISO/IEC 9126 definition.

Dimension 3 is compatible with two subcharacteristics of efficiency in the ISO/IEC 9126 definition.

Dimension 4 comprises three subcharacteristics of functionality; however, the security subcharacteristic isn't closely related to the other functionality subcharacteristics.

Dimension 5 has a single subcharacteristic: security. A dimension of a single subcharacteristic isn't recommended because of the likely measurement error.[7] Therefore, if further studies reach the same classification, security should be classified as a characteristic rather than a subcharacteristic, and additional security subcharacteristics should be defined.

Internal-consistency reliability

Measurement always includes some amount of random error.[7,8] Subcharacteristics are relatively reliable if random measurement error minimally affects them. Ambiguities in wording and inconsistencies in respondents' interpretations can affect the measurements' consistency. Therefore, to give confidence to our results, it's crucial to estimate the amount of error.

Cronbach's alpha coefficient is a popular method used to estimate the internal-consistency reliability of measurements in software engineering.[7,11] Because Cronbach's alpha is applied to unidimensional constructs, it's more appropriate to compute the internal consistency for each dimension separately. In the early stages of the research, a Cronbach's alpha of 0.7 or higher is considered sufficient.[11] Table 2 shows that Cronbach's alpha for each dimension is greater than the recommended value.

The percentages of variance in Table 2's second-to-last row are summary measures indicating how much of the total original variance of all subcharacteristics each dimension represents. The first dimension explains 18.67 percent of total variance of the 18 subcharacteristics. As the last row shows, all of the five dimensions explain 67.61 percent (cumulative percent of variance) of total variance.

Discussions

Users' perceptions about product quality vary across user types. For example, end users typically value usability more than developers do. As we noted previously, however, we couldn't find any significant differences in their responses.

Recall also that we omitted the ISO/IEC 9126 reliability and compliance subcharacteristics because of the difficulty our pretesters had in relating those concepts. We believe that definition of reliability in the standard is clear and that the reliability subcharacteristics probably do constitute a distinct dimension. In contrast, including the six compliance subcharacteristics leads to two possible results. The first possibility is that the compliance subcharacteristic for each characteristic belongs to its characteristic as defined in ISO/IEC 9126. The second possibility is that the six compliance subcharacteristics constitute a separate dimension.

Our study's limitations, although characteristic of most empirical studies, should be made clear in interpreting our results and are worth explaining here. Our results are from evaluating a single packaged-software product. No single study can be fully definitive and widely generalizable. Further empirical work with larger and broader samples is necessary to help us better understand the underlying dimensional structure. Future studies should also examine other products to see if the dimensions are consistent across different software types.

In addition, we based our study on the reports of those who use or modify a software product. In large software acquisitions, software buyers might be different from the users who participated in this study. Thus, future work should examine the difference in dimensionality between buyer and user.

We also based our study on a sample data collected in Korea. Cultural differences might exist in recognizing and making judgments about the ISO/IEC 9126 quality subcharacteristics. Because the standard is intended for worldwide use, we hope to see this study replicated in other countries.

Our results reveal ambiguities in the way that ISO/IEC 9126 is structured in terms of characteristics and subcharacteristics. Although the results also provide evidence of much of that structure's validity, additional empirical work is necessary for clarity.

Systematic sample surveys can provide meaningful and widely generalizable statistical results at a moderate cost. However, the survey data should be augmented with more comprehensive measures of product quality in future studies. Replications of our study using other statistical analytic methods such as confirmatory factor analysis are also necessary to substantiate or clarify the present results.

About the Authors

Ho-Won Jung is a professor in the Department of Business Administration at Korea University. He's also the international SPICE research coordinator for the emerging ISO/IEC 15504 (Process Assessment) international standard and an authorized instructor for introductory courses in the SEI Capability Maturity Model Integration approach. He received his PhD in management information systems from the University of Arizona. He's a charter research member of the International Process Research Consortium and is on the *Software Quality Journal* editorial board. Contact him at the Dept. of Business Administration, Korea Univ., Anam-dong 5Ka, Sungbuk-gu, Seoul 136-701, Korea; hwjung@korea.ac.kr.

Seung-Gweon Kim is a PhD candidate in business administration, majoring in management information systems, at Korea University. He also worked at the Korea IT promotion agency. His research interests include software quality management and IT performance analysis. He's also the assessor of ISO/IEC 15504 and Capability Maturity Model Integration. Contact him at the Dept. of Business Administration, Korea Univ., Anam-dong 5Ka, Sungbuk-gu, Seoul 136-701, Korea; sunggweon@korea.ac.kr.

Chang-Shin Chung is a principal member of the technical staff for the Software Quality Evaluation Team at Telecommunications Technology Association, Korea. He received his PhD in computer engineering from Chungbuk National University. His research interests are software product quality testing and evaluation. Contact him at the Telecommunications Technology Assoc., 267-2 Seohyeon-dong, Bundang-gu, Seongnam City, Kyunggi-do, 463-824, Korea; cschung@
tta.or.kr.

References

1. G. Shiffler et al., *Gartner Dataquest Market Databook, March 2003 Update*, Gartner, 2003.
2. IDC, *Worldwide Black Book*, vol. Q2, 2003.
3. C. Kaner, "Software Engineering and UCITA," *The John Marshall J. Computer and Information Law*, vol. 18, no. 2, 2000.
4. C.N. Madu, C.H. Kuei, and R.A. Jacob, "An Empirical Assessment of the Influence of Quality Dimensions on Organizational Performance," *Int'l J. Production Research*, vol. 34, no. 7, 1996, pp. 1943–1962.
5. S.L. Pfleeger, N. Fenton, and S. Page, "Evaluating Software Engineering Standards," *Computer*, vol. 27, no. 9, 1994, pp. 71–79.
6. T.L. Saaty, *Fundamentals of Decision Making and Priority Theory*, RWS, 1994.
7. J.P. McIver and G.E. Carmines, *Unidimensional Scaling*, Sage Univ. Paper Series on Quantitative Applications in Social Sciences, Sage, 1981.
8. S. Sharma, *Applied Multivariate Techniques*, John Wiley & Sons, 1996.
9. R.W. Lissitz and S.B. Green, "Effects of the Number of Scale Points on Reliability: A Monte Carlo Approach," *J. Applied Psychology*, vol. 60, no. 1, 1975, pp. 10–13.
10. A. Carlos and F. Harrell, *An Introduction to S and the Hmisc and Design Libraries*, ch. 4, 2004, http://biostat.mc.vanderbilt.edu/twiki/pub/Main/RS/sintro.pdf.
11. J.C. Nunnally and I.H. Bernstein, *Psychometric Theory*, McGraw-Hill, 1994.