# ISO 20282: Is a Practical Standard for the Usability of Consumer Products Possible?

Nigel Bevan[1] and Simon Raistrick[2]

[1] Professional Usability Services, 12 King Edwards Gardens, London W3 9RG, UK
[2] Foviance, 14 Bonhill Street, London, EC2A 4BX, UK
mail@nigelbevan.com, simon.raistrick@foviance.com

**Abstract.** The current ISO 20282 standard is intended to provide a test method that can be used to assess the ease of operation of walk-up-and-use products and consumer products. The standard is currently being revised to improve the cost-effectiveness of the test method, to broaden the scope to include aspects of usability that go beyond ease of operation, and to more clearly define how to obtain reliable results. But the challenge remains to provide a standard that will be cost-effective and useful to manufacturers, purchasers and test houses.

**Keywords:** standards, usability, consumer products, measurement, test method.

## 1 Why Aren't Consumer Products Usable?

There are many reasons why a product may not be usable. Some manufacturers try to make products usable but lack the skills or measures to do so, whereas others consciously decide not to adopt usability best practise. This can be due to internal politics, or due to a customer experience management strategy that prioritizes other aspects of the customer experience over usability. However, one common factor across many industries is that senior executives are often unaware that the products their companies produce are unusable, due to a lack of effective comparative measures.

In contrast, the market is deeply concerned with usability, as is often evidenced by poor results in product reviews. For example, the ViaMichelin X-930 portable GPS navigator received a very poor review from PC Magazine [6]:

> "Overall, I was sorely disappointed with the ViaMichelin X-930. Its menus were difficult to navigate and inefficiently laid out. For example, to find a list of the nearest gas stations from the map view takes a time-consuming ten screen taps. … For entry-level GPS devices, there are others on the market that are better and easier to use."

Almost all consumer organisations and magazine product reviews rate usability as an important factor, so why do they not use more thorough and comparable usability measures, and why do manufacturers not give usability a higher priority?

Unfortunately people have to purchase many products without the opportunity to try using them, and shop assistants only have very limited knowledge about usability [3]. A UK TV program recently highlighted customers' frustration at not being able to try mobile phones before purchase. This led to a redesign of a store to allow more interaction with products before purchase [14].

A. Marcus (Ed.): Design, User Experience, and Usability, Pt I, HCII 2011, LNCS 6769, pp. 119–127, 2011.

But even when it is possible to try products in a shop, it is difficult to assess the usability because usability can only be fully experienced by using a product in real life. And this is not feasible within the short time frame of the purchasing decision. It is a challenge to systematically identify the important tasks and try them all. At best, it is usually only possible to practice some basic tasks with a device [11].

In general, there seems to be little incentive for a manufacturer to worry about whether consumers can use their products. Indeed, the early adopters of innovative technology are more concerned with the excitement and status of owning and using an innovative device, than with whether it is actually useful.

As far as the market is concerned, this is often true of technologies that are in their infancy, but for users of technologies which have gone through several iterations, lack of usability is a time-consuming frustration which gradually moves higher up their agenda. Some consumers return products they have purchased complaining they are faulty, when the real explanation is that they were unable to work out how to use the product. For example, a study of mobile device returns in the United Kingdom showed that 1 in 7 mobile phones was returned within the first year of purchase as faulty [12]. Of these returns, about 63% had no hardware or software fault but the reported problems related to usability, mismatch with user's expectations or issues relating to the configuration of the handset.

It would be a great benefit if reliable, comparable information about usability could be provided to potential purchasers. This was the original objective of the ISO 20282 standard.

## 2 Standards for Usability

The ultimate test of the usability of a product is that users are effective, efficient and satisfied when using it [1]. These criteria were established in ISO 9241-11 in 1998, and subsequently formed the basis for the Common Industry Format for Usability Test Reports [3], which became ISO/IEC 25062. The intended audience for this standard was corporate purchasers, providing a standard format for suppliers to provide usability information to potential purchasers.

In 2000, work started on a proposed standard for assessing consumer product usability [2]. Gradually the work focused on walk-up-and-use products (such as ticket machines) as these were identified as being easier to test reliably as they have well defined user groups and tasks.

The standard ISO TS (Technical Specification) 20282-2 published in 2006 specifies in detail the procedure for summative testing intended to ensure that results are reliable and consistent.

To cover consumer products, two further standards in the 20282 series were developed, ISO PAS 20282-3 Test method for consumer products, and ISO PAS 20282-4 Test method for the installation of consumer products. These were published with the lower status of Publicly Available Specifications to reflect the need for the validity of the methods for consumer products to be checked.

## 3 User Requirements for Measuring Consumer Product Usability

With little evidence emerging of the practical use of ISO 20282, in 2009 a new ISO group was established [8] to review the approaches to assessing consumer product usability.

The potential benefits of providing a standard summative test method for different stakeholder groups was reviewed:

a) For corporate and government purchasers of walk-up-and-use vending machines, usability is essential for their success. Information about usability could contribute to purchasing decisions, support a more effective tendering process, and improve user adoption.
b) Manufacturers of both walk-up-and-use and consumer products could benefit from using usability test results in marketing to differentiate their products as well as in quality control.
c) Organizations commissioning or carrying out tests (typically consumer organizations) could benefit from a valid test that provided more reliable results, although it might be more expensive than current test methods.
d) Consumers would have objective, comparable information about the level of usability of a product at the pre-purchase stage.

The test method has to be cost-effective to implement, and there was some concern that the test method in 20282-2 may be viewed as unacceptably expensive because of the large number of participants needed to obtain statistical significance. In order to estimate a population success rate greater than 75% with 95% confidence, which was seen as a suitable success rate, a maximum of one unsuccessful user out of 17 would be tolerable, requiring a test of a minimum of 17 users.

Investigation showed that the most recent statistical procedures (the adjusted Wald test [13]) required fewer participants, and the less stringent criterion of an estimated population success rate greater than 75% with 80% confidence could be achieved with one unsuccessful user out of 9. This would bring the test method in line with common practice of testing 10-12 participants to assess usability.

## 4 Summative Test Procedure

The test procedure proposed is specified in sufficient detail to ensure reliable and consistent results. (The revised version of 20282-2 has made few changes.) The steps are:

a) Ensure that the people who will carry out the test have an acceptable level of knowledge in usability testing.
b) Identify the product to be tested.
c) Identify whether the product is within the scope of the standard.
d) Identify the main goals of use of the product.
e) Define the main goals to be tested.
f) Establish criteria for main goal achievement.
g) Identify the tasks.
h) Specify the user groups to be used for testing.

i) Specify relevant environmental characteristics which affect usability.
j) Check that the product is compatible with intended user characteristics.
k) Decide whether to test one or more user groups.
l) Identify which measures are required, and whether there is a particular value which defines a success criteria for each measure, or whether it is a purely comparative measure.
m) Decide the desired confidence level for the testing.
n) Specify test scenarios and conditions, with clear criteria for goal achievement.
o) Recruit a representative sample of users (that represents the intended user group(s) of the product).
p) Test the product in an environment that resembles as closely as possible the environment in which the product would be used.
q) Establish a written test procedure using best practice usability testing, for example ensuring that the user is not led or prompted.
r) Measure success rate, task time and satisfaction.
s) Calculate effectiveness (percentage success rate), efficiency (median task time) and satisfaction (mean questionnaire scores).
t) If success criteria for task time and satisfaction were identified, the proportion of users that met these should be calculated.
u) Prepare a full report and, if required, a short summary.

## 5 Reliability of Usability Measures

As part of the development of ISO20282-3, a detailed analysis was made of the factors that would determine the reliability of the test method for a range of products and systems used by consumers. The conclusion was that it would be difficult to obtain reliable measures of usability for certain types of “complex” products where any of the following are true:

- It is not possible to define the user's goals in a clear and repeatable way (for example where these goals vary considerably between users).
- The criteria for success are difficult to define (for example where they include an element of subjectivity or creativity).
- It is not possible to reliably measure the success of the outcome in a repeatable way.
- Success is highly dependent on the particular data or subject matter (for example when the content of the tasks includes highly variable parameters, such as when booking flights).

Based on these criteria, it was concluded that the following are examples of products that could in principle be tested reliably:

- Fire extinguisher.
- Butterfly identification website.
- Pregnancy scanning machine.
- House paintbrush.
- Camera: for common goals only.

- A simple ecommerce shopping website (e.g., bookshop/rental car booking).
- Sewing machine (although quality criteria are hard to define).
- Oven - for the simpler task of reheating food only.
- Mobile phone - only for simple and well-defined goals.

The following products could not be tested reliably:

- Complex ecommerce website (e.g., purchasing a laptop, airline booking), excluding the less complex parts (e.g., basket and checkout) which do not include selection processes.
- Oven – for the more complex tasks of cooking food from ingredients.
- Microsoft Word – the content variability is too large and the quality of the main success criteria are not easily measurable (although it might be possible to test it for specific purposes).
- Car – too many goals and success measures (although car entertainment and navigation devices could be tested).

Note that the products that could not be tested reliably in their entirety, could be tested for specific goals e.g. the checkout part of a complex ecommerce site, or the starting of the car, or the creation of a new blank document in Microsoft Word, but in their entirety the products cannot be reliably tested.

In complex products like a flight booking web site, the usability is difficult to measure because it depends on the precise goal (for example there may be trade-offs between price and convenience), the particular range of potential alternative flights available to satisfy a particular query, and the features and ease of operation of the web site to support the particular goal. Although it is possible to compare aspects of the usability of different airline booking sites, there may be a temptation to select unrealistic or unrepresentative simple benchmark tasks, whereas in reality different users may have different complex goals. One could give examples of specific tasks that are easy or difficult to achieve on different sites, but it would be very difficult to obtain reliable and consistent measures.

In the first instance, the scope of 20282-2 will be for consumer products that do not have complex goals, but additional parts of 20282 could be produced for other types of products and systems that do not have complex goals, such as some types of web sites.

## 6 Assessing Goal Achievement

### 6.1 Defining Clear Success Criteria

Even for consumer products that are not "complex", one challenge is to find clear criteria for goal achievement. The criteria for what constitute successful achievement of a goal should match as closely as possible the criteria that would be applied by a typical user, rather than be limited by any constraints of the technology.

For some consumer products, there may only be one acceptable result (e.g., setting an alarm correctly). For others there may be a range of results that are acceptable to the user (for example a range of temperatures on an oven, or dryness of clothes from a tumble dryer).

The experts in the ISO group discussed the method that they would recommend using when evaluating the usability of an oven, which produced a surprising range of approaches to assessing goal achievement:

a) As the ultimate goal is to cook food correctly, some people suggested that the success measure should be that the food is cooked acceptably (e.g. a cake). Some thought that the quality of cooking should be judged by an expert, some that it should be judged by each user, and some that the method would be unreliable without an objective test.
b) Others thought that the goal should only be to set the oven to the intended temperature, but as the controls might be inaccurate, the actual temperature of the oven should be independently measured using a thermometer.
c) Another group thought that only the correct setting of the oven controls needed to be assessed (i.e. the scope of the testing should be the user interface only, rather than the whole product)

These issues with defining clear success criteria raise a much wider issue of scope: should the scope of the standard just be ease of use of the interface (called "ease of operation" in the current 20282-2), or should it be achievement of the user's goal (as in the ISO 9241-11 definition of usability)? For example, is an oven usable if it cooks at 200°C when set to 180°C? Or is an oven usable if it cooks at the correct temperature but the cakes are burnt at the end of cooking?

### 6.2 Who is Best Qualified to Judge Success?

In cases where the success criteria depend on some level of judgment, a new question arises: who should judge whether a goal has been achieved? For example, when cooking cakes, should it be an experienced cook or each user?

a) If expert judgment is used to assess goal achievement, would two experts make the same judgment?
b) If individual users are to judge, this introduces a new variable into the measures, and factors such as cultural variation could cause different tests to be incomparable. On the other hand, if experts are to judge, a standard measure of acceptability must be agreed on, and industry standard measures do not always exist, depending on the product being tested.

This issue of who should judge the results is compounded by potential variability in the skill or experience of the user. For example, some irons or sewing machines may make it difficult for less experienced users to produce acceptable results. In most cases the target audience will includes a mix of experienced and inexperienced users, and in order to be a realistic test, the variations in experience of these groups should be accounted for. Since these groups will have different expectations not only in usage, but also in quality of output, this further complicates the question of who should judge success.

### 6.3 Technical Tests and Usability Tests

An investigation into the testing methods used by the UK Consumers Association [5] revealed the following tests are performed:

- Oven: technical tests of heating accuracy, and expert assessment of ease of use and quality of cooked food.
- Camera: technical tests of image quality and ease of operation of the interface.
- Iron: expert review of speed of ironing and quality of results, and user review of ease of use.
- Sewing machine: expert and novice reviews of ease of use, and expert reviews of quality of results for experts and novices.

In these examples there are two independent sets of tests:

a) Technical or expert assessment of the capability of the product to produce acceptable results.
b) Expert and/or novice ratings of the ease of use.

These two separate tests should in principle produce results that are just as reliable as a single combined test, except when there is interaction between user operation and successful goal achievement. For example, some cameras may only achieve sharp telephoto pictures when held very steadily, while others (with an effective image stabilizer) will also achieve good results despite camera shake. In this case, measuring the ease of use of the controls independently of the technical quality of the photos would be insufficient to assess the ability to achieve the goal of a clear photo.

The potential inaccuracy in such cases could be circumvented either by performing a technical test to simulate the effect of camera shake, or by assessing the quality of photos produced during real use.

### 6.4 Approaches to Measuring Quality as Part of Usability

To resolve these problems, the ISO group is currently considering inclusion of the following procedure in the standard:

a) If a goal has only one outcome (for example, turning a product on) or has easily identified outcomes (for example, being woken at a particular time), achievement of the intended outcome shall be used as the basis for goal achievement.

   For example, the goal would not be met if a user believes that they have successfully set an alarm to ring at a particular time, but because they omitted the final step of pressing the alarm-on button, have failed to achieve the goal.

b) If the adequacy of the intended outcome of using the product can depend on how the product is used or operated (for example, if the quality of a picture is influenced by the steadiness with which a camera is held), the following tests for goal achievement should be considered in order of preference:

1. Carry out technical tests of the adequacy of goal achievement, if this is cost-effective and there are standards or other published criteria for what constitutes an acceptable level of quality in the resulting outputs.
2. Carry out expert assessment to define the quality criteria for adequate goal achievement, if the results would be reliable and cost-effective. These criteria should be based on standards or other published criteria for technical results that fall within an acceptable range.
3. Ask users to assess the acceptability of the quality of the outcome for defined purposes, if this is easy for a user to judge consistently by inspection.

Each of these methods has advantages and disadvantages, in particular, different users may use different criteria, and when using technical criteria these can be unwieldy and costly, and may not cover the same range as the criteria a user or expert would use.

For example, the quality of a photo could in principle be assessed by technical tests of the sharpness and color rendering, or by assessment by an expert, or by asking the user to assess the adequacy for a specific purpose (for example producing an A4 print). The quality of ironing could in principle be assessed by expert assessment, or by asking the user to assess the adequacy of the smoothness of particular clothes for a specific purpose.

The disadvantage of providing three options for the methods that can be used is that independent users of the standard might use different test methods. However, whenever comparisons are needed, the same method could be selected.

### 6.5 Big or Little Usability?

One objective of revising the standard was to broaden the scope from "ease of operation" (ease of use of the interface) to usability (effectiveness, efficiency and satisfaction when achieving goals). These could be regarded as "big" and "little" usability, referring to the wide scope versus the narrow scope.

The analysis above indicates that in some cases, in order to assess effectiveness reliably (as is required to cover "big" usability) complex technical tests may be required. These technical tests may be part of a broader test program (for example by a consumer organization), but could significantly increase the costs if this is not the case, and usability was being tested in isolation. In such situations, it may be desirable to use the standard to limit the scope to "small" usability, i.e. to just test the usability of the user interface, determining whether the controls can be set correctly regardless of the technical quality of the results.

While this type of testing is consistent with the popular interpretation of usability as ease of use, it does mean that even if a product is in this sense usable, it might also be useless, as it may not be able to technically perform the functions it was designed for. One objective of the definition of usability in ISO 9241-11 was to promote the practical and business relevance of usability interpreted as the quality of the product from a user perspective [1].

## 7 Conclusions

The current ISO 20282-2 standard contains a more rigorously specified procedure for obtaining reliable measures of usability than is available elsewhere, and is even more specific in the revised standard. The ISO 20282 series is intended to give acquiring organizations the opportunity ask for evidence of the usability of walk-up-and-use products and products for consumer use, and to give manufacturers the opportunity to publicize the results of usability test results. The method could also be adopted for use in certification schemes.

Some aspects of the method are still being developed, and academic and research organizations are invited to contact the ISO group (via the authors) to arrange trial use of the method, particularly to assess the consistency of results when the method is carried out by different test organizations.

Commercial organizations are invited to contact the ISO group (via the authors) to discuss the potential benefits of adopting the method.

The content of the standard is expected to be finalized and the revised standard published by ISO by 2013.

**Acknowledgements.** Thanks to Timo Jokela for helpful clarifications, and to the other contributors to the standard that have recently included: Tomas Berns, Susan Harker, Ran Linghua, Zhengjie Liu, Hans Persson and Lonneke Spinhof.

## References

1. Bevan, N.: Quality in use: meeting user needs for quality. Journal of Systems and Software 49(1), 89–96 (1999)
2. Bevan, N., Schoeffel, R.: A proposed standard for consumer product usability. In: Proceedings of 1st International Conference on Universal Access in Human Computer Interaction (UAHCI), New Orleans (2001)
3. Bevan, N., Claridge, N., Maguire, M., Athousaki, M.: Specifying and evaluating usability requirements using the Common Industry Format: Four case studies. In: Proceedings of IFIP 17th World Computer Congress, Montreal, Canada, August 25-30, pp. 133–148. Kluwer Academic Publishers, Dordrecht (2002)
4. Carswell, C.M., Lio, C., McNally, J.: How Knowledgeable are Salespeople about the Usability of Their Merchandise? In: Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting, v. 48, pp. 990–994 (2004)
5. Consumers' Association, http://www.which.co.uk
6. Ellison, C.: ViaMichelin X-930. PC Magazine. 2091601,00.asp (2007), http://www.pcmag.com/article2/0,2817
7. ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11 Guidance on usability (1998)
8. ISO TC159/SC4/WG11: Ease of operation of everyday products (2011), http://www.iso.org/iso/iso_technical_committee.html?commid=53372
9. ISO/IEC 25062: Software Engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Common Industry Format (CIF) for Usability Test Reports (2006)
10. ISO/TS 20282-2: Ease of operation of everyday products – Part 2: Test method for walk-up-and-use products (2006)
11. Jokela, T.: When Good Things Happen to Bad Products: Where are the Benefits of Usability in the Consumer Appliance Market? ACM interactions XI.6, 28–35 (2004)
12. Overton, D.: 'No Fault Found' returns cost the mobile industry $4.5 billion per year, http://www.wdsglobal.com/news/whitepapers/20060717/20060717.asp
13. Sauro, J.: Confidence Interval Calculator for a Completion Rate (2010), http://www.measuringusability.com/wald.htm
14. Withers, P.: Fonehouse revamps store look and brand. Mobile News Online (2011), http://www.mobilenewscwp.co.uk/2011/02/fonehouse-revamps-store-portfolio-and-brand