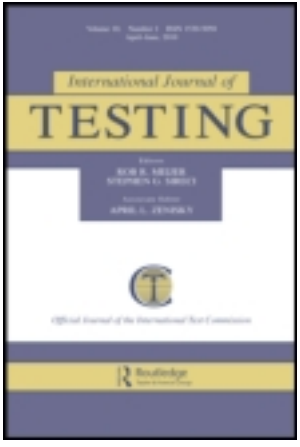

International Journal of Testing
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hijt20

Test Reviewing in Germany

Carmen Hagemeister [a] , Martin Kersting [b] & Gerhard Stemmler [c]

[a] Department of Psychology, Technische Universität Dresden, Dresden, Germany

[b] Department of Psychology, Justus Liebig University Giessen, Giessen, Germany

[c] Faculty of Psychology, Philipps-Universität, Marburg, Germany

Available online: 24 Apr 2012

To cite this article: Carmen Hagemeister, Martin Kersting & Gerhard Stemmler (2012): Test Reviewing in Germany, International Journal of Testing, 12:2, 185-194

To link to this article: http://dx.doi.org/10.1080/15305058.2012.657922

International Journal of Testing, 12: 185–194, 2012

ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2012.657922

Routledge
Taylor & Francis Group

Test Reviewing in Germany

Carmen Hagemeister
Department of Psychology, Technische Universität Dresden, Dresden, Germany

Martin Kersting
Department of Psychology, Justus Liebig University Giessen, Giessen, Germany

Gerhard Stemmler
Faculty of Psychology, Philipps-Universität, Marburg, Germany

In 2006, a (new) German standard for test reviewing was passed (Testkuratorium, 2006). There was already a European standard in place (European Federation of Psychologists' Associations, 2008). This article presents the German standard for test reviewing and explains how the German test review system was derived from demands in the German standard DIN 33430 for proficiency assessment procedures. Crucial decisions that were made prior to the construction of the standards are clarified, including why other (e.g., European/Dutch) existing standards were not simply translated. Furthermore, advances, difficulties, challenges, and perspectives regarding the test reviewing system in Germany are discussed.

Keywords: DIN 33430, quality management, testing, test manual, test review

THE GERMAN STANDARD DIN 33430

As the German test reviewing system is based on the German standard DIN 33430 (DIN, 2002),[1] it is necessary to give a brief introduction of this standard before presenting the German test review system. DIN 33430 was passed in 2002, and contains demands relating to instruments and their application in proficiency assessment procedures. It is important to note that, unlike professional

Correspondence should be sent to Carmen Hagemeister, Technische Universität Dresden, Department of Psychology, Assessment and Intervention, Dresden 01062, Germany. E-mail: Carmen.Hagemeister@tu-dresden.de

standards, which only hold for certain groups such as psychologists (Berufsverband Deutscher Psychologinnen und Psychologen, 2005), the demands of DIN 33430 cover all proficiency assessment procedures, irrespective of the profession of the assessor. As most proficiency assessment procedures in Germany are carried out by persons of other professions, this is an important step toward improving the quality of such procedures.

At first glance, however, it is not clear why a test review system should be based on the 33430 standard. First, DIN 33430 only covers proficiency assessment procedures, not assessment procedures in other fields. Second, DIN 33430 is a standard for the assessment process and not for the products used in this process. Third, DIN 33430 states no prerequisites specifically for tests. The requirements apply equally to all instruments used in the assessment process, that is, interviews and assessment centers as well as tests. Nevertheless DIN 33430 contains demands that have to be met by tests and their documentation (test manuals). These demands are necessary, but not in themselves sufficient to allow an assessment procedure in which the respective tests are used to be deemed DIN conform. This means that a test that in itself meets the demands of the DIN 33430 can nevertheless be used in an assessment procedure in an improper manner; the result is an assessment process that is not DIN conforming.

Kersting (2006, 2008) extracted 318 statements from the DIN 33430 standard. He combined these statements into checklists. One hundred and forty of these statements refer to test manuals and can be found in DIN Screen Checklist No. 1. These statements referring to test manuals can be applied to any field of assessment. Many of the other statements in DIN 33430 only refer to proficiency assessment procedures.

DIN 33430 demands that test authors and test publishers provide the required information about the construction of the test and about empirical studies. They must provide information on how the test is applied, scored, and interpreted. If a test does not meet these demands it is not deemed to be in accordance with DIN 33430. But DIN 33430 is a standard for the assessment process. So no test (or any other instrument) used within this process can in itself be in accordance with DIN 33430. And yet any process in which a test with insufficient information in the test manual is applied will not be in accordance with DIN 33430. A deficit in one single instrument can suffice to spoil the DIN conformity of the process as a whole.

In DIN 33430, standards were set for proficiency assessment procedures, irrespective of the profession of the assessor. In Germany, nearly 90% of proficiency assessment procedures are conducted by nonpsychologists (Bartram, 2001), and it is important that standards for assessment—and testing—are not only set for assessors of a certain profession (e.g., psychologists). The standards for tests in DIN 33430 can be applied not only to tests constructed by psychologists but to all tests. Tests about which insufficient information is provided for the test user can be rated as insufficient for any assessment process.

DIN 33430 does not allow any conclusion about the quality of a test in a positive sense. No test can be classified as good without due respect to the question, the circumstances, the tested person, and the qualification of the testing person. DIN 33430 refers to the assessment process as a whole. Only if the necessary information about a test is provided in the test manual can the assessor decide if it is appropriate to use this particular test to test this particular person for this particular question under these particular circumstances. If such information is lacking, the process does not conform to the DIN standard.

THE CURRENT GERMAN TEST REVIEW SYSTEM

Prior to 2006, there was a list of criteria to use for test reviewing, but this list mainly provided definitions of the criteria. One exception was some remarks about what should be mentioned when describing on what the construction of a test was based (Testkuratorium der Föderation deutscher Psychologenverbände, 1986). This list did not describe the review process nor was it intended that reviews should result in ratings that would render tests comparable.

The German test review system was constructed in order to combine the advantages and avoid the disadvantages of other test review systems, including DIN 33430. The review process covers the following steps:

1. The Board of Assessment and Testing chooses the test to be reviewed. Anyone can propose tests for review.
2. The Board of Assessment and Testing mandates two reviewers to analyze the tests. The Board vouches for the independence and impartiality of the two reviewers.
3. The Board makes sure that the reviewers are provided with the test to be reviewed and with other necessary material. If the test is "confidential," the Board warrants the confidentiality of information that might be relevant under the law of competition.
4. The reviewing process itself consists of three steps:
   4.1. Checking if the test is eligible for review
   The reviewers check whether the test manual contains all pieces of information required according to DIN 33430. These demands can be applied to tests of all fields. The demands are operationalized in the DIN Screen Checklist No. 1 (Kersting, 2008). This checklist is the standard for the information that has to be provided for tests assessing human experience and behavior. This checklist should have already been filled in by the test publisher and should indicate on which page of the manual information can be found. The reviewers check this information and correct it, if necessary. The checklist is part of the

reviewing process, but it is not published as part of the review. So it is only available to test users if it is incorporated into the manual (as, for example, in the Wilde-Intelligenz-Test 2; Kersting, Althoff, & Jäger, 2008).

Based on the information in the DIN Screen Checklist No. 1, the reviewers decide whether the test is eligible for review. A test cannot be reviewed if crucial pieces of information which must be provided according to DIN 33430 are missing. In this case, the result of the review process is the statement "The test does not meet the demands regarding information and documentation as required in DIN 33430." No further reviewing is necessary.

4.2. Categorizing the test

In the second step, the test is categorized according to the systems of European Federation of Psychologists' Associations (EFPA, 2008) and Leibniz-Institute for Psychology Information (ZPID). Formal features of the test are described for data banks. For this step, the ZPID system and parts of the EFPA system (1.10.1–1.10.3 and 1.11; European Federation of Psychologists' Associations, 2008) are used. This information should also be provided by the test publisher, with references to the pages of the manual where the information can be found. The reviewers check this information and correct it, if necessary.

4.3. Reviewing according to the categories of the German test review system

The third and final step is the actual review by the reviewers. This review is based on the information provided in the test manual. The guidelines for this process explain which questions of the DIN Screen (Checklist No. 1) belong to which review category (e.g., DIN Screen question no. 1 "A test manual exists" belongs to the category "General information about the test in the manual and description of the test and its diagnostic aim"). The reviewers can use further information if it is available to the public. This procedure differs from the EFPA system (2008), where the evaluation of the test material allows the inclusion of material that is not public. Here, the rationale of the German review system is the same as that of DIN 33430, namely that relevant information about the test must be available to the "normal" user of the instrument.

The review covers seven categories and ends with a summarizing final evaluation. The categories and the kind of evaluation (only free text, or free text and structured format) are shown in Table 1. The formal evaluation for the categories 1, 3, 5, and 6 is given on a four-point

TABLE 1
Categories for the Review, Evaluation, and Maximum Text Length

Category	Evaluation	Maximum Number of Characters (Including Spaces) for the Free Text Evaluation
1. General information about the test, description of the test, and its purpose in assessment	Free text and structured format	1000
2. Theoretical basis of test construction	Free text	1000
3. Objectivity	Free text and structured format	1000
4. Norms	Free text	1000
5. Reliability	Free text and structured format	1000
6. Validity	Free text and structured format, including fairness (if requested)	1000
7. Further quality criteria (susceptibility to failure, non-fakability, and scaling)	Free text	1000
8. Final evaluation	Free text	2000

scale, "The test fulfills the demands completely/mostly/partly/not at all."

5. The review steps 4.1 to 4.3 are performed by the two reviewers independently.
6. When the Board of Assessment and Testing receives the two reviews, the mutual anonymity of the reviewers is revoked, and the two reviewers are asked to create a common version of the review.
7. If the reviewers cannot agree on a common version, the relevant differences of their position are presented. The Board can augment the maximum length of the complete review up to 12,000 characters. If the reviewers do not agree on whether the test is eligible for review, or on the formal ratings, the Board decides.
8. The Board then sends the review to the authors of the tests, who are given a defined period of time within which they can state their position on the evaluation. The Board decides whether the reviewers will be asked to modify their review. The Board itself has the right to modify the review if it is not satisfied with any modification the reviewers make.
9. The reviews are then published in *Report Psychologie* and *Psychologische Rundschau* as well as in other journals that may have cooperated in the review process.
10. The authors are listed in alphabetical order unless they have made other arrangements. Each author has the right to remain anonymous.

CHANGES IN THE TEST REVIEW SYSTEM FROM 2006 TO 2009

When the first version was written, there seemed to be the prospect of the reviewers receiving a fee for their work (Testkuratorium, 2006). As it turned out, however, no funds were available for paying such fees. For this reason, this information was deleted in the second version (Testkuratorium, 2009).

The first version of the German test review system required at least one of the reviewers to have a license according to DIN 33430 (Testkuratorium, 2006). This license shows a "professional qualification in independently planning and carrying out job-related proficiency assessments according to DIN 3340." To obtain this license, an examination has to be passed, and practical experience has to be proven. This turned out to be an unrealistic demand, and was later changed (Testkuratorium, 2009). Such a license is typically held by psychologists working in the practical field of personnel selection. First, it was not the intention to exclude experts from other fields (e.g., clinical psychology). Second, demanding that an expert in personnel selection should take part in every review would seriously limit the number of potential reviewers. Third, it was likely that the test review system would be less accepted if such expertise in a specialist field was considered more important than expertise in any other field of psychology.

The first version intended the test to be reviewed by two persons (Testkuratorium, 2006). In several reviews, it turned out that the person addressed wanted to do the review together with a colleague. For this reason, the option of having a co-reviewer was permitted as an exception (Testkuratorium, 2009). In other words, a review can be written by at most four persons working in pairs.

SIMILARITIES AND DIFFERENCES BETWEEN THE GERMAN AND THE EUROPEAN AND DUTCH TEST REVIEW SYSTEMS

Before the German test review system was passed there had been a discussion in the German Board of Assessment and Testing about whether the systems of the EFPA or the COTAN (Commissie Testaangelegenheden Nederland, a commission of the Psychologists' Association of the Netherlands) were viable alternatives to Germany having a system of its own. All three systems are standardized in so far as they require the reviewers to cover central topics. All tests are rated in the same categories and can thus be compared easily.

The most important difference between the German test review system, on one hand, and the European and the Dutch test review systems, on the other hand, is that the German system has no regulation on how to evaluate criteria on the basis of the size of coefficients (other differences are described in the

paper "Internationalization of Test Reviewing" by Arne Evers, this issue). The test review systems of the EFPA (2008) and COTAN (Evers, Lucassen, Meijer, & Sijtsma, 2010) contain criteria for evaluating, for example, the size of reliability coefficients. DIN 33430 contains no comparable numerical criteria, nor does the German test review system. There were several reasons for this choice in the DIN 33430 standard and in the German test review system.

First, no system states reasons for its choice of the numbers, which are the thresholds for the rankings. Depending on the assessment situation in which a test is used (decision of different "importance" about individual persons or studies on group level), the COTAN system gives different ratings for the same reliability coefficient. This is plausible, but shows that fixed numbers are not adequate.

Second, reliability and validity coefficients themselves depend not only on the quality of the test but also on the sample. Artificially heterogeneous samples lead to higher reliability coefficients. Such coefficients would result in better ratings in the EFPA and the COTAN system, but they are neither realistic nor informative. Results of studies conducted with samples of typical tested persons have lower coefficients and result in less favorable ratings.

Third, it is desirable that the full range of coefficients (e.g., validity coefficients) that can be expected is reported. If several coefficients are determined, for example, in a validation study, only the most favorable coefficients may be reported. This improves the rating but does not give the test user all possible useful information. Criteria like "success" can be operationalized in many different ways. If the correlations of many scales of a test with these operationalizations are calculated, a test author gets the best results when only the highest coefficients are reported. Such a strategy would result in better ratings in cases where the size of validity coefficients is relevant.

The second and third argument are based on the reasoning that the main demand is not for the coefficients in a test manual to be as high as possible but rather to contain as much information as possible for the test user.

PRACTICAL MATTERS

Up to now, 11 reviews have been published and can be retrieved from http://www.bdp-verband.de/psychologie/testrezensionen/index.html. Just as many reviews are currently in different stages of the review process.

On average, it takes almost a year from the first letter to a potential reviewer until the review is completed. The process of publishing is very fast. *Report Psychologie* has 10 issues per year; *Psychologische Rundschau* has 4 issues per year. Once the publications have come out in print, the review becomes available online free of charge.

At present, the German Board of Assessment and Tests distributes responsibility for being the senior editor of the reviews to several members, whereas previously this task was carried out by a single member (two in succession). This development allows more tests to be reviewed within the same period of time.

Reviewers can choose to remain anonymous. But so far no reviewer has ever opted for anonymity, presumably because being named in the publication is the only tangible compensation that reviewers receive.

WHAT IS STILL NECESSARY

We need the cooperation of test publishers. At present, information on a test is sometimes scattered all over the test manual and thus is difficult to find. In some tests, information that is probably available to the test author is not presented in the test manual. For example, whenever the test-retest reliability of a test is determined, the effect of repeated test taking can be calculated. But many authors remain quiet about this effect. If test publishers required the authors to use the DIN Screen to indicate where the necessary information can be found in the manual, test reviewing would become much easier. In this way, the tests themselves might be improved in the process of being published: Test authors might become aware of which information is still missing and add at least the information that is available without conducting further studies. This would help all test users.

We need the cooperation of journals. Even in journals whose editors presumably share our interest in improving tests and testing by test reviewing, the German test review system has not been made the standard for reviews. According to the German test review system, reviewers are to be chosen by the Board of Assessment and Testing. Reviewers who evaluate manuscripts for journals are usually chosen by the editors of the journal. Even so, these reviewers may of course use the German test review system. Reviews might then become clearer and easier to compare, even if formal differences remain, such as longer or shorter texts due to limitations imposed by the journal, or if there is only a single reviewer.

We need the cooperation of experts in different fields—experts who are well informed about test construction and are willing and able to review tests in addition to their daily work and are motivated by the prospect of adding test reviews to their list of publications. In other words, we need to build up a tradition of voluntary work, with psychologists who have a more profound knowledge in test theory helping other psychologists—and persons of any profession—in the interests of the persons being tested. We are currently building up a pool of potential reviewers. However, one practical problem when choosing potential reviewers is the fact that

experts in a certain field have often either cooperated with the test author or are competitors because they themselves are authors of a test in the same field.

The more popular the German test review system becomes, the easier these remaining tasks might become.

NOTE

1. DIN means "Deutsches Institut für Normung" (German Institute for Standardization). The typical example of a standard for a product, which is known by most Germans, is DIN A4, a standard for the size of a sheet of paper. The DIN 33430 belongs the Services Standards, which are less known to the public.

REFERENCES

Bartram, D. (2001). Guidelines for test users: A review of national and international initiatives. *European Journal of Psychological Assessment*, *17*, 173–186.

Berufsverband Deutscher Psychologinnen und Psychologen (Ed.). (2005). *Ethische Richtlinien der Deutschen Gesellschaft für Psychologie e.V. und des Berufsverbands Deutscher Psychologinnen und Psychologen e.V.* [Ethical standards of the Deutsche Gesellschaft für Psychologie e.V. and the Berufsverband Deutscher Psychologinnen und Psychologen e.V.] Retrieved from http://www.bdp-verband.de/bdp/verband/ethik.shtml

DIN. (2002). *DIN 33430: Anforderungen an verfahren und deren einsatz bei berufsbezogenen eignungsbeurteilungen* [DIN 33430: Requirements for proficiency assessment procedures and their implementation]. Berlin, Germany: Beuth. English translation. Retrieved from http://www.bdp-verband.org/bdp/politik/clips/din33430en.pdf

European Federation of Psychologists' Associations (Ed.). (2008). *EFPA review model for the description and evaluation of psychological tests*. Retrieved from http://www.efpa.eu/download/9044bd41c7953b956876e06c797f8c9f

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit von tests* [COTAN review system for the quality of tests]. Retrieved from http://www.psynip.nl/cms/streambin.aspx?requestid=287E1703-37F7-4499-894C-E7482C38B714

Kersting, M. (2006). *"DIN SCREEN"–Leitfaden zur kontrolle und optimierung der qualität von verfahren und deren einsatz bei beruflichen eignungsbeurteilungen* [DIN Screen–guide line to control and improve the quality of instruments and their application in professional personnel selection]. Lengerich, Germany: Pabst.

Kersting, M. (2008). DIN Screen, Version 2. Leitfaden zur kontrolle und optimierung der qualität von verfahren und deren einsatz bei beruflichen eignungsbeurteilungen [Guideline for monitoring and optimizing the quality of proficiency assessment procedures and their implementation]. In M. Kersting (Ed.), *Qualität in der diagnostik und personalauswahl - der DIN Ansatz* (pp. 141–210). Göttingen, Germany: Hogrefe.

Kersting, M., Althoff, K., & Jäger, A. O. (2008). *Wilde-Intelligenztest 2* [Wilde Intelligence Test 2]. Göttingen, Germany: Hogrefe.

Testkuratorium. (2006). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen [TBS-TK. Test review system of the German Board of Testing of the Federation of German Psychologists' Associations]. *Report Psychologie*, *31*, 492–499.

Testkuratorium. (2009). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09, September 2009 [TBS-TK. Test review system of the German Board of Testing of the Federation of German Psychologists' Associations. Revised version, September 9, 2009]. *Report Psychologie*, *34*, 470–478.

Testkuratorium der Föderation deutscher Psychologenverbände. (1986). Beschreibung der einzelnen kriterien für die testbeurteilung [Description of the individual criteria for test reviewing]. *Diagnostica*, *32*, 358–360.