On the psychometric quality of new ability tests administered using the WWW

Oliver Wilhelm[1], Michael Witthöft[2], Andreas Größler[3], & Patrick McKnight[4]

Key words: Calibration, Equivalence, Validity, knowledge tests

Abstract:

Internet or ordinary ability testing methods differ with respect to the nature of the testing situation on several important factors (e.g., person mediated communication and subject behavioral control). Ordinary testing methods ought to be preferred in psychological experiments, however, there are important considerations for choosing alternative (i.e., internet) testing methods. The conditions that may warrant the use of internet testing include necessary resources, availability of experts and bridging geographical distances.

In some applied settings, e.g., personel selection, the goal of measurement is to select the best person from a pool of highly capable subjects. With ordinary testing methods there are frequently not enough experts available or the experts are too costly to assist in constructing, calibrating, evaluating, and validating measurement instruments. New knowledge tests were developed for internet distribution to make use of available experts and gather the largest sample possible for item evaluation and test construction. The new knowledge tests were developed in science and economy - two domains where, to our knowledge, no convenient instruments are available. Each domain had two parallel tests in two languages and used self selected samples participating via internet.

The quality of data was critically evaluated on several levels with different methods. In addition to procedures from classical test theory, probabilistic

1 Lehrstuhl Psychologie III, Universität Mannheim, D-68131 Mannheim, Phone ++49 (0) 621 181 2143, Fax, ++49 (0) 621 181 3997, email: wilhelm@tnt.psychologie.uni-mannheim.de

2 Lehrstuhl Psychologie II, Universität Mannheim, D-68131 Mannheim witthoeft@tnt.psychologie.uni-mannheim.de

3 Industrieseminar, Universität Mannheim, D-68131 Mannheim agroe@is.bwl.uni-mannheim.de

procedures were applied to the data. Both domain tests were evaluated for a priori internal structure. Specifically, difficulties and validities on an item level could be compared to expert ratings of difficulty and validity for the economics tests. In both domains data gained by internet administration could be compared with data gained using ordinary methods. In the science test the internal structure was tested for robustness using a quasi-experimental manipulation. Finally the equivalence of parallel tests was determined comparing relevant statistics. To preliminarily validate the new measurement instruments, biographical questions (e.g., education level, relevant prior knowledge, and proxy variables including reading of newspapers and journals) for the interest in the respective domains were used. The results of the analysis essentially support the supposition that internet administration yields quality data, hence, it seems appropriate to use the internet as a testing medium for the means outlined here. Future investigations include calibrating test results to real world behavior. Additionally, future efforts to vary the item format from traditional yes/no or multiple choice response choices will be tested along with an expansion of the knowledge structure.

Introduction:

Using the internet as a medium to apply psychometric tests seems to be an appealing method to gain information. This is evident with respect to the big variety of online recruiting sites in the world wide web. The information collected there ranges from ordinary biographic questions to personality questionnaires adapted for use via the WWW to questions intending to measure knowledge or ability. Decisions based on those data include preselections for employment interviews. Consequently, as with all other methods used to draw employment decisions, the utility and many more properties of the measurement instrument are to be investigated empirically. However despite the monetary important judgments the quality of the data is usually unknown.

While biographic information can usually be checked easily, it is harder to get a reliable picture on the abilities of an applicant, especially if cheating can not be excluded. Apart from the cheating and faking problems there are long standing debates in psychology about the equivalence of results gained by different methods (e.g. computerized versus paper pencil testing) (Maiwald & Conrad, 1993; Mead & Drasgow, 1993). It is hard to ensure the equivalence of data gained by independent subjects with distinct methods, the data gained with participants taking the tests via WWW seem to be less noisy and unreliable (Wilhelm, Witthöft, & Größler, 1999; Wilhelm & McKnight, in prep.).

Major arguments in favor of the testing situation offering less control are, that special groups of subjects could be accessible, the costs of research could be reduced

An additional advantage can be, that subjects with diverse cultural backgrounds, speaking different languages, could participate in the research and thereby increase the external validity of the conclusions drawn from the results (Reips, 1999).

The last point was of primary interest in the research to be reported here. Test translations and their quality are a complex problem. There are a number of aspects that have to be considered and a big variety of methods that can be applied (Ellis, 1993; Hambleton, 1993; Wilhelm & McKnight, in prep.). A major goal in cross cultural measurement is to measure the same ability, to tap he same latent construct in all cultures. It is possible to test that equivalence and to explore possible causes for deviations from equivalence, if the samples are easily comparable. If the samples however are not easily comparable there is an alternative explanation for deviations from equivalence: Its not the test translation that manipulates the properties of the test, its the respective sample. The harm done to equivalence can take a variety of forms. The groups could be distinguished for example only with respect to the overall level of performance achieved. However more severe violations of equivalence include biases in dispersions, item and test reliabilities and validities and finally the structure of the answer vectors given by persons of different groups (Rost, 1990; Steyer & Eid, 1993).

On the other side when new measurement instruments should be designed it is a good starting point to develop the new measurement instruments in the medium and in the languages intended for later use instead of translating readily developed measures. We decided to explore tests from two domains of human knowledge that, despite their significance, are widely neglected applied areas of psychology[5]: general business administration knowledge and general science knowledge. Both areas obviously play an essential role in a great variety of jobs but still there are no wide spread used or well known tests (in German language). An available general business administration test (Krumm & Seidel, 1970) is rather outdated. Besides this test only recently has a German adaptation of an American economics test (Soper & Walstad, 1987) been published (Beck & Krumm, 1999). Especially the latter of the two tests is essentially a test on economics rather then general business administration. In the area of science there are to our knowledge no published measurement instruments.

Methods

Materials and Procedure

We designed two general business administration method tests, that included items from the following nine content areas: general management, cost accounting, financial accounting, production management, finance, strategic management,

marketing, commercial law, and taxation. From a pool of 60 items two halves were divided randomly on two tests holding content domain and expert rated difficulty parallel across the two tests. Those two tests in German language were translated into English with adaptations made were necessary[6].

For the science test the direction of translation was inversed. Starting from two English tests, designed to be parallel, the respective German versions were generated. Problems with the language differences here centered around the use of Latin words in the German versions, because those words frequently have a German translation. In table 1 a sample item of both tests is given. The default answer in all test versions was “omitted”. Preceding the questions in all tests general and specific instructions were given and biographic questions were asked. Testing time was unlimited from our side and the use if auxiliary means was said to render the feedback useless. Besides the knowledge questions participants were asked to answer biographical questions including relevant prior education, consumption of corresponding media for the respective tests, private interest in the respective domains and more general questions like education, profession, age, and sex.

Table 1: Sample items for the business administration and the science test.

General Science Test
How are animals genetically engineered to produce proteins?
① A process alters protein synthesis via changes in DNA structure
② A process alters the amount of RNA, thus reducing proteins
③ A process alters protein synthesis via changes in DNA coding
④ A process alters protein synthesis via changes in gene coding
⑤ not answered (default)
Business administration test
The contribution margin is equal to the revenues less the variable cost
① True
② False
③ not answered (default)

Participants

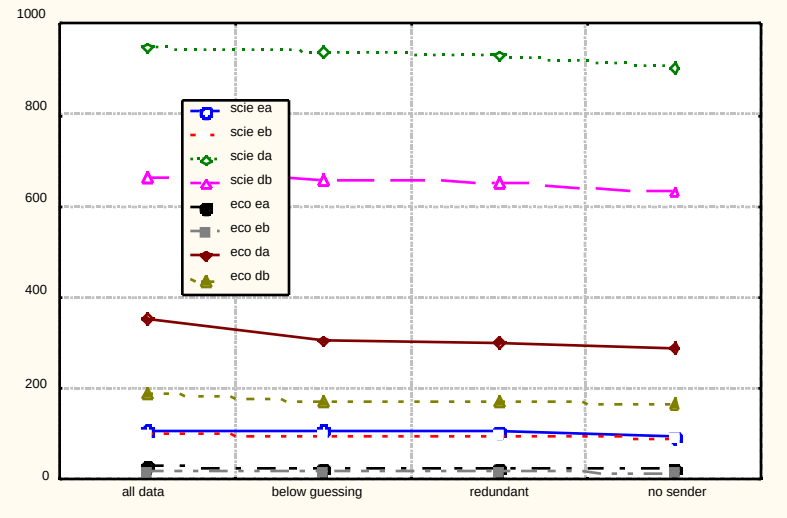
A total of 2217 participants contributed to this research. The only incentive given was feedback about performance in the tests, absolute and relative to other participants. Slightly above 20% of those subjects took two different tests of the same domain, enabling us to compute retest reliabilities. The educational and

6 We assume participants who have taken the English forms of all tests to come from a more diverse national and cultural background as the German speaking samples. Adaptations made to

professional background of the subjects is comparable to the mixtures found before (e.g. Wilhelm & McKnight, in prep., Musch & Klauer, in prep.).

The strictness of inclusion criteria had very little influence on the data. In figure 1 cumulative stricter criteria were applied to all tests. From analysing all data, over excluding all data sets with performance below the guessing probability and removing all data with equal email address and observations without valid sender the size of each group does not alter substantially. This is valid too for the means and standard deviations.

Figure 1: Alterations in N under variations of inclusion criteria



Results

Although the tests show comparable means and standard deviations across languages and test versions the alphas within the tests are very low (see table 2). For those subjects who took more then one test retest reliabilities were computed. The values are -.03 for the English science tests, .41 for the German science tests, and .56 for the German business administration tests. Opposed to the devastating values found for the science tests we found strong correlations between the item means across languages and with the values found in a paper-pencil study conducted with 85 seniors in high school that were about to start an accelerated science program in college (none of the intercorrelations is below.86 – while the intercorrelations of part whole corrected item test correlations range between -30 and .49). Due to the small group size of the participants in the English business administration test the same procedure could not be applied to the business knowledge test.

Table 2: Descriptive statistics for all test versions.

Test	N items	N subjects	m	sd	alpha
Science e a	18	95 (28)	10,7	1,9	,22
Science e b	11	91	8,9	2,4	,44
Science d a	18	904 (325)	10,3	2,1	,35

Economy e a	30	23 (2)	18,5	2,3	,17
Economy e b	30	17	18,8	2,4	.03
Economy d a	30	290 (94)	20,1	2,8	,27
Economy d b	30	167	22,0	3,7	,64

Legend: e=English test, d=German test, a=first parallel test, b=second parallel test, numbers in parentheses are the number of participants who took the second version of each test.

For brevity further analysis can only be presented for the German economy tests. The data from 94 persons who took both tests were subjected to a confirmatory factor analysis. In order two explore the internal structure of the two parallel tests (all analysis were restricted to the six scales with more then three items) a first measurement model contained only one general factor, on which all scales loaded. A second model allowed correlated errors between corresponding scales from the two tests. The difference between model one and two is not significant. Due to the already acceptable fit of the general factor model, it is a priori not likely that substantially correlated errors as indicators of specific variance on a group factor level can occur. For parsimony and for the better descriptive indices we accept the general factor model. Data with more items from the subdomains and a bigger sample might well find a distinction between different areas of business administration knowledge.

Table 3: Confirmatory models of the German business knowledge tests

Model	Chi^2	df	p	CFI	AGFI	RMSEA
1	57	54	,36	,96	,87	,027
2	53,5	48	,27	,93	,87	,037

Discussion

The analysis of the data collected so far showed a bunch of problems. The test data are not very reliable and the biographic questions have not demonstrated their usefulness in predicting the test scores. For the economics test the psychometric quality improved compared to previous versions (Wilhelm & McKnight, in prep.). Still the measure is not satisfying. We have shown elsewhere (Wilhelm, Witthöft, & Größler, 1999) that the psychometric problems can not be reduced to the internet as test administration medium for knowledge tests, because the distinguished psychometric properties of a computer knowledge test are an apparent counterexample to that hypothesis. Additionally the psychometric problems we found for the business administration test occur independent of test medium. The problems we observe are tied to the specific tests we use.

What then will be the future efforts we undertake to improve the tests currently under investigation? The business administration test obviously needs to broaden its

needs to be provided with a different answer format in order to decrease guessing probability and the associated noise in the data. Besides deepening our understanding of the internal structure of business knowledge by building up on the results of the measurement models will be a major aim. For the science knowledge test the role of the text sections seems debatable to us. Additionally we do not have strong prior theory for postulating an internal structure so far. Adding further items with a more general content could improve the internal consistency and move the measured dimension nearer to the well known crystallized intelligence. Ironically through attempts to ameliorate poor tests, it seems to us, that we can learn more about the WWW as test administration medium and its influences then by using tests showing satisfying properties.

References

Beck, K., & Krumm, V (1999). Wirtschaftskundlicher Bildungs-Test (WBT) [Economics Education Test]. Göttingen Hogrefe.

Ellis, B. B. (1993). A partial test of Hulin's psychometric theory of measurement equivalence in translated tests. European Journal of Psychological Assessment, 11, 184-193.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 57-68.

Krumm, V., & Seidel, G. (1970). Wirtschaftslehretest BWL [Business administration test]. Weinheim: Beltz.

Maiwald, J., & Conrad, W. (1993). Entwicklung und Evaluation des MTP-C: Mannheimer Test zur Erfassung des physikalisch-technischen Problemlösens als Computerversion [Development and evaluation of the MTP-C: Mannheim Test to measure physical-technical problem solving abilities computer assisted]. Diagnostica, 39, 352-367.

Mead, A., D., & Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. Psychological Bulletin, 114, 449-458.

Reips, U. (1999). The web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.) Psychology experiments on the internet. San Diego, CA: Academic Press.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. Applied Psychological Measurement, 14, 271-282.

Soper, J. C., &Walstad, W. B. (1987). Test of economic literacy (2nd edition): Examiners manual. New York: Joint Council on Economic Education.

Steyer, R., & Eid, M. (1993). Messen und Testen [Measuring and testing]. Berlin: Springer.

Wilhelm, O., Witthöft, M., & Grössler, A. (1999). Comparisons of paper-and-pencil

Mathey, A. Jaillet, & E. Nissen (Eds.) Proceedings of IN-TELE 98 (pp. 439-449), Berlin: Peter Lang.