Behavior Research Methods
2009, 41 (4), 1183-1189
doi:10.3758/BRM.41.4.1183

Comparison of ability tests administered online and in the laboratory

Jan Marten Ihme
University of Kiel, Kiel, Germany

and

Franziska Lemke, Kerstin Lieder, Franka Martin, Jonas C. Müller, and Sabine Schmidt
University of Jena, Jena, Germany

As the Internet, the global medium of the future, expands exponentially, it has become increasingly relevant to scientific research. So far, there is but little evidence that online testing is suitable for collecting ability-test data. The present article aims to shed light on some aspects of the issue by comparing the performance in a computer-administered ability test of one lab sample and two online samples using a quasi-experimental design. Mean score differences appeared, but can be explained by differences in age and education, and were not due to the test setting (online vs. laboratory). Also, there were no structural differences between the achievement scores of both samples. Some limitations on generalizability are discussed.

In the last decades, the Internet has become an increasingly popular medium. It plays a larger and larger role in our lives, which is one of the reasons it is presently at the heart of scientific research. The rising number of users visiting platforms like the Web Experimental Psychology Lab (Reips, 2001) recently clearly illustrates the growing interest in online research. Many diverse research questionnaires and ability tests can already be found online; it is important, however, to determine the quality, validity, and reliability of such data. Therefore, the aim of this study is to investigate whether data collected online and offline from a computer version of a matrices intelligence test are of comparable quality. This would be the case if the achievement scores of the participants were independent of the setting (online vs. offline).

Online Testing

From a researcher's perspective, online testing applied to wide fields of current scientific research is an attractive procedure, because it is a very economic strategy in all stages of testing (Klinck, 1998). Huge samples can be recruited with low expenditure (Buchanan & Smith, 1999b); therefore, the cost of such investigation is less than with conventional research procedures. Another crucial advantage of online testing is that large and heterogeneous as well as very particular communities of participants can be reached (Buchanan & Smith, 1999a). Consequently, it is possible to raise the power of a test in a dramatic way (Reips, 2000).

Even though the integration of the World-Wide Web into scientific research is highly desirable, some specialties of this powerful medium have to be taken into consideration. An important requirement for scientific research is the guarantee of identical testing conditions for all participants (Buchanan & Smith, 1999b); this might be difficult, considering the large variety of browsers and connections with different features that are frequently used. For instance, layout and downloading times for different participants could differ and prove difficult to control, resulting in distorted comparisons of participants' performance. Furthermore, compared with conventional computer testing, online testing is characterized by a lack of direct contact between participant and researcher, as well as by diminished experimental control. This raises the question of how to deal with data collected online, and to what extent such data can be applied to research. First, it is unclear whether the participants fully understand the instructions. In contrast to the laboratory situation, participants' effort, concentration, attention, and compliance cannot be easily assessed. Second, the motivation and test situation of the participants in the study remains unclear; for example, interference such as noise, listening to music, and so on are not controllable. On the one hand, it is assumed that motivation among online users is higher than it is among participants recruited for a classical research procedure in the laboratory because the online users are self-selected, and invest time and effort (Wilhelm & McKnight, 2002); on the other hand, some users might just glance through the test simply to acquire information about it, and make no attempt to solve it seriously. This could happen if a participant becomes frustrated with the difficulty of the

J. M. Ihme, ihme@ipn.uni-kiel.de

test. In summary, it can be stated that a lot of important variables, such as technical conditions and the situation of a participant, cannot be controlled by the experimenter using online tests.

Measuring General Mental Ability

According to Jensen (1998), the general component of cognitive activity is mental ability, which is not dependent on sensory or output features in any critical way. Figural matrices tasks (Raven tasks) are known to be the best single marker for general mental ability (Carroll, 1993; Jensen, 1998). Therefore, the assessment of a figural matrices test is a good opportunity to compare the quality of data collected online with that collected offline.

Raven items typically consist of a 3×3 matrix of figural elements, arranged according to one or two rules. The bottom right cell is left blank and the participant is asked to select the solution from eight alternatives. According to the processing theory of figural matrices (Carpenter, Just, & Shell, 1990), two steps of cognitive processing are necessary to solve such items: At first, the figural elements arranged by the same rule have to be identified. This process is called *correspondence finding*. Second, the participants have to test their hypothesis concerning different subproblems and store these results in working memory. This is called *goal management*. Working-memory load is affected by the number of rules, their complexity, and their distinctiveness (Carpenter et al., 1990; Embretson, 1998).

Raven items are suitable for rule-based item construction because of their definite structure; that is, by combining one or two rules, different modes of plotting, and directions of rule application, a virtually infinite repository of variably difficult items can be constructed.

Quality of Online Assessed Data

In the research literature there are a few studies dealing with the quality of online assessed data. In online tests, experimenters have less control over the setting and, because they are not physically present, should be more concerned about data quality; however, there are compensatory techniques (Buchanan, 2002; Reips, 2002). A study conducted by Bartram and Brown (2004) indicates that lack of supervision has little if any impact on scale scores. A further study demonstrates that differences in data in an online and a paper–pencil version of a test battery can be ascribed to the computerization process rather than to the lack of supervision or to the differences between online and offline modes (Coyne, Warszta, Beadle, & Sheehan, 2005). This research indicates that lack of supervision while a test is being taken should not have an impact on data quality.

A study by Preckel and Thiemann (2003) provides evidence for the comparability of the quality of online and offline collected data of a matrices intelligence test. Highly gifted participants completed either a paper–pencil version or an online version of a test consisting of twenty-six 4×4 matrices. The findings of the study support the assumption that valid and reliable data can be collected online (Preckel & Thiemann, 2003). The internal consistencies of both versions are comparable, and correlations with different criteria do not differ significantly. Mean differences in item difficulty can be explained by sampling effects (e.g., differences in dropout rates and prior experience with test taking). These encouraging findings could be observed despite differences in test media (paper–pencil vs. computer).

As already stressed by Preckel and Thiemann (2003), differences in characteristics of online and offline samples (e.g., age, gender, and education) can influence the comparability of data quality. Mean score differences between online and offline samples were found and could not be resolved. In many cases, differences in score distributions can be attributed to differences between samples tested online and offline (Buchanan, 2003); characteristics of samples should, therefore, be examined accurately and this issue should be carefully considered when the collected data are analyzed. Differences in data quality between online and offline samples have also been found in randomized experiments in which differences in the characteristics of the samples should not exist (Joinson, 1999). However, the tests applied in these studies were personality-related tests, or tests in which social desirability played an important role. In contrast, effects of social desirability should not arise in ability tests, in which the aim is to demonstrate maximal performance. Thus, the data quality of online ability tests is likely to be unaffected by social desirability and, therefore, easier to maintain.

To ensure appropriate data quality, a number of techniques can be used. Examples proposed by Reips (2002) are warm-ups, subsampling, multiple site entry, controlling for multiple submissions, controlling for motivational confounding, providing contract information, and dropout-reducing design.

Hypotheses

The data quality of online and offline collected data of a matrices intelligence test is equally high. Differences in test performance among subsamples are due to demographic distinctions such as age, gender, educational level, time taken to complete the test, concentration, test experience, and expectation.

The hypotheses were chosen for various reasons: First, if an ability test is applied, effects such as social desirability involved in personality testing, for example, should not occur. Therefore, fewer problems have to be taken into consideration when applying ability tests. Second, if both the offline and the online version of a test are presented via the same medium (computer), differences in data quality, which can arise through differences in test media, should not play a significant role. Third, if differences between subsamples, sometimes responsible for differences in test performance, are included in statistical analysis, comparability of data collected online and offline can be assessed more accurately. Fourth, if several techniques commonly used to ensure data quality of online tests (e.g., proposed by Reips, 2002) are considered while a test is being constructed, comparability of online and offline samples should be possible.

METHOD

Design and Recruitment

The study was designed as a quasi-experiment. Reips (2001) stated that external validity of questionnaires is higher when a participant can choose where to take the test—at home or in a test lab, for example. External and internal validity have to be balanced. The aim of this study was to compare online and lab testing with an ability test, so samples were chosen that could easily be drawn but could be assumed to be typical of each method. This made a quasi-experiment preferable to a randomized experiment.

A laboratory sample and an online sample were compared. The lab sample was recruited at the University of Jena. To enroll psychology students for the lab sample, lists were handed out in several lectures. The undergraduates volunteered in exchange for course credit. The online sample was recruited in two ways: First, in order to obtain participants comparable to the psychology students of the lab sample, participants for the online sample were recruited from mailing lists of psychology students at the universities in Cologne, Düsseldorf, Greifswald, Kiel, and Mannheim, who found the psychology students' mailing list sample (pml sample). Second, in order to recruit additional participants usually found in online samples, a standardized invitation to participate in an online study, and the test link, were posted in thematically appropriate forums (forums sample). All samples participated voluntarily. The testing period for all samples lasted from November 2006 to January 2007. This research design makes it possible to clarify whether mean differences in sum scores can be ascribed to demographical distinctions of participants or to type of setting (online vs. offline).

Participants

Altogether, 698 participants were assessed with the first task of the ability test; 490 (70.2%) of them completed the whole test and claimed to have seriously worked on it. In detail, 220 of the 364 (60.4%) in the forums sample, 213 of the 276 (77.2%) in the pml sample, and 57 of the 58 (98.3%) in the lab sample fulfilled these criteria. Of these, 68% were women, who were more likely to work on the test once started, as shown by the fact that only 53% of the participants who started the test but did not finish it were female. This sex-specific dropout rate might be due to reaction against forced responses (Stieger, Reips, & Voracek, 2007). There was no evidence for further demographical differences between participants and dropouts. Four hundred eighty-one of the 490 data sets (98.2%) were included because these participants took at least 6 min to solve the test. This time appeared to be necessary for serious work on the test, given the task and the distribution of it. It was determined from the coded IP address and the personal code of each participant that there had been no multiple submissions. Table 1 shows the descriptive statistics of the three subsamples.

Lab sample. Fifty-seven participants (13 male, 44 female) who completed the test were on average 21.7 years old (age range, 18–46 years; $SD = 5.4$). As for educational levels and current occupations, the sample was quite homogeneous. The majority of the participants were students who had obtained the Abitur (German high school diploma) as their highest formal qualification.

Pml sample. Two hundred twelve participants (55 male, 157 female) who completed the test were on average 24.7 years old (age range, 18–50 years; $SD = 5.4$). The sample showed a low variation of educational level and current occupation. The majority of the participants were students with the Abitur as the highest formal qualification.

Forums sample. Two hundred twelve participants (84 male, 128 female) who completed the test were on average 27.0 years old (age range, 18–56 years; $SD = 8.9$). The variation of educational level and current occupation was relatively high for this subsample.

Table 1
Descriptive Statistics of the Three Subsamples

Sample	*N*	Female	Age		With Abitur
			M	*SD*	
Lab	57	44 (77.2%)	21.7	5.4	91.2%
Pml	212	157 (74.1%)	24.7	5.4	87.3%
Forums	212	128 (60.4%)	27.0	8.9	48.5%

Note—Abitur, German high school diploma.

Procedure

The test was programmed in PHP, and all participants' input was stored in a MySQL database. The test was almost identical for all three samples, with the exception of slight differences in detail and a different hyperlink referring to the test for each sample. (The test was located at the Web server www.uni-jena.de.) The time, date, and encoded IP address were stored at the start of each test run. At the beginning, a written introduction was presented to the participants, stating that the upcoming test, intended to measure intelligence, would take 20 to 40 min to complete. Furthermore, some instructions concerning optimal conditions for solving the test (e.g., avoiding any kind of distraction) were given. Demographical data (e.g., gender, age, level of education, current occupation, and state of origin) were collected, and a personal code was registered for each participant. Afterward, participants solved the intelligence test. To prevent them from answering items by mistake, no item could be accessed until the previous one had been answered. After completion of the test, participants answered questions about their test experience and their reported concentration while taking the test. This was important, because these factors can evoke differences in performance among the three subsamples. Participants' e-mail addresses were registered and saved separately, so that individual data could not be identified and anonymity was guaranteed. After having answered all questions, each participant was given feedback consisting of the number of properly solved items and a comparison with participants who had already taken part in the study. In addition, it was pointed out that test performance could depend on several factors such as current mood or distraction while working on the test. After the test, the participants were thanked for taking part.

Such immediate feedback could have led to repeated participation (Reips, 2002), but this was controlled by saving the coded IP addresses: Data sets with similar content coming from the same IP address could be identified without endangering anonymity. The decision to give honest feedback to all participants, including those whose scores were comparatively low, was made carefully. It was found more important to give feedback to all participants than to protect those who scored low. Some hints for self-esteem maintaining attribution were given if a participant was having a bad day.

Despite the testing procedure being the same for all participants, specific situational settings and degrees of concentration on the test could have varied for different participants; this could threaten the comparability of data of the different subsamples. In the laboratory, the working environment was controlled by the experimenter, so that optimal conditions for filling in the test were provided. The test was carried out in a laboratory at the University of Jena. One to 3 participants were tested simultaneously. Participants were allocated their workstations. All further instructions were presented on the monitor. In the presence of the experimenter, participants carried out the test. When finished, participants were credited for having taken part in the study. In contrast to the lab sample, the conditions under which the participants from the two online samples carried out the test online were not controllable. Several items were used to judge the influence of reported concentration, test experience, and expectation on test performance (e.g., "Did the test meet your expectations?").

Materials

Intelligence was measured by a figural matrices test with 22 items chosen from a larger pool of rule-based self-constructed items designed with the program ITEMGENERATOR (Ihme, 2007).

These items were used instead of an established measure because the access to the test was not limited, so it was not possible to use a copyright-protected test. Computer-based construction allowed items to be designed specifically for this purpose, without concerns of copyright and item protection. The level of difficulty of the items varied widely. For each item, participants had to choose between 8 different possible answers: the correct solution and 7 distractors. Selecting 1 of the alternatives was always necessary in order to proceed to the next item. The time the participants took to complete the test was registered. The number of correctly solved items served as the dependent measure.

RESULTS

Item Selection

Since the matrices test used in this study had not yet been evaluated, item analyses were conducted first. Three items were excluded because more than 95% of the sample solved them correctly. Another item was excluded because less than 12.5% solved it, making its relative frequency less than the guessing probability.

To avoid artificial difficulty factors, it is always necessary to consider the categorical structure of data when dealing with dichotomous items. Factor analysis for categorical data uses polychoric correlations instead of Pearson correlations; this provides a better reflection of the true relation. An exploratory factor analysis for categorical data was conducted with the remaining 18 items using Mplus (Muthén & Muthén, 1998). As expected, the scree plot criterion affirmed the one-factor solution. Three items were excluded because of standardized factor loadings of less than .3 on the general factor. All subsequent analyses were conducted with the remaining 15 items.

Dimensionality

An item response theory (IRT) Rasch model was specified with the remaining 15 items using MULTIRA (Rost & Carstensen, 1998). The program allows computing a sufficient value for test performance simply by adding up all item values. The fit of Rasch models can be tested by simulating data with the same model parameters and comparing the data thus obtained with the original data, a procedure called *bootstrapping* (von Davier, 1997). The Cressie–Read statistic CR(2/3) of the original data set valued 7.771, the mean of 999 bootstrap samples valued 6.554. The rank of the original data set was 893, so the p value for an equal or better fit of the real data was $p = .11$. The data fit a Rasch model, which means that the number of correct answers for the 15 items can be used as the indicator for test performance (dependent variable).

Table 2
Descriptive Comparison of the Three Samples

Sample	*N*	*M*	*SD*
Lab	57	11.02	2.30
Pml	212	11.18	2.48
Forums	212	9.71	2.94

Comparison of Samples

Descriptive comparison. Table 2 illustrates the descriptive statistics for the total scores of the three samples. The pml sample achieved the best overall result with a mean score of 11.18 points. The forums sample had the most difficulties solving the items. Participants in this sample a person achieved, on average, 9.71 points. With an average score of 11.02 points, the lab sample is in between.

At first, only the two samples consisting of students (i.e., the lab sample and the pml sample) were compared. Variance homogeneity is given [$F(1,267) = 0.63$, $p = .427$]. The overall means of the lab sample and the pml sample were not significantly different [$F(1,267) = 0.21$, $p = .648$, $\eta^2 = .001$]. Inserting age, gender, and education in a first step, and test experience, expectation, concentration, and time taken to complete the test in a second, revealed no hidden effects of the setting. Age, time taken to complete the test, concentration, and experience all had an effect on test scores, but gender, education, and expectation did not (see Table 3). Time taken to complete the test, concentration, and experience all had a positive effect on the test score, but the effect of age was negative.

Adjusted comparison. In a second analysis, the lab sample was compared with both online samples, the pml and the forums sample. Again, the setting (offline vs. online) served as the independent variable; pml and forums samples were, therefore, combined into one sample. The overall test score was the dependent variable; age, gender, and education served as covariates. Education was dichotomized (whether the Abitur had been obtained or not) to enter the analysis as a covariate. The variances of errors were not heterogeneous [$F(1,479) = 2.78$, $p = .096$]. The overall model turned out significant [$F(4,476) = 5.59$,

Table 3
ANCOVA for Comparison of Test Scores Between Lab Sample and Pml Sample

	Model 1					Model 2				
Covariate	*F*	df_1	df_2	*p*	η^2	*F*	df_1	df_2	*p*	η^2
Model	1.48	4	264	.210	.022	8.75	8	260	<.001	.212
Setting	0.94	1	264	.334	.004	0.231	1	260	.631	.001
Education	0.12	1	264	.726	.000	0.03	1	260	.861	.000
Age	4.07	1	264	.045	.015	7.72	1	260	.006	.029
Gender	0.28	1	264	.596	.001	2.85	1	260	.093	.011
Time						40.37	1	260	<.001	.134
Concentration						10.15	1	260	.002	.038
Test experience						4.72	1	260	.031	.018
Expectation						0.61	1	260	.434	.002

Table 4
ANCOVA for Comparison of Test Scores Between All Three Subsamples

	Model 1					Model 2				
Covariate	F	df_1	df_2	p	η^2	F	df_1	df_2	p	η^2
Model	5.59	4	476	<.001	.045	15.72	8	472	<.001	.210
Setting	0.53	1	476	.458	.001	0.98	1	472	.322	.002
Education	10.38	1	476	.001	.021	5.73	1	472	.017	.012
Age	4.08	1	476	.044	.008	9.32	1	472	.002	.019
Gender	1.30	1	476	.254	.003	0.58	1	472	.446	.001
Time						63.09	1	472	<.001	.118
Concentration						15.70	1	472	<.001	.032
Test experience						12.25	1	472	.001	.025
Expectation						0.08	1	472	.778	.000

$p < .001$]. Education and age had an effect on the test score; setting and gender had none.

Four additional covariates were included: concentration, test experience, expectation, and time taken to complete the test. The overall model did not change much [$F(8,472) = 15.73, p < .001$]. Education and age still had an effect on the test score, but gender did not. In addition, time working on the test, concentration, and test experience all had an effect, but expectation did not (see Table 4 for both models).

Comparison of IRT Models

To examine the structural equivalence of the test in both lab and online samples, two separate IRT models were estimated and the item parameters were compared. The correlation between the item parameter estimations was .95, so the item parameters were basically the same in both conditions. In both conditions, bootstrap tests of the Rasch model support the data fit to this model. Table 5 shows the comparisons among the correlations of sum score with age, gender, time needed to complete the test, test experience, and concentration in both conditions. Education was left out of this analysis because of its very low variance in the lab sample. None of these correlations differed between the conditions.

DISCUSSION

The aim of this study was to investigate whether online and offline data collected from a computer version of a matrices intelligence test are of comparable quality if certain issues are taken into consideration. Since online testing is an attractive procedure that combines many advantages, such as low costs and the opportunity to recruit large samples, more and more researchers are interested in integrating this new option into their research; therefore, this raises the question of the comparability of data collected online with that of data collected conventionally in laboratory settings. The advantages of online research are only worthwhile if no significant losses in data quality are involved in the application of this procedure; that is why research is needed to investigate whether data collected online and offline have the same quality, or whether certain aspects have to be taken into consideration in order for comparability to be assumed.

To find out whether it is possible to apply a matrices intelligence test online without losses in data quality, two online samples were collected, one that consisted of psychology students, and so was very similar to the laboratory sample, and another one that consisted of forum participants and therefore represented a group easily reached via the Internet, to improve the ecological validity. The results of the study show that there are indeed no differences when only those two samples consisting of students are compared. Comparing all three subsamples, the model adjusted for demographic variables showed no significant effect of setting (online vs. offline); the effect size was very small. Inserting into the model the covariates—time needed to complete the test, test experience, expectation, and concentration—the overall model effect size increased, but the effect of setting decreased. In detail, differences in education and age between the offline and online samples led to differences in test performance. The time needed to complete the test, test experience, and concentration explain additional variance of the test performance, but these covariates had no effect on the relation between setting and test performance. Beyond an equivalence of corrected means, structural equivalence was found. The estimation of Rasch model parameters led to similar results, and the identity of correlations of demographical and test variables with the test score supports the assumption of equal validity under both laboratory and online-testing conditions. This means that data quality achieved by online testing can be comparable to the quality of offline-assessed data, although it should be noted that the demographical structure of online samples can differ from the structure of offline samples acquired

Table 5
Correlations With Test Sum Score Compared Between Both Conditions

Correlation of Sum Score	Laboratory Sample	Online Sample	Fisher Z
Age	−.219	−.120	−0.706
Gender	.151	−.063	1.489
Time taken to complete the test	.141	.350	−1.546
Expectation on test	−.318	−.083	−1.703
Concentration	.419	.177	1.851

Note—The Fisher Z statistic is normally distributed; an absolute value of 1.96 would indicate that two correlations are significantly different.

by conventional procedures. This issue should be taken into consideration when analyzing online data.

Some limitations of this study should be discussed, too. Due to the use of self-constructed, nonevaluated data, some items had to be excluded. However, this does not affect the validity of the other items. The 15 chosen items all load on the same factor. The number of correct answers for the 15 items constitutes a sufficient statistic of test performance, because the 1 PL model fits the data well.

The test applied includes some special features that could limit the generalizability of the results of this study. First, the test is a power test, not a speed test; that is, there are no time restrictions. The measurement of time taken to complete the online test, or using items with time restrictions, imply difficulties, because the test situation is not controllable under these circumstances. In addition, technical reasons might restrict time measurement, so it might be more difficult to reach high data quality in online applied speed tests. Second, the test does not contain any verbal items that could have been identified with a simple Internet search. Cheating at such an item in the online setting might influence data quality, because test performance is not necessarily attributable to a person's ability. Third, the test was not used for aptitude testing, so invalid data sets were excluded; this would not be possible when testing individuals, each collected data set would have to be examined. Fourth, the test performance had no consequences for the participants, so the motivation to cheat, and consequently reduce data quality, was low. Since participants had neither the possibility nor the motivation to cheat in the applied matrices intelligence test, specific sources influencing data quality did not have to be taken into consideration. Online testing in other contexts might be connected to other problems that did not play a role in this study, so whether data collected online contains sufficient quality, despite additional uncertainties in other contexts, is an issue to be examined further.

Some information regarding dropout effects is also remarkable. Altogether, 66.3% of all participants of both online samples, and 98.3% of the participants in the offline sample, were included in the examinations, having filled in the test seriously and having spent at least 6 min on the test. However, 33.7% of the participants in the online samples did not finish the test or did not work seriously on it. Since failing completion of a test indicates an absence of motivation, or too much stress, the effects of self-selection, especially in the online samples, are evident here. Despite the dropout rates, online samples are still more variable than are many conventionally recruited samples in psychological research. The forums sample, especially, has a nearly balanced ratio of gender distribution and a wide age spread. Hence, online samples are a good alternative to the commonly used samples in psychological research, often consisting of 20-year-old female psychology students participating for course credit.

In conclusion, online ability testing can be highly recommended without reservations concerning data quality, as long as some conditions, such as testing without time restrictions and preventing participants from cheating, are met. When an online test is being constructed, techniques to ensure data quality should be employed. Furthermore, demographical characteristics of online samples, which can differ from characteristics of offline samples, should be kept in mind and considered when online data is analyzed. For future research, online samples are a good additional possibility for data collection.

Future research can concentrate on further validation of the applied matrices intelligence test—for example, by including external criteria such as math grades or final-exam grades. This would emphasize the findings of this study, if they were replicated in a randomized experiment in which every participant were in the online or lab sample by chance. Moreover, it would be interesting to find out whether the data quality of tests containing additional characteristics likely to influence data quality, such as time restrictions, or tests with knowledge tasks, can be compared with the data quality of offline versions of the same tests.

AUTHOR NOTE

Correspondence concerning this article should be addressed to J. M. Ihme, Leibniz Institute for Science Education (IPN), Olshausenstrasse 62, 24098 Kiel, Germany (e-mail: ihme@ipn.uni-kiel.de).

REFERENCES

Bartram, D., & Brown, A. (2004). Online testing: Mode of administration and the stability of OPQ 32i scores. *International Journal of Selection & Assessment*, **12**, 278-284. doi:10.1111/j.0965-075X.2004.282_1.x

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research & Practice*, **33**, 148-154. doi:10.1037/0735-7028.33.2.148

Buchanan, T. (2003). Internet based questionnaire assessment: Appropriate use in clinical contexts. *Cognitive Behavior Therapy*, **32**, 100-109. doi:10.1080/16506070310000957

Buchanan, T., & Smith, J. L. (1999a). Research on the Internet: Validation of a World-Wide Web mediated personality scale. *Behavior Research Methods, Instruments, & Computers*, **31**, 565-571.

Buchanan, T., & Smith, J. L. (1999b). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, **90**, 125-144. doi:10.1348/000712699161189

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, **97**, 404-431. doi:10.1037/0033-295X.97.3.404

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Coyne, I., Warszta, T., Beadle, S., & Sheehan, N. (2005). The impact of mode of administration on the equivalence of a test battery: A quasi-experimental design. *International Journal of Selection & Assessment*, **13**, 220-224. doi:10.1111/j.1468-2389.2005.00318.x

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, **3**, 380-396. doi:10.1037/1082-989X.3.3.380

Ihme, J. M. (2007). Ergänzung des Itempools des Matrizentests der adaptiven eignungsdiagnostischen Testung [Extension of the item pool of the matrices test of the adaptive aptitude test]. *Untersuchungen des Psychologischen Dienstes der Bundeswehr*, **42**, 93-110.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers*, **31**, 433-438.

Klinck, D. (1998). Papier-Bleistift versus computerunterstützte Administration kognitiver Fähigkeitstests: Eine Studie zur Äquivalenzfrage

[Paper–pencil vs. computerized administration of cognitive ability test: A study of the equivalence question]. *Diagnostica*, **44**, 61-70.

MUTHÉN, L. K., & MUTHÉN, B. O. (1998). *Mplus: The comprehensive modeling program for applied researchers user's guide*. Los Angeles: Muthén & Muthén.

PRECKEL, F., & THIEMANN, H. (2003). Online- vs. paper–pencil version of a high potential intelligence test. *Swiss Journal of Psychology*, **62**, 131-138. doi:10.1024/1421-0185.62.2.131

REIPS, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. M. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89-118). San Diego: Academic Press.

REIPS, U.-D. (2001). The Web Experimental Psychology Lab: Five years of data collection on the Internet. *Behavior Research Methods, Instruments, & Computers*, **33**, 201-211.

REIPS, U.-D. (2002). Internet-based psychological experimenting: Five dos and five don'ts. *Social Science Computer Review*, **20**, 241-249. doi:10.1177/08939302020003002

ROST, J., & CARSTENSEN, C. H. (1998). MULTIRA [Computer program]. Kiel: IPN–Institute for Science Education.

STIEGER, S., REIPS, U.-D., & VORACEK, M. (2007). Forced-response in online surveys: Bias from reactance and an increase in sex-specific dropout. *Journal of the American Society for Information Science & Technology*, **58**, 1653-1660.

VON DAVIER, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data—Results of a Monte Carlo study. *Methods of Psychological Research Online*, **2**, 29-48.

WILHELM, O., & MCKNIGHT, P. E. (2002). Ability and achievement testing on the World Wide Web. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 151-180). Ashland, OH: Hogrefe & Huber.

(Manuscript received August 7, 2008;
revision accepted for publication May 7, 2009.)