International Journal of Internet Science
2014, 9 (1), 52–63 ISSN 1662-5544

International Journal of Internet Science

IJIS.NET

Method Effects on Assessing Equivalence of Online and Offline Administration of a Cognitive Measure: The Exchange Test

Michael Schreiner[1], Siegbert Reiss[2], Karl Schweizer[2]

[1]University of Education, Heidelberg, Germany;
[2]Goethe University Frankfurt, Frankfurt a. M., Germany

Abstract: Internet administration of established offline measures has become more and more common over the past years. Research has shown that online administration of a measure does not necessarily change its psychometric properties but one cannot simply assume that online versions of any measure are equivalent to offline counterparts. In a counterbalanced test-retest-design the web-enabled version of the Exchange Test was administered offline and online. Data were analyzed by means of repeated measurement ANOVA. Results indicated that both administration forms were parallel at the first measurement. Hence, online and offline administration seemed equivalent. However, data from the second measurement show they were not: Analyses revealed a main effect of repeated measurement and a systematic interaction of "repeated measurement" and "order of administration" (online-offline vs. offline-online). Comparable articles often report analyses based on one measurement only or aggregate data from repeated measurements. The present study shows by counterbalanced test-retest-design that there is the possibility of method effects that can only be detected in a cross-mode test-retest design and with appropriate analyses. Potential reasons for the significant differences between offline and online administration at the second measurement are discussed and theoretical explanation approaches are presented that may guide further research into the observed lack of equivalence.

Keywords: Equivalence, Internet, online, working memory, counterbalanced, method-effects

Introduction

Computer technology has not only changed our everyday life completely, but also influenced the administration of psychological assessment. Originally, psychological assessment took place in a controlled and standardized face-to-face setting. Today we are able to collect data via Internet and we don't usually see face to face the people who contribute to our data sets. These "online laboratories" have 24h access (Reips, 2002); hence there is almost no limitation to where and when someone participates in an online experiment or questionnaire. This provides us with access to a large participant pool and in turn increases the generalizability of results and the power of statistical tests dramatically (Reips, 2000).

There are a great number of potential advantages and disadvantages connected with the online administration of a measure (see Birnbaum, 2004; Reips, 2000). One of the more apparent disadvantages of Internet research we have to face is the impaired standardization and controllability: Just a few years ago the idea of someone taking a test online was composed of a person sitting at home in front of their own personal computer. Today,

Address correspondence to M. Schreiner, Department of Psychology, University of Education, Keplerstraße 87, 69120 Heidelberg, Germany, Phone: (+49) 6221 477 629, Schreiner@ph-heidelberg.de

participants using a wireless Internet-connection on a mobile device may sit in their doctor's waiting room or lie at the beach–and more changes are to be expected as technology will continue to progress.

Honaker (1988) focused on test takers' subjective perception of test situations and stated that the perception of dramatically different test-taking situations might result in non-equivalent results. The example set above makes it clear that there are dramatic differences to a controlled laboratory environment. With respect to situational differences, we have to prove equivalence of any measure originally administered in a face to face / laboratory setting (offline) with its web administered equivalent (online). In contrast, after reviewing the literature addressing the question of equivalence of online measures with their offline versions, Epstein and Klinkenberg (2001) came to the broadly shared belief, that "new Internet-based investigations need not be so focused on demonstrating the equivalency" "because research has demonstrated that translating a measure into a computerized format does not necessarily change its reliability and validity" (p. 310). This statement obviously still refers to the controlled laboratory where a computerized version of a measure is tested against the paper and pencil counterpart in a standardized fashion. With the step to the Internet, a range of sources of disturbances one usually controls in laboratories needs to be considered. Acoustic and visual distractions are simply not controllable in Internet administrations. "Much of the situation cannot be controlled in Web experiments" concluded Reips (2000) and especially in the context of speed tests and cognitively demanding measures, these disturbances are likely to divert the participant's attention.

Equivalence

Nonetheless, it is "a fundamental assumption of Internet (online) research […] that the results obtained are comparable to in-person (offline) research" (Meyerson & Tyron, 2003, p. 614). After reviewing the literature, they came to the conclusion that "investigators have taken different approaches to this matter, but none have specifically examined psychometric equivalence" (p. 614).

Psychometric equivalence of two forms of a test is given when both forms are parallel (Honaker, 1988). The International Test Commission (2005) defined requirements for establishing parallel forms: (1) means should be equal, (2) variances should be equal, and (3) both forms should correlate to the same degree with external criteria (see also Allen & Yen, 1979; Ghiselli, Campbell, & Zedeck, 1981). In 1986 the American Psychological Association defined two slightly different requirements for considering forms being parallel: (a) the rank orders of scores of individuals tested in alternative forms should closely approximate each other, and (b) the score distributions should be approximately the same.

If two forms of a measure have different means, they can be made parallel by adding a constant. Forms with different variances can be transformed into being parallel by an equi-percentile transformation (see Meyerson & Tyron, 2003, p. 615). Since offline instruments do not necessarily maintain their psychometric properties when administered online, the documentation of potential differences in psychometric properties is an essential requirement in establishing an online version of an offline measure.

Buchanan et al. (2005) failed to demonstrate equivalence of the online and the paper-and-pencil version of the Prospective Memory Questionnaire. In an online study, half of the measure's subscales proved to be "essentially meaningless" (p. 148). Based on this result, they take the examination of an online test's psychometric properties as a matter of course. Barak and Cohen (2002) administered a career assessment online and offline and found differences between administration forms for three of the six subscales of the measure. Ployhart, Weekley, Holtz, and Kemp (2002) investigated by means of multiple group confirmatory factor analysis whether an Internet administration of a Big Five-type personality inventory yielded comparable psychometric characteristics to the offline version. The means of the online version were lower than the means of the offline version. Additionally, more severe, factor loadings were not comparable across administration formats. Klinck (1998) administered a computerized version of a cognitive measure called Performance Testing System (Leistungsprüfsystem, LPS; Horn, 1983). In comparing means and correlations between offline and online administration, "doubtful medium-of-administration effects" were found. By comparing data of a problem-solving task collected in laboratory and via Internet, Dandurand, Shultz and Onishi (2008) found online participants to be significantly less accurate.

In the case of minor differences between means of offline and online administration, some researchers claim demographic differences between subsamples tested offline and online (e.g. Buchanan, 2003; Ihmke et al., 2009) to be the source of these differences.

Study designs and effects of order

Another potential source of systematic variation in studies assessing equivalence is an effect of order. Data of studies, in which a sample group first completes an offline and later an online version of a measure–without

administering the same tests counterbalanced in reverse order to a second group–is likely to be impaired by effects of order. Therefore, Klinck (1998, 2002) claims a counterbalanced test-retest-design to be the approach of choice in assessing equivalence since this design enables the systematic disclosure of effects of order. (Additionally, this design allows the comparison of rank orders of individuals tested in offline and online administrations.)

A fair amount of study includes a counterbalanced design (e.g. Buchan, DeAngelis, & Levinson, 2005; Hays & McCallum, 2005; Pedersen et al., 2012), however, the studies often do not report explicitly on the detailed analysis of effects on data caused by study design. For example, the authors named above did use a counterbalanced study design, but they seem to have aggregated data by administration form for the purpose of statistical investigation. This procedure follows the assumption that potential differences between sub groups annul each other as a result of an aggregation. Possible effects of the study design can't be revealed by such analysis; moreover they are masked by the aggregation of discrete data.

The Exchange Test goes online

The present paper is about the comparability of data collected offline and online by a measure of working memory capacity: the Exchange Test (Schweizer, 1996). Initially, in the 1990s, the Exchange Test was run in DOS mode in laboratories. In order to be able to administer this measure via the Internet and in a more contemporary design, a web-enabled version of the Exchange Test with a more appealing user interface was programmed. The conduction of this ability test on the Internet was mainly guided by a step-by-step-guide provided in *Internet-based ability testing: Problems and opportunities* (Schroeders, Schipolowski, & Wilhelm, 2009). Schreiner, Altmeyer, and Schweizer (2012) administered this web-enabled version in a controlled laboratory setting (offline). The investigation of consistency indicated acceptable to good quality, comparable means were found and an internal structure according to theory was demonstrated. Most importantly, as in the previous DOS version, the online versions' scores proved to be highly correlated with fluid intelligence. A detailed description of the Exchange Test's web-enabled version can be found in Schreiner et al. (2012) and the online version is available here[1]. For a better understanding of the measure in the context of this article, a brief introduction is given in this article's method section.

As cited above, Schreiner et al. (2012) already proved the web-enabled version to be a close match of the DOS version. However, a limitation of their work is the offline administration of an online measure in a laboratory. They controlled the test environment and are therefore unable to judge the actual equivalence of the Exchange Test's offline and online administrations in the sense of an Internet administration. They explain their approach with the risk of confounded effects: if the investigation of the web-enabled version had failed in an online administration, they would not have been able to name a specific determining factor, doubtlessly (see Schreiner et al., 2012). Therefore, given non-equivalence, possible alternative reasons would have been the characteristics of the online version, characteristics of the sample, or interferences due to the lack of standardization and of controllability. They close their discussion with the sequacious advice to check equivalence by administering the web-enabled version online.

Aims of the present study

This paper is about an experimental approach to clarify the unanswered question of equivalence of offline and online administration of the Exchange Test. In other words: in the sense of parallel forms, psychometric properties are expected to be comparable across the administration forms "Laboratory" and "Internet". (1) Means and variances of offline and online administration are expected to be equal, (2) score distributions of data collected in the laboratory should correspond to those collected via Internet, (3) rank orders of participants tested offline and online should approximate each other, and (4) both administration forms should correlate to the same degree with external criteria.

Methods and Materials

Exchange Test

The Exchange Test is a measure of working memory capacity that requires the reordering of four symbols until the sequence of these symbols corresponds to the sequence of another given reference list composed of the same symbols. Only adjacent symbols are allowed to be exchanged stepwise. Participants have to perform this reordering of adjacent symbols mentally and have to count the number of exchanges required to attune both lists. Figure 1 provides an illustration of one item.

[1]URL to Exchange Test: http://www.exchangetest.uni-frankfurt.de

Figure 1. Stimuli of an Exchange Test item of the 5th treatment level demanding four exchanges.

The Exchange Test includes 60 items evenly distributed over 5 treatment levels. Each treatment level comprises 12 items, characterized by a specific number of required cognitive operations to be performed based on the increasing number of involved symbols and exchanges. Users only need a keyboard to complete the Exchange Test since there are no mouse click sensitive areas and only keystrokes are being stored.

The exchanges have to be performed mentally. Therefore, all intermediate configurations of symbols need to be upheld in mind. The cognitive load increases over the treatment levels and therefore the load on working memory. The individual's capacity of working memory manifests as the number of correct counts of exchanges. Due to the limited capacity of working memory, the probability of miscounts rises with the treatment levels. This effect is assured through the fact that participants are asked to come to a solution as quickly as possible ("speed test"). Participants are instructed to press the response key "ENTER" by the time they come to the solution. This keystroke removes the respective item and the always same response screen appears. Figure 2 provides a screenshot of this response format.

Figure 2. Screenshot of response format.

When this screen with the numbers 0 to 6 appears, none of the numbers is highlighted since highlighted numbers indicate the chosen response. Therefore, by pressing a number on the keyboard, the corresponding number on the screen is visualized by framing it. Participants can change their answer. With the next stroke of the enter key the solution is being finally stored. A following intermediate screen confirms that the response was saved successfully and asks to press "ENTER" to present the next item. While an item is being presented participants have a timeframe of 30s (countdown implemented on client-side) to come to a solution and to press "ENTER". In case they don't do so, no reaction time and accuracy score are recorded (coded as –99) but an intermediary screen appears, which brings the time limit back to mind and calls to work faster. The next item appears once "ENTER" is pressed.

A user specific MySQL data set is being compiled on the server before the first trail starts. All intermediate results are stored client-sided as long as the Exchange Tests' 60 trails haven't been completed and will be transferred to the server at the end of the testing.

Participants

Participants were 220 college students (Another 12 students agreed to attend the study but since none of them completed the Exchange Test–neither online nor offline–they are not included in the final sample). Participants received either a financial reward (6€) or extra credit. Participants were mainly students of majors (31% psychology, 10% educational studies, 8% educational sciences, and 51% others) with an uneven gender ratio whereby more females (151) than males (69) attended this study. The sample ranged from 18 to 49 years of age (mean age of 23.1, $SD = 4.7$). Participants were randomly assigned to two conditions: One group ($N = 113$) was supposed to complete the first administration of the Exchange Test in an offline laboratory setting and the second online via Internet whereas the other group ($N = 107$) attended both settings in reverse order. The two groups did not differ in age, $t(218) = -.13$, $p = .90$, or gender composition, $\chi^2(1, N = 220) = 1.86$, $p = .17$.

A subsample of 137 participants attended a second independent study and in this context completed Raven's (1962) Advanced Progressive Matrices (APM). The APM is widely recognized as a measure of fluid intelligence (see Jensen, 1998) and was therefore expected to correlate strongly with accuracy scores collected in offline as well as in online administration. 48 participants of this subsample had the first administration of the Exchange Test via Internet whereas the other 89 started with the laboratory testing. In both groups the APM was administered in the laboratory after the Exchange Test had been completed at first.

Procedure

In order to control differences in test media and in characteristics of samples, an identical web-enabled version of the Exchange Test was administered to all participants twice in two different settings: an offline laboratory administration and an online Internet administration. Hence, this procedure ensures that a possible non-equivalence cannot be explained by sample characteristics or by design of the stimuli.

As suggested by Klink (1998, 2002), a counterbalanced test-retest design was chosen. This design allows the comparison of mean scores and variances between offline and online administration, and can reveal possible effects of order between groups due to the counterbalanced administration contexts. Usually, in this design a test is given to four groups: to two groups in both contexts but in different orders (offline-online vs. online-offline) and, furthermore, to two groups each time in the same context (online-online & offline-offline). The realization of the two last conditions is of interest when addressing the question of reliability and has been skipped for this equivalence study. Therefore, only the core-conditions including both environments have been realized. Hence, this study is based on a two factorial design of repeated measurement. Figure 3 visualizes the study design.

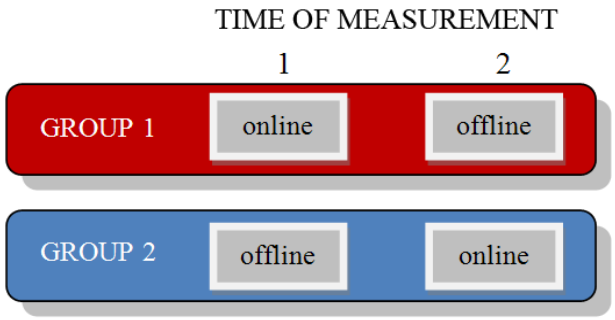

Figure 3. Study design: two factorial design of repeated measurement with two experimental conditions (GROUP 1 & 2) and two measurements (1st & 2nd TIME OF MEASUREMENT).

Participants were recruited in lectures and by postings at Goethe University Frankfurt. Participants were told that for successful participation it was necessary to take part in a repeated measurement design and that remuneration was only given when *both* administrations were completed. Once participants agreed to complete the Exchange Test twice, they were assigned randomly (arising from order, the first measurement was completed online or offline) to one of the two conditions.

Since the laboratory was only available for a couple of weeks, towards the end of the testing period the second offline measurement would have fallen in the time where the laboratory was already booked by another research team. As a consequence, during this period, already recruited participants were only assigned to the condition that was supposed to start with the offline measurement in the lab (N = 41). Therefore, more participants attended the offline-online condition (N = 132 vs. N = 88).

Participants were asked personally or by phone to name two days they would be able to compete the Exchange Test online and offline (or offline and online). The target was an average testing interval of 10 days. Therefore participants were instructed to choose a retest interval not less than 9 days and not longer than 11 days. Subsequently, participants received an e-mail containing the confirmation of dates and a six-digit unique ID as a login for the two upcoming measurements. This way the data sets could be matched, as the sets were linked to the same ID. The overall average test-retest interval was 11.4 days (SD = 4.6). The two groups did not differ in this interval statistically, $t(218) = 1.27$, $p = .21$. On average, group 1 completed the online measurement 10.9 days (SD = 5.3) before they came to the laboratory, whereas group 2 completed the Exchange Test online 11.7 days (SD = 4.2) after they have been tested offline.

The Exchange Test runs without bugs on the most common browsers. For the purpose of this study, the execution in both situations was restricted to Internet Explorer's Kiosk Mode (full screen). Hence, no mobile

devices were used. On request, two participants (both randomized to group 1 “online-offline”) received an extra link to run the Exchange Test on a Macintosh Computer.

Participants received two days before the offline testing an individual reminder by e-mail containing date and time as well as directions to the laboratory. The laboratory was composed of three cabins each with a computer workstation. Each workstation was an AMD Sempron 2.7 GHz (2 GB DDR3 SDRAM) connected to a 19" monitor.

The examiner welcomed the participants and asked them to choose one of the spare cabins where the start page with the login mask was already opened in the laboratory testing situation. There was no instruction since the same self-explaining version as online was used and additional interaction could have had effects on performance.

For the online measurement an e-mail with a link and detailed step by step instruction to access the Exchange Test via Internet was sent to each participant one day before the individually arranged online testing period started.

Data Treatment

The mean of reaction time over the treatment levels serves as the dependent variable "reaction time". The "accuracy score" is defined by the total of false responses.

Data analysis revealed that 9 participants' (4%) mean reaction time in at least one treatment level was 1 second or shorter. Since it is very unlikely to perform the required mental exchanges this quickly, it is reasonable to believe that these participants either misunderstood the instruction or were non-compliant. This impression is strengthened by the fact that these participants’ answers (number of exchanges) were incorrect in at least 50%. Data from these participants were excluded from further analysis since they obviously emphasized speed over accuracy. Therefore, their accuracy scores can’t be treated as a measure of working memory capacity. Three out of the 9 excluded participants belonged to the group that had the first measurement in the laboratory. In the end, the reported results are based on the remaining 211 participants ($N_{\text{offline-online}} = 129$, $N_{\text{online-offline}} = 82$).

Table 1 provides an outlier analysis for both dependent variables over time and administration contexts.

Table 1
Outlier Analysis for Reaction Time and Accuracy over Both Measurements (t_1 & t_2) and Administration Contexts (Online & Offline)

	Reaction time				Accuracy			
	t_1		t_2		t_1		t_2	
	offline	online	offline	online	offline	online	offline	online
Far outlier 3 IQR above 75th percentile	-	1	-	-	-	-	-	-
Outlier 1.5 IQR above 75th percentile	1	3	-	3	-	-	-	-
Within Interquartile Range (IQR) +/-1.5 IQR	210	207	211	208	209	211	211	209
Outlier 1.5 IQR below 75th percentile	-	-	-	-	2	-	-	2
Far outlier 3 IQR below 75th percentile	-	-	-	-	-	-	-	-

Table 1 shows all in all (considering reaction time and accuracy over both measurements and administration contexts) there are 12 outliers (referable to 11 participants). These participants were not excluded for the following reasons: Four of them produced outliers on accuracy by performing very well. They gave not one single incorrect answer. The other eight outliers represent reaction times well below average, which go hand in hand with accuracy scores above average ($AM_{t1} = 4.6$, $AM_{t2} = 3{,}7$). Therefore these data were not excluded from analysis since they represent a reflective (slow and accurate) cognitive style (see Kagan, 1965).

An analysis of variance (ANOVA) can be presented in terms of a linear model. Therefore, the distribution of the residuals needs to be normal. Check of residuals revealed the violation of the assumption of normality for accuracy scores, $z_{t1} = 1.18$, $p_{t1} = .12$ & $z_{t2} = 2.00$, $p_{t2} < .01$, but not for reaction time, $z_{t1} = 1.03$, $p_{t1} = .24$ & $z_{t2} =$

.83, p_{t2} = .50. Thus, a root transformation of accuracy scores was applied and led to a Gaussian distribution, z_{t1} = 1.03, p_{t1} = .24 & z_{t2} = .87, p_{t2} = .43. In this way, the ANOVA with repeated measurements of the accuracy scores is based on the root transformed data.

Results

Figure 4 illustrates the arithmetic means and standard errors of the means for reaction time and accuracy scores by experimental condition and time of measurement.

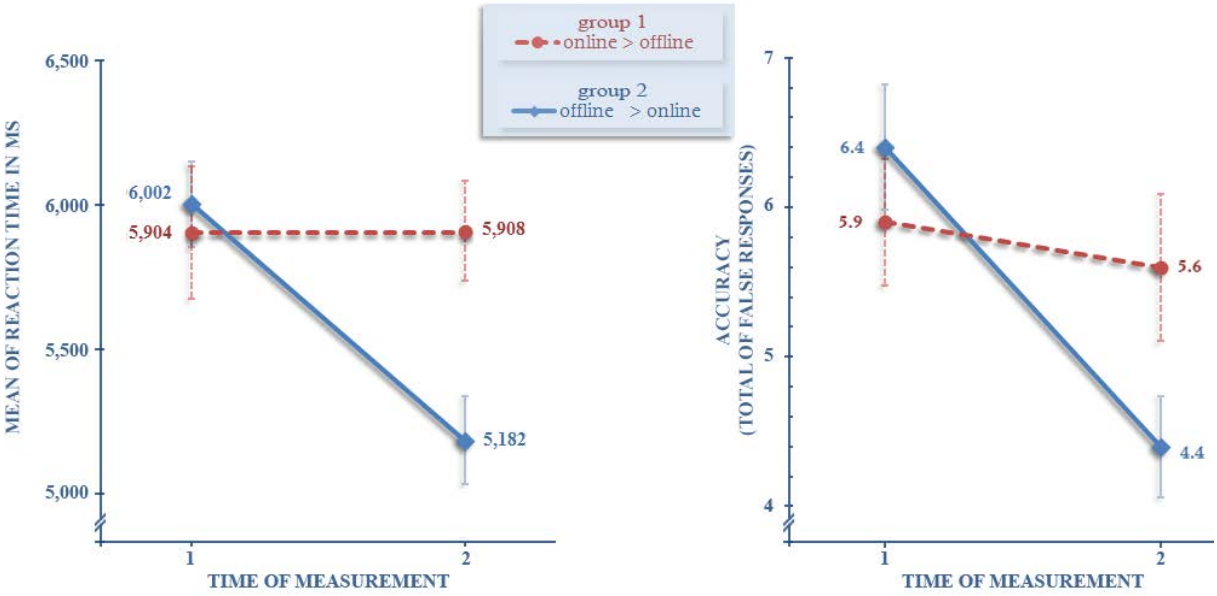

Figure 4. Means and variances for reaction time and accuracy by experimental condition and time of measurement.

Descriptively, group 1 (online-offline) shows almost identical scores on both dependent variables at both times of measurement. The mean reaction time increases slightly from 5,904ms (*SD* = 2,080) to 5,908ms (*SD* = 1,574). For the accuracy scores there is a marginal decrease from 5.9 (*SD* = 3.4) to 5.6 (*SD* = 4.0) false responses. Statistically, within this group there is no difference between online and offline administration for reaction time, $t(81) = -.19$, $p = .99$, and for accuracy scores, $t(81) = -.51$, $p = .61$.

In contrast, group 2 (offline-online) shows descriptively a consistent decrease on both variables: on average, the reaction time decreases from 6,002ms (*SD* = 1,683) to 5,182ms (*SD* = 1,723) and the number of false responses decreases by 2 from 6.4 (*SD* = 4.3) to 4.4 (*SD* = 3.5). On average, compared to the offline administration, group 2 performs online significantly quicker, $t(128) = 7.11$, $p < .01$, and with fewer errors, $t(128) = 4.39$, $p < .01$.

The overlap of error bars indicates that the groups differ neither in their reaction times, $t(209) = .37$, $p = .71$, nor in their accuracy scores, $t(20.70) = 85$, $p = .40$, at the first time of measurement. However, at the second time of measurement, group 2 scores significantly better on reaction time, $t(209) = -3.09$, $p < .01$, as well as on accuracy, $t(209) = -2.13$, $p = .04$.

So far, only a comparison of means has been presented, since this representation reveals nicely how different study designs affect the scientific outcome: a mere comparison of two groups with only one measurement as well as the unparalleled repeated measurement of group 1 would have led to the assumption of equivalent means. Only the counterbalanced test-retest design exposes more detailed effects: ANOVA reveals, the differences found for group 2 are strong enough to cause a main effect within all subjects for repeated measurement of reaction time, $F(1,209) = 11.86$, $p < .01$, $\eta^2 = .054$, and accuracy, $F(1,209) = 9.25$, $p < .01$, $\eta^2 = .042$. By assessing the interaction of the factors measurement and group, the question of equivalence of offline and online administration was finally addressed. For both dependent variables the interaction is highly significant: within the design the location of administration affects reaction time, $F(1,209) = 11.13$, $p < .01$, $\eta^2 = .055$, and accuracy, $F(1,209) = 10.23$, $p < .01$, $\eta^2 = .047$.

Rank orders

Table 2 shows results of the analysis of stability of rank orders between both locations.

Table 2
Correlation Coefficients for Reaction Time and Accuracy by Administration Order. Statistical Parameters Are Related to Differences Between Administration Orders

	N	*r*	*z_diff*	*p*
REACTION TIME				
offline-online	129	.71**	3.59	<.0
online-offline	82	.35*		1
ACCURACY				
offline-online	129	.67**	2.37	.02
online-offline	82	.43**		

For reaction time, group 2 shows a considerably higher stability of ranks, $r(129) = .71$, $p < .01$, than group 1, $r(82) = .35$, $p = .01$. Likewise, for accuracy scores group 2, $r(129) = .67$, $p < .01$, shows fewer deviations in rank order than group 1, $r(82) = .43$, $p < .01$.

Fisher r-to-z transformation reveals that this difference in stability of rank order is significant for reaction time, $z = 3.59$, $p < .01$, as well as for accuracy, $z = 2.37$, $p = .02$.

Equivalence for 1st time of measurement

As shown above, investigations of the data by means of repeated measurement ANOVA revealed a main effect of repeated measurement and a systematic interaction on both dependent variables. Both effects will be discussed later, however, with respect to the question of equivalence it seems to be most reasonable to focus only on the first time of measurement. This elicitation enables the mere comparison of both administration forms unaffected by learning or any effect of order.

It has already been demonstrated that for the first time of measurement there were no differences in means for both dependent variables. Another requirement for equivalence is that variances are approximately the same. Results of F-tests (single factor ANOVA) show that the score distributions of offline and online administration do not differ for reaction time, $F(1,210) = .14$, $p = .71$, or accuracy respectively, $F(1,210) = 3.06$, $p = .08$.

The shape of distributions coincides for both dependent variables: for reaction time, $\chi^2(210, N = 211) = 211.0$, $p = .47$, and for accuracy, $\chi^2(22, N = 211) = 22.7$, $p = .42$, there is no significant divergence between the shape of distributions.

Last but not least the accuracy scores of both administration forms correlated strongly with APM scores. For group 2 (offline-online) a Pearson's, $r(48) = .56$, $p < .01$, was observed, for group 1 (online-offline) it was, $r(89) = .52$, $p < .01$. Fisher r-to-z transformation revealed there were no statistical differences in correlations of online and offline data with APM, $z = .33$, $p = .74$.

Evaluation of user feedback on technical or other difficulties

After completing the Exchange Test, participants had the opportunity to report any technical or other difficulties. For the offline situation in the laboratory there was not a single report. For the online context four participants (1.9%) reported irregularities. Two of them reported that they had to restart the Exchange Test since it locked up when switching from the practice to test mode. One participant reported the flickering of their monitor and one other participant reported that the test locked up once in a while.

Additional information on outliers

All reported analyses include outliers which weren't excluded for the reasons specified above. Figure 5 shows that the exclusion of the people concerned doesn't change the basic pattern of data.

As before, for both groups and dependent variables there is a significant interaction of factors and a significant increase in performance from t_1 to t_2 for group 2, all $p \leq .01$. Furthermore, there is no significant difference in performance between both groups at t_1 and group 1s' reaction time and accuracy scores don't change significantly, $.17 \leq p \leq .72$.

Summary of results

To sum up, the Exchange Test proved to be user friendly and technically stable. At the first measurement both dependent variables do not differ between online and offline administration–there is no effect of administration mode. On average, same reaction time and accuracy scores result for the second measurement (offline) when the Exchange Test previously has been completed online. In contrast, the average participant completes the Exchange quicker and more accurately at the second measurement (online) when the first administration has taken place in the laboratory (offline). All these findings can be found independently from the treatment of outliers, because their exclusion doesn't change this pattern. Notably, in both dependent variables, rank correlations between administration modes are more stable for the group that performs better at 2nd administration in contrast to the group that shows on average no change in performance. When the Exchange Test is administered for the first time, there is a strong correlation with APM for online as well as for offline administration.

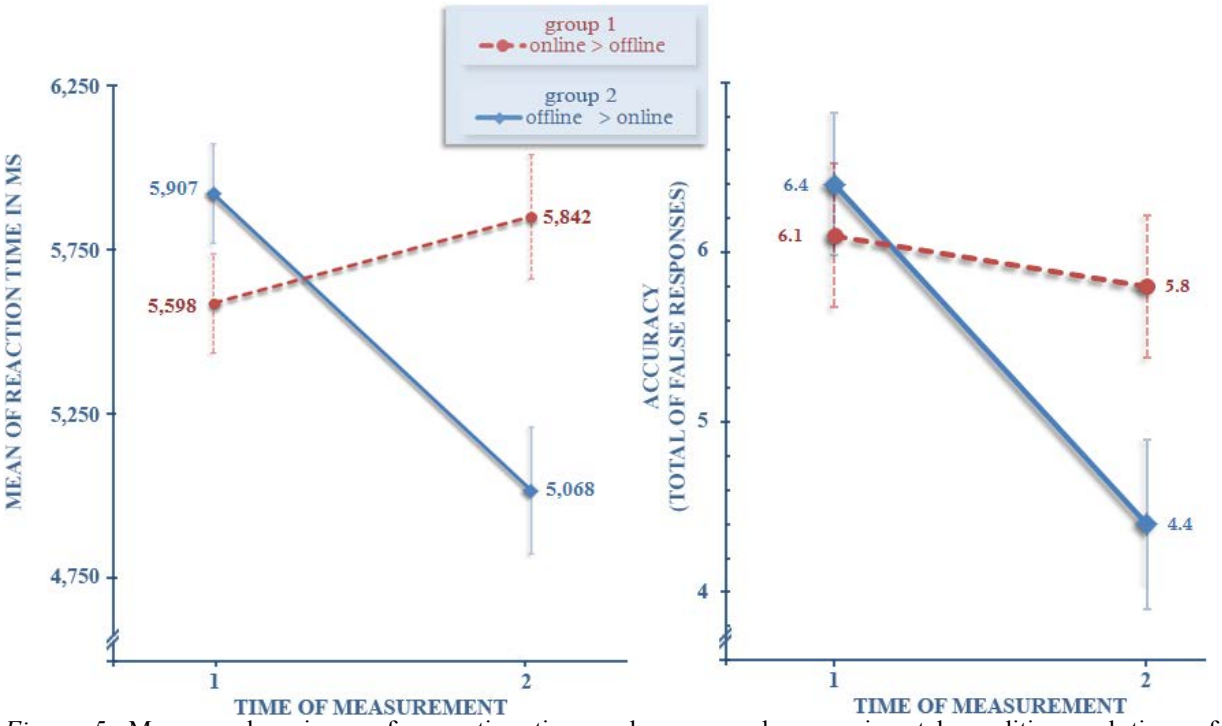

Figure 5. Means and variances for reaction time and accuracy by experimental condition and time of measurement (outliers excluded).

Discussion

There is a lot of support for online versions to yield comparable psychometric properties to already established offline measures. Notwithstanding, since equivalence of administration forms cannot be taken for granted, online measures need to be proven equivalent to their offline counterparts. In testing equivalence, researchers use counterbalanced designs frequently. In aggregating data by administration forms, only part of the methodological options is exhausted and possible interaction effects may remain undetected.

The present study addressed the question of equivalence of offline and online administration of the web-enabled version of the Exchange Test in a counterbalanced test-retest-design. By means of repeated measurements ANOVA, differences between counterbalanced subgroups were described in detail and an interaction effect appeared: Online and offline administration resulted in comparable test characteristics at the first, but not at the second administration (between subjects). For the group that had the first administration via Internet, accuracy scores and reaction time remained stable across offline and online administration, whereas the group who started in the laboratory scored better on both variables at the second online administration (within subjects). These results show that the aggregation of data may conceal interactions and produces means, which insufficiently describe differences between counterbalanced subgroups. Hence, uncovering this method-effect is an advantage of using the chosen combination of design and analysis.

But to go beyond statistics: what determines this difference at the second administration? First of all, it should be pointed out that the stability in test scores (group 1) rather than the improvement (group 2) is notable: the

Exchange Test is an achievement test and therefore learning effects are not unusual. Ackerman (1988) found reaction times in particular to decrease notably throughout practice sessions. Concerning accuracy, a meta-analysis over 40 aptitude and achievement tests conducted by Kulik, Kulik and Bangert (1984) demonstrated that the number of false responses can decrease by practice. In the light of learning effects one can discuss the usefulness of the repeated measurement within an experimental design.
Nonetheless, it is remarkable to observe that only group 2, who had the first contact with the Exchange Test in a controlled laboratory setting, shows learning effects as well as considerably higher stability of rank orders over time. In accordance with the theory of limited working memory capacity, the elimination of disturbances in a laboratory offline setting is a plausible moderator of this effect: under optimized conditions in a laboratory, participants do not need to spend limited cognitive resources on the suppression of interfering disturbances. Cognitive load theory (Sweller, 1988, 2005) is devoted to the question of how to deal effectively with our limited cognitive resources. To achieve best results in learning, learning situations and environments should help individuals to focus on relevant stimuli. Following this reasoning, in uncontrolled online environments disturbances would be more likely and should impair learning processes more than optimized offline settings. This theory *may* describe the data: Regarding the first contact to the Exchange Test as a learning situation and the second as a performance review, participants of the controlled offline learning situation scored better than the self-regulated online learners.

But in the end we need to emphasize that this is a post hoc explanation. Unfortunately, the current study cannot ascertain for sure what lead to the observed interaction. One possible cause has been stated above but other explanations beyond the theory of limited working memory capacity need to be considered as well. A useful hint can be found at Ollesch, Heineken and Schulte (2006). They varied not only the physical environment but also the mode of presence of an experimenter and therefore the social situation. They found an effect of social situation on participants' efforts when giving a written report. In their terms in the present study the experimenter attended the offline testing situation but was not actively involved (participants knew the experimenter waited for them behind the door to complete the Exchange Test). Social facilitation theory (the tendency to perform better in speed and accuracy at simple tasks in the mere presence of other people but to impair in performance at less familiar tasks) suggests that the social environment affects participants' efforts in achievement tests like the Exchange Test. Therefore, despite all attention on an objective procedure in the current study, there might be an effect of the mere contact with the experimenter which may have affected the repeated measurement asymmetrically.

In the end, to reduce the uncertainties connected with the view on data in its entirety, only data of the first measurement were taken into consideration when analyzing equivalence, since (a) there were significant differences between both groups in rank orders from first to second measurement and (b) the second measurement of both groups seems to be affected to a varying degree by a learning effect. Therefore, the assessment of equivalence isn't based on results achieved by repeated measures ANOVA.

All requirements for establishing parallel forms are met at the first time of measurement: (1) there are no differences in means and variances, (2) score distributions correspond, and (3) both administration forms correlate to the same degree with APM.

A limitation of this study might be the controlled recruiting of participants (see Riva, Teruzzi, & Anolli, 2003): the sample in the present study was obviously comprised of highly motivated students; they were recruited personally and once they attended the first measurement there were no dropouts. In contrast, Internet samples are usually known to be more heterogeneous in motivation and demographical characteristics.

There are two messages: The present study and findings from Schreiner, Altmeyer, and Schweizer (2012) contribute substantially to investigate the equivalence of offline and online administration of the web-enabled version of the Exchange Test. And in taking heed of solutions for web experiments offered by Reips (2000), one can be optimistic about unattended online data collection, as there is already strong support for equivalence.

In addition to the question of equivalence there is also a methodic aspect. The chosen cross-mode test-retest study design revealed an interaction effect that would have remained undetected with most widely used methods. Unfortunately, based on the limited information from the current study we cannot explain this effect yet. Our discussion provides at least two reasonable and theoretically well-founded explanations, but further research should (1) clarify whether this kind of interaction can also be found when the same method is applied to other web-enabled achievement tests and (2) help to understand what determined the present data pattern.

In doing so, similar studies should realize the two missing conditions (online-online & offline-offline) as well. This should help to understand what produced different results at second administration. Additionally, a systematic variation of cognitive load and social situation might help to understand the interaction effect.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*, 288–318.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, *55*, 803–832.

Buchan, B. D., DeAngelis, D. L., & Levinson, E. M. (2005). A comparison of the web-based and paper-and-pencil versions of the career key interest inventory with a sample of university women. *Journal of Employment Counseling, 42*(1), 39–46.

Buchanan, T. (2003). Internet based questionnaire assessment: Appropriate use in clinical contexts. *Cognitive Behaviour Therapy*, *32*, 100–109.

Buchanan, T., Ali, T., Heffernan, T. M., Ling, J., Parrott, A. C., Rodgers, J., & Scholey, A. B. (2005). Nonequivalence of on-line and paper-and-pencil psychological tests: the case of the prospective memory questionnaire. *Behavior Research Methods, 37*, 148–54.

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428–434.

Epstein, J., & Klinkenberg, W. D. (2001). From Eliza to Internet: A brief history of computerized assessment. *Computers in Human Behavior*, *17*, 295–314.

Ghiselli, E. E., Campbell, J. P. & Zedeck, S. (1981). *Measurement theory for the behavioural sciences.* San Francisco, CA: Freeman.

Hays, S., & McCallum, R. S. (2005). A comparison of the pencil-and-paper and computer-administered Minnesota Multiphase Inventory–Adolescent. *Psychology in the Schools, 42*, 605–613.

Honaker, L. (1988). The equivalency of computerized and conventional MMPI administration: a critical review. *Clinical Psychology Review*, *8*, 561–577.

Horn, W. (1983). *Leistungsprüfsystem* [Performance Testing System]. Göttingen. Hogrefe.

Ihme, J. M., Lemke, F., Lieder, K., Martin, F., Müller, J. C., & Schmidt, S. (2009). Comparison of a computeradministered ability test used online and in laboratory. *Behavior Research Methods, 41*, 1183-1198.

International Test Commission (2005). *International Guidelines on computer-based and Internet delivered testing*. Retrieved June 12, 2013 from http://www.intestcom.org/guidelines

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

Kagan, J. (1965). Impulsive and Reflective Children: Significance of Conceptual Tempo, in J.D. Krumboltz (Ed.), *Learning and the Educational Process* (pp 133-161). Chicago: Rand McNally.

Klinck, D. (1998). Papier-Bleistift- versus computergestützte Administration kognitiver Fähigkeitstests: Eine Studie zur Äquivalenzfrage [Comparison of paper-and-pencil vs. computerized administration of the NEO-Five-Factor-Inventory (NEO-FFI)]. *Diagnostica, 44,* 61–70.

Klinck, D. (2002). *Computergestützte Diagnostik* [Computer-based diagnostics]. Göttingen: Hogrefe.

Kulik, J. A., Kulik, C.-L. C., & Bangert-Drowns, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*, 435–447.

Meyerson, P., & Tryon, W. W. (2003). Validating internet research: A test of the psychometric equivalence of internet and in-person samples. *Behavior Research Methods, Instruments, & Computers, 35*(4), 614–620.

Ollesch, H., Heineken, E. & Schulte, F.P. (2006). Physical or Virtual Presence of the Experimenter: Psychological Online-Experiments in Different Settings. *International Journal of Internet Science, 1*, 71–81.

Pedersen, E. R., Grow, J. Duncan, S., Neighbors, C., & Larimer, M. E. (2012). Concurrent validity of an online version of the Timeline Followback assessment. *Psychology of Addictive Behaviors*.

Ployhart, R. E., Weekley, J., Holtz, B., & Kemp, C. (2002). Web-based vs. paper and pencil testing: A comparison of factor structures across applicants and incumbents. In F.L. Oswald & J.M. Stanton (Chairs and Eds), *Being virtually hired: Implications of web testing for personnel selection*. Symposium presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Canada, April.

Raven, J. C. (1962). *Advanced progressive matrices*. London: Lewis & Co.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–114). San Diego: Academic Press.

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*, 243–256.

Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the internet in psychological research: Comparison of online and offline questionnaires. *CyberPsychology & Behavior, 6*, 73–80.

Schreiner, M., Altmeyer, M., & Schweizer, K. (2012). The web version of the Exchange Test: Description and psychometric properties. *European Journal of Psychological Assessment*, *28*(3), 181–189.

Schreiner, M., & Schweizer, K. (2011). The hypothesis-based investigation of patterns of relatedness by means of confirmatory factor models: The treatment levels of the Exchange Test as example. *Review of Psychology, 18*(1), 3–11.

Schroeders, U., Schipolowski, S., & Wilhelm, O. (2009). Internet-based ability testing: Problems and opportunities. S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for behavioral research on the Internet* (pp. 131-148). Washington, DC: American Psychological Association.

Schweizer, K. (1996). The speed-accuracy transition due to task complexity. *Intelligence, 22,* 115–128.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 19-30). Cambridge, MA: Cambridge University Press.