Psychological Test and Assessment Modeling, Volume 59, 2017 (2), 229-243

Test order effects in an online self-assessment: An experimental study

Lisbeth Weitensfelder[1]

Abstract

While computer based online self-assessments become increasingly popular, not much research can be found yet on how such test batteries should be designed regarding test order, though effects on motivation and performance can't be ruled out. The given study examines in an experimental setting how test order influences dropout rates and fatigue in a computer based online self-assessment. Test order could at the most reduce overall dropouts to a limited degree – especially a combination of a warming up phase (e.g. a short questionnaire) with a subsequent high hurdle (e.g. a strenuous test) can be recommended. Contrary to studies that focused on high stakes situations, fatigue effects in strenuous tests could be found.

Key words: Context effects; self-assessment; test order; fatigue; dropout rates

[1] *Correspondence concerning this article should be addressed to:* Lisbeth Weitensfelder, doctoral student at the University of Vienna, Faculty of Psychology (Division of Psychological Assessment and Applied Psychometrics), Liebiggasse 5, 1010 Vienna, Austria; email: lisbeth.weitensfelder@univie.ac.at

Computer based online self-assessments, where testees can assess their abilities, proficiencies, interests or attitudes without supervision of trained test administrators, are becoming increasingly popular in educational contexts: When it comes to provide support for students to select the field of study that best matches their interests and abilities, online self-assessments are already an important assessment tool in German-speaking countries (for examples see Hornke, Wosnitza & Bürger, 2013, or Kubinger, Frebort, Khorramdel & Weitensfelder, 2012, 2013). For the design and development of such online test batteries some recommendations exist with regard to the use of certain tests and instruments: Kubinger (2015) suggests the use of objective personality tests based on experimental-based behavior tasks in addition to personality questionnaires based on self-ratings, he also enunciates against too brief test batteries, which might not meet psychometric quality standards. However, not much research can be found yet on how such online-assessments should be put together regarding the order of the contained tests – this paper aims to close this gap.

Context effects regarding test order and test length

While also the presentation mode of online vs. traditional testing might have an effect, making it possible that online versions measure partly different constructs (e.g. Buchanan, 2001, 2002), the following study focuses solely on online self-assessments. Previous research regarding test order effects or effects of test length usually focuses on "traditional" testing situations. Hereby, the results are contradictory. While for the change of item order, resulting context effects can be seen as proven (e.g. Franke, 1997; Knowles, 1988; Ortner, 2008), results about effects caused by test order are inconsistent. Khorramdel and Frebort (2011) showed in a sample of managers that test order might have effects on the performance of experimental-based behavior tasks, but they found no effects on cognitive ability tests. The authors concluded that test order rather affects simple than complex tasks. However, these results could not be reproduced in a later study, where no test order effects could be found (Schünemann, 2013).

Test order effects might be closely related to test length effects, since with ongoing test time users might experience fatigue. But several studies find no (e.g. Tulsky & Zhu, 2000) or only small (e.g. Zhu & Tulsky, 2000; Ryan, Glass, Hinds & Brown, 2010) fatigue effects on performance scores. It might sound surprising that longer test length does not necessarily lead to fatigue effects, but that goes along with findings of Davis (1946) who found three patterns of performance change with ongoing test time of a 70-minute task: A stable performance (shown by about three-quarters of the sample), an increase of performance (approx. 17%) and a decrease (less than 8%). A study of Ackerman and Kanfer (2009) affirms that performance does not necessarily have to go down with long test length conditions: Testing three versions of the SAT, a test for scholastic aptitude widely used in the U.S. (with test lengths between 3.5 and 5.5 hours), they found that performance even increased in the longer versions, being dissociated with subjective fatigue which rather increased with longer test length. Similar results could be found in a small sample earlier (Liu, Allspach, Feigenbaum, Oh & Burton, 2004), where prolonging the SAT with an essay did not affect performance negatively. Jensen, Berry and Kummer (2013) even showed that a lengthier version of a biology exam resulted in better performance than a shorter version. However, that does not lead to conclusions for the context of self-assessments. In high stakes tests as the SAT, the results have a powerful implication on the testees' educational pathway, which does not necessarily apply to self-assessments.

Effort in low stakes testing

One possible explanation for the dissociation between subjective fatigue and performance as found by Ackerman and Kanfer (2009) was seen in compensation mechanisms: The authors suggested that people realized their fatigue and therefore put in more effort – an explanation that makes sense in high stakes testing situations. In low stakes testing situations, where test results have no or only a small implication, the situation could be different. Self-assessments for counseling purposes surely can't be classified as high stakes situations (like entrance exams or the SAT test), but they also differ from "classical" low-stakes testing situations (like PISA) since users undergo the assessment for counseling purposes. Lack of self-relevance has been identified as one of the problems of low-stakes testing (e.g. Finn, 2015, for an overview). Therefore, the wish for appropriate feedback could be a motivator to show some effort in the test battery.

As Wise and DeMars (2005) pointed out, motivation has a big impact on performance in low stakes testing, with more motivated students scoring an average *g* effect size of .59 better than less motivated students. But this might not apply to everyone and every sort of test equally: Barry, Horst, Finney, Brown and Kopp (2010) identified three types of effort in low-stakes testing. Two groups differed in their overall level of effort, but both showed less effort in a cognitive test than in non-cognitive ones (test-taking effort was assessed via a set of items that participants responded to after each contained test). A third group showed moderately high effort over both cognitive and non-cognitive tests. Hints in literature show that such differences could at least partly be caused by relatively stable traits: E.g. Brown and Finney (2011) found that non-responders (representing a group with low motivation) in a low stakes assessment show a different level of reactance than responders. When domain-specific motivational traits (like self-concept) are controlled, the correlation between situation-specific motivation and performance turns out to be smaller, though it still exists (Penk, 2015).

Dropout rates in online surveys and self-assessments

While online research offers several benefits like easy access for testees or greater external validity (e.g. Reips, 2002), one of its disadvantages can be seen within high dropout rates (Frick, Bächtiger & Reips, 2001; Reips, 2002). There has been some research on how to reduce those dropout rates in online research (Frick et al., 2001; Göritz & Stieger, 2008): Personal information asked in the beginning can reduce dropout rates, as well as incentives like the possibility of winning a lottery (Frick et al., 2001). However, in the case of self-assessments, incentives are no realistic possibility, and asking personal information in the beginning could bear the risk of scaring away users who are concerned about their privacy. Therefore, another method seems promising: Putting high hurdles, which means motivational adverse factors, close to the beginning, so that most dropouts happen in the beginning (Reips, 2002). Hurdles can come in diverse forms, such as the need for patience (e.g. via loading times or long texts in the beginning), or technical pretests (Reips, 2002). But as Göritz and Stieger (2008) showed, not all high hurdles might indeed be beneficial: Increasing loading times of webpages to filter out less motivated participants and reach a higher data quality did not only fail to reach the intended effect, but even led to counterproductive results (a lower likelihood of responding to the study).

Reips (2002) suggests a high hurdle in combination with a warm-up phase for internet-based assessment. Seen in the light of the theory of sunk costs (Arkes & Blumer, 1985), such a suggested warm-up phase could also have the benefit that testees have already invested time and effort in the test battery, making it harder for them to quit.

Results of a pre-study. In the *Viennese Self-Assessments* (Kubinger et al., 2012, 2013), dropout rates are especially high for two technical fields of study, namely *Architecture* and *Mechanical Engineering*, where they outnumber finishing rates. As shown in a pre-study (Weitensfelder, Frebort, Müller & Mitschek, 2011), most of the dropouts happen at the beginning of the test batteries, but with different gradients: While a majority of dropouts in *Mechanical Engineering* happens during the first test, the dropout process in *Architecture* is more lingering (figure 1). The test batteries are similar, yet include some

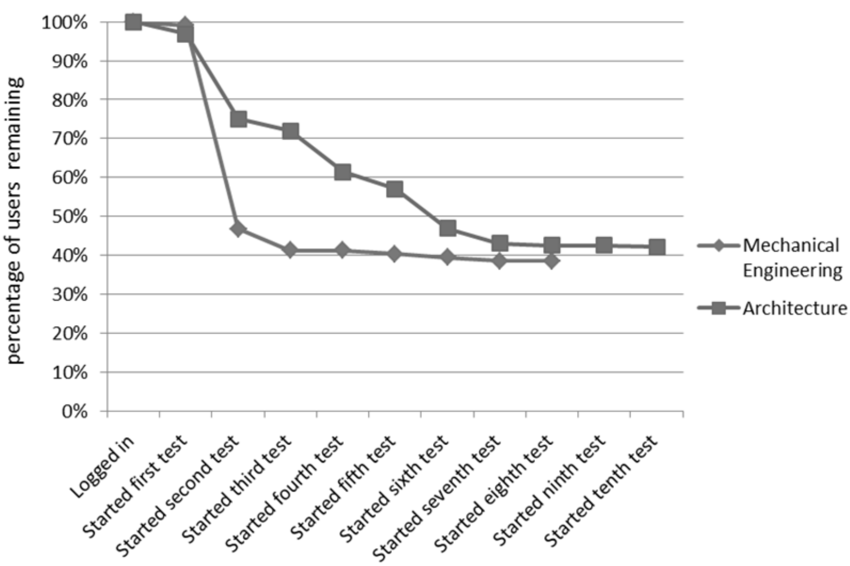

Figure 1:
The dropout process of two different test batteries in a pre-study (modified from Weitensfelder, Frebort, Müller & Mitschek, 2011)

differences regarding contained tests/items and a different test order, where the position of two strenuous tests is switched: The test battery in *Mechanical Engineering* starts with a (high-hurdle-like) demanding learning test (Kubinger, Haiden, Karolyi & Maryschka, 2012), which is combined with a number series test (Poinstingl, in prep.; see Berndl, Steinfeld & Poinstingl, 2012), while the test battery in Architecture starts with a matrices test (Undeutsch, 2012); the learning test appears much later in the test battery.

The comparison of the dropout processes leads to the question whether the differences might be caused by test order and in how far a changed test order can alternate dropout rates. Regarding reasonability, as being the amount that a testee's resources are preserved (Testkuratorium, 1986), the dropout process in *Architecture* seems especially disadvantageous for users, where they invest much time working on tests before dropping out to the same amount as testees in *Mechanical Engineering*. The ex-post-facto-situation does not allow a causal interpretation (e.g. there might be a chance of trait differences between the groups), so the previous study examines the effect of test order in an experimental setting. (The formulation of hypotheses refers to the study design and therefore can be found in the method section.)

Method

Design

For an experimental design, test order for the test battery in *Architecture* was varied weekly for a survey period of 18 weeks in 2012. Three different test orders were tested:

Version A: Starting the test battery with a matrices test, placing the demanding learning test (combined with the number series test) in the middle of the test battery, as it happened in the pre-study. This test order offers the benefit of a rather high hurdle (the matrices test) in the beginning, yet leading testees to invest already some time before the highest hurdle, so that they might feel determined to finish the test battery.

Version B: Starting with the test that seemed to be the most strenuous one regarding the dropout rates in the pre-study.

Version C: Starting with a short questionnaire about expectations for the field of study, afterwards following the test order as in version B. This test order should have the benefit of a very high hurdle rather in the beginning, but also having a warm-up-phase up front.

Table 1 shows an overview of the test order in the three different versions including short test descriptions. Please note that the learning test is conjoined with a filling test (the number series test, which also serves as an intermediate task between different recalls) which is listed as a separate test, therefore test order in version B seems to be shifted by one position.

Table 1:
Contained tests and short test descriptions

	Version A	Version B	Version C
First	**Matrices Test** (Undeutsch, 2012)	Learning Test	Expectancy Questionnaire
Second	**Coding Test:** An adapted short version of one subtest of the experimental-based behavior task "Work Styles" (Kubinger & Ebenhöh, 2011; see also Kubinger & Ebenhöh, 2012), where symbols have to be coded in a simple task and testees have to estimate how much they will be able to do	Number Series	Learning Test
Third	**Spatial Abilities Test:** Comparing 2D-plans with 3D-images for accuracy (Weitensfelder, 2012)	Coding Test	Number Series
Fourth	**Interest Questionnaire:** Contains study-specific interest questions (Weitensfelder, Undeutsch, Khorramdel & Useini, 2012), reporting two different interest scores	Spatial Abilities Test	Coding Test
Fifth	**Learning Test:** A short version of the learning test LAMBDA (Kubinger, Haiden, Karolyi & Maryschka, 2012), where testees have to learn a company's organization chart	Interest Questionnaire	Spatial Abilities Test
Sixth	**Number Series** (Poinstingl, in prep.; see Berndl, Steinfeld & Poinstingl, 2012), assessing numeric reasoning, also serves as filling test in the learning test. In the second half of the test, a strain condition is added.	Matrices Test	Interest Questionnaire
Seventh	**Personality Questionnaire,** assessing several study-relevant traits (Khorramdel & Maurer, 2012)	Personality Questionnaire	Matrices Test

➡

Eight	**Knowledge Test:** Based on the concept of the LEWITE (Wagner-Menghin, 2004), already existing pre-knowledge in the field of architecture is assessed (see Frebort, Gleeson & Weitensfelder, 2012)	Knowledge Test	Personality Questionnaire
Ninth	**Expectancy Questionnaire,** assessing relevant expectancies and frequent false expectations for the aspired field of study (see Weitensfelder et al., 2012)	Expectancy Questionnaire	Knowledge Test
Tenth	**Questions of Surrounding Conditions,** also for self-reflection of the testees	Questions of Surrounding Conditions	Questions of Surrounding Conditions

Hypotheses

Due to the results of the pre-study, it is not assumed that a simple position exchange of the two strenuous tests (versions A vs. B) might be able to reduce overall dropout rates, yet that shall be tested. However, it seems plausible that a warm-up phase (as suggested by Reips, 2002) lowers the dropout rate. Therefore the first hypotheses are:

Hypothesis 1: Test order of version A leads to a significantly lower dropout rate compared to version B.

Hypothesis 2: A short warm-up phase as in version C can significantly lower the dropout rate compared to version B.

Apart from the consequences on dropout rates, it shall be researched whether test order influences test results. Feedback, especially when being criterion oriented, does not take test order into account, implying that there are no mentionable score differences depending on test order. Therefore test results of the first six tests as in version A should closely be looked at regarding order effects:

Hypothesis 3: There are significant mean differences in test scores between testees from version A and testees from version B.

Hypotheses 1 and 2 are to be tested with simple Pearson chi-square-tests (one-sided). Regarding Hypothesis 3, a multivariate analysis of variances (MANOVA) is conducted for the results of the first test. Additionally, proficiency level is added as a variable in the coding test, even if it is not reported in the feedback for the users: It represents a very simple task and therefore could show different fatigue levels than complex tasks (Khorramdel & Frebort, 2011). All variables are used in raw scores, alpha is set to 5%.

Sample

In the 18 weeks lasting survey period, 676 users logged in for the test battery (users with prior or multiple logins as well as test users and users, where research purposes were known or assumed, were deleted beforehand). A closer look at the data showed that a considerable amount (namely 100 users) seemed to have paused the test battery for a short or longer period of time (of those who finished the test battery, only the ones with incomplete test parts and the ones who took 4 hours or longer for the test battery were manually controlled). Due to their considerable amount, interrupters were not automatically erased from the sample. They were only erased if their break caused problems like a meanwhile changed test order or technical problems as incomplete tests/test batteries (which appeared to n = 62 interrupters and happened especially when the learning test was interrupted). The remaining interrupters were considered as finishers, if they finished within the survey period (n = 5) and were considered as dropouts, if they did not finish while the survey period lasted (n = 33), no matter whether they still logged in or finished after.

After these adjustments 614 users remain in the sample, including 419 dropouts.

Results

Results regarding dropouts

Crosstabs and descriptives for all three versions are shown in table 2. As expected, chi-square-tests do not show significant differences between versions A and B (one-sided Pearson chi-square: p = .165). The process (figure 2) seems to show differences though and is comparable to the one found in the pre-study, with a more lingering dropout process in version A. Investigating the effect of a short questionnaire up front, the one-sided hypothesis shows a significant difference between versions B and C (one-sided Pearson chi-square: p = .041), meaning that a preceding presentation of a short expectancy questionnaire lowers overall dropouts. The overall effect with approx. 8% less dropouts after presenting a short preceding questionnaire before the highest hurdle can be seen as modest. It should be noted though, that the effect would remain insignificant, if all disrupters had been expelled from the sample (remaining samples then would be: n = 164 in version B with 116 dropouts and n = 201 in version C with 129 dropouts; one-sided Pearson chi-square: p = .093).

Descriptively, version B has the highest dropout rate and version C the lowest, but only the difference between B and C remains significant (a comparison of A and C is not part of the study design, but shows only insignificant differences; Pearson chi-square = .721, two-sided p = .396). Regarding the dropout process, B and C look comparable (figure 2) with only a small degree of dropouts happening after the highest hurdle.

Table 2:
Dropouts in all three versions

	Version A		Version B		Version C		Total
	%	*n*	%	*n*	%	*n*	
Dropout	68.3%	157	72.7%	128	64.4%	134	419 (68.2%)
Completed	31.7%	73	27.3%	48	35.6%	74	195 (31.8%)
Total	230 (100%)		176 (100%)		208 (100%)		614 (100%)

Chi-square test for versions A vs. B (n = 406): Pearson chi-square = .951, p (1-sided) = .165; chi-square test for versions B vs. C (n = 384): Pearson chi-square = 3.033, p (1-sided) = .041.

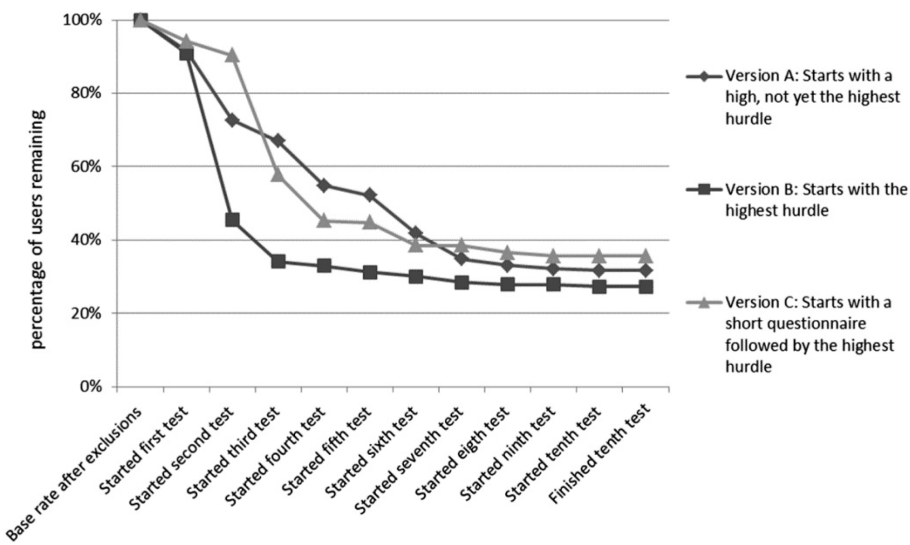

Figure 2:
The dropout process in all three experimental versions

Results regarding performance effects

To investigate hypothesis 3, a multivariate analysis of variances (MANOVA) was conducted to compare means in test scores between version A and version B. Results of the first six tests of version A served as variables. To ensure that the groups were comparable and that breaks wouldn't level out test order effects, only data from users who completed the test battery continuously was used. The Box's M Test proved to be insignificant (p = .076), so that the resulting F-values of the MANOVA can be interpreted fairly.

Table 3:
Multivariate analysis of variances and descriptive statistics for versions A and B

	MANOVA			Version A (n = 67)		Version B (n = 44)	
	F	p	part. η^2	M	SD	M	SD
Matrices test: Items solved	5.092	.026	.045	5.537	3.076	4.205	2.993
Coding test: Proficiency level	.003	.958	.000	46.24	11.960	46.11	12.429
Coding test: Aspiration level	.409	.524	.004	-.165	.187	-.189	.203
Spatial abilities test: Items solved	4.628	.034	.041	2.42	1.281	1.89	1.262
Interest questionnaire: Learning interest (items approved)	.805	.372	.007	13.90	2.432	14.30	2.075
Interest questionnaire: Topic interest (items approved)	1.122	.292	.010	12.90	2.856	13.48	2.791
Learning test: Learning time needed	2.886	.092	.026	280.67	157.944	331.27	146.451
Number series: Items solved (part 1)	1.497	.224	.014	4.85	1.769	4.43	1.757

Table 3 shows the results of the MANOVA and descriptive statistics. Two variables show significant test order effects: the matrices test score (p = .026) and the test for spatial abilities (p = .034). Testees did score a bit lower in both tests, when they appeared later in the test battery (part. η^2 = .045 and Hedges g = .44 for the matrices test; part. η^2 = .041 and Hedges g = .42 for the test for spatial abilities). The other tests did not show effects on performance.

Discussion

The results of an experimental variation of test order in an online self-assessment show that test order has none (hypothesis 1) or only a small impact (hypothesis 2) on overall dropout rates. When comparing two different versions that both start with a high hurdle, there is no difference in overall dropout rates (therefore, hypothesis 1 has to be rejected). However, differences in the dropout process seem visible: When starting with a high, not yet the highest hurdle, more dropouts happen along the ongoing test battery (figure 2). Therefore, regarding reasonability, it seems advisable to put the highest suspected hurdle rather in the beginning of a test battery, so that users who might drop out at a later stage already drop out sooner. A high hurdle rather soon could also be a method to filter out less motivated students; especially for low-stakes assessment, motivation filtering seems advisable anyway (Wise and DeMars, 2005). Of course a very high hurdle as a start might also discourage some users: Descriptively (but not to a significant amount), the version with the highest hurdle in the very beginning has more dropouts than the one that starts with a high, not yet the highest hurdle.

While the position exchange of two strenuous tests did not affect dropout rates, a short expectancy questionnaire up front did (affirming hypothesis 2). The effect of about 8% less dropouts in version C compared to version B can be seen as small, but satisfying given the background that motivational differences in low-stakes assessment might also be caused by relatively stable traits (e.g. Brown & Finney, 2011). However, it remains unclear whether the different dropout rate between versions B and C is caused by a drop-out increase after the highest hurdle in the beginning, a dropout decrease after the short preceding questionnaire or a combination of both. It is possible that putting the highest hurdle in the very beginning (as it happened in version B compared to version A) leads to a dropout increase that is too small to become significant. But even in this case, a possible negative impact of the highest hurdle up front would be outbalanced with a positive impact of the preceding questionnaire. The effect of the questionnaire could be caused either by a general positive effect of a "warming up" phase, or because of the questionnaire content which (by asking for practical expectations for the aspired field of study) might have appeared especially relevant to the testees and increased their com-mitment. A third explanation, namely that the lower dropout rate is caused by a sunk cost effect, does not seem plausible given the background that another strenuous and time-consuming test in the beginning (as it happened in version A) did not lower dropout rates for the later coming highest hurdle – even though that represents much higher sunk costs in case of a dropout.

Overall, dropout rates are still very high and test order as investigated in the study can reduce dropouts only to a small amount. It is possible that other methods (e.g. incentives, continuous feedback) might be more effective in reducing dropouts, but not all methods can realistically be implemented in self-assessment test batteries as the one investigated. Test orders however can be changed easily, so if other methods are not available, at least a limited decrease of dropouts can be accomplished.

Apart from an effect on dropout rates, test order led to fatigue effects in a matrices test and a spatial abilities test (hypothesis 3). Testees who had already completed some tests scored .44 (matrices test) and .42 (spatial abilities test) standard deviations lower (Hedg-es *g*) – a difference that indeed could be relevant. This is inconsistent with previous results, where no (relevant) fatigue effects (e.g. Liu et al., 2004) or even an increase of performance (Jensen et al., 2013) in longer test-taking times were found. It also does not go along with the assumption of Khorramdel and Frebort (2011) that fatigue effects might show rather in simple than complex tasks. However, the test-taking situation in the given study differs from previous studies: In high-stakes testing situations, testees might realize their feelings of fatigue and compensate it with even more effort, which might be one reason for the dissociation between performance and subjective fatigue (Ackerman & Kanfer, 2009). In contrast, testees in self-assessments might not feel urged to put in compensatory effort to counterbalance fatigue effects.

Limitations and implications for future research

As a limitation, it has to be noted that the researched population in the present study is very specific, consisting of users who underwent an online counseling for architectural studies. Though not being probable, it cannot be ruled out that testees interested in completely other fields of study might show a different motivation or dropout behavior.

Another limitation lies within the fact that demographic information of dropouts is not available, since demographic information is only asked in the very end of the test battery. So both for the fatigue effects as well as for the effects on dropout rates, it remains open whether the observed effects do apply equally to men and women. Previous results suggest that this might not be the case: Attali, Neeman and Schlosser (2011) found that men show a larger differential performance than women between low and high stakes assessments. The authors assumed that at least a part of the differential performance was due to the fact that men invested less effort in low stakes assessment. Therefore it might also be possible that in the given study the performance differences or the effects on dropouts might not apply equally to men and women. A detailed examination of possible sex differences in the dropout process could be a matter for future studies. Putting a demographic questionnaire up front, it would be interesting to research whether demographic information as warming-up leads to a similar decrease of dropouts like the expectancy questionnaire, and whether a combination of both can decrease dropouts even further.

A major restriction lies within the necessary exclusion of users with technical problems, which might have caused a small bias: Not all tests caused technical difficulties when interrupted. Especially the learning test could not be completed after taking a break, so that exclusions of users with incomplete test batteries consist mainly of users who interrupted the learning test. Users who interrupted other tests did not always have to be excluded and were then added either to the dropouts or to the finishers. However, it would also be possible, that interrupters represent a separate group regarding their test taking effort. This could also be a matter for future studies.

Regardless of these limitations the present study gives some suggestions on how to design online test batteries. When it comes to dropout rates, the findings confirm suggestions that have already been made by Reips (2002). Additionally, they underline the necessity for further research of test order effects in settings that are not high-stakes: Since the given results in self-assessments do indeed find fatigue effects, such effects should not automatically be ignored when implementing test batteries.

Acknowledgements

All data refers to the “Viennese Self-Assessment” project (Kubinger, Frebort, Khorramdel & Weitensfelder, 2012, 2013).

Special thanks for the good collaboration to the *Technische Universität Wien*, represented by then vice chancellor Prof. Adalbert Prechtl, Ph.D., MSc., Ilona Herbst, MSc., and Franz Reichl, Ph.D.

I also want to show my gratitude to the former staff of the *Centre for Testing and Consulting* of the University of Vienna, especially Prof. Klaus D. Kubinger, Ph.D., MSc., Lale Khorramdel, Ph.D., and Martina Frebort, Ph.D., as well as Sandra Ztrivas, MSc., Fabian Becker, MSc., Thomas Scheck, MSc. and then-intern Leonard Schünemann, MSc.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology, 15,* 163–181.

Arkes, H. R., & Blumer, C. (1985). The Psychology of Sunk Cost. *Organizational Behavior and Human Decision Processes, 35,* 124-140.

Attali, Y., Neeman, Z., & Schlosser, A. (2011). *Rise to the Challenge or Not Give a Damn: Differential Performance in High vs. Low Stakes Tests.* IZA Discussion Paper No. 5693. Institute for the Study of Labor (IZA), Bonn. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1842090 [Feb 20, 2016]

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do Examinees Have Similar Test-Taking Effort? A High-Stakes Question for Low-Stakes Testing. *International Journal of Testing, 10,* 342-363.

Berndl, G., Steinfeld, J., & Poinstingl, H. (2012). Schlussfolgendes Denken numerisch: Der Wiener Zahlenreihentest [Numeric reasoning: The Viennese number series test]. In: K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment:Theorie und Konzepte* (pp. 161-169). Lengerich: Pabst.

Brown, A. R., & Finney, S. J. (2011). Low-Stakes Testing and Psychological Reactance: Using the Hong Psychological Reactance Scale to Better Understand Compliant and Non-Compliant Examinees. *International Journal of Testing, 11,* 248-270.

Buchanan, T. (2001). Online Personality Assessment. In U.-D. Reips & M. Bosnjak (Eds.), Dimensions of Internet science (pp. 57–74). Lengerich: Pabst.

Buchanan, T. (2002). Online Assessment: Desirable or Dangerous? *Professional Psychology: Research and Practice (33),* 148-154.

Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the Effects of Exam Length on Performance and Cognitive Fatigue. *PLoS ONE, 8:* e70270. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0070270 [March 4, 2016].

Davis, D. R. (1946). The disorganization of behaviour in fatigue. *Journal of Neurology, Neurosurgery and Psychiatry, 9,* 23-29.

Finn, B. (2015). *Measuring Motivation in Low-Stakes Assessments*. ETS Research Reports Series ISSN 2330-8516. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ets2.12067/epdf [Feb 20, 2016].

Frebort, M., Gleeson, R., & Weitensfelder, L. (2012). Wissenstest zur Erfassung des bereichspezifischen Vorwissens [Knowledge test to assess field-specific pre-knowledge]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment: Theorie und Konzepte* (pp. 83-91). Lengerich: Pabst.

Franke, G. H. (1997). “The Whole is More than the Sum of its Parts”: The Effects of Grouping and Randomizing Items on the Reliability and Validity of Questionnaires. *European Journal of Psychological Assessment, 13,* 67-74.

Frick, A., Bächtiger, M.-T., & Reips, U. (2001). Financial incentives, personal information and dropout in online studies. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet Science* (pp. 209-219). Lengerich: Pabst.

Göritz, A. S., & Stieger, S. (2008). The high-hurdle technique put to the test: Failure to find evidence that increasing loading times enhances data quality in Web-based studies. *Behavior Research Methods, 40,* 322-327.

Hornke, L. F., Wosnitza, M., & Bürger, K. (2013). SelfAssessment: Ideen, Hintergründe, Praxis und Evaluation [SelfAssessment: Ideas, Backgrounds, Practical Experience and Evaluation]. *Wirtschaftspsychologie, 15,* 5-16.

Khorramdel, L., & Frebort, M. (2011). Context Effects on Test Performance. What About Test Order? *European Journal of Psychological Assessment, 27,* 103-110.

Khorramdel, L., & Maurer, M. (2012). Das Wiener Studieneignungs-Persönlichkeitspotenzial [The Viennese personality potential inventory for study aptitude]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment:Theorie und Konzepte* (pp. 103-118). Lengerich: Pabst.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Kubinger, K. D. (2015). Kritische Reflexion zu Self-Assessments im Rahmen der Studienberatung [Critical Reflection of Self Assessments in the Context of Study Counselling]. *Das Hochschulwesen, 63,* 76-80.

Kubinger, K. D., & Ebenhöh, J. (2011). *Arbeitshaltungen (AHA)* [Work styles], Version 27. Mödling: Schuhfried.

Kubinger, K. D., & Ebenhöh, J. (2012). Experimentalpsychologische Verhaltensdiagnostik des Anspruchsniveaus: Der Untertest Symbole Kodieren aus den „Arbeitshaltungen“ [Experimental-based behavior assessment of the aspiration level: The symbol-coding subtest of the „Arbeitshaltungen“]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment:Theorie und Konzepte* (pp. 93-101). Lengerich: Pabst.

Kubinger, K. D., Frebort, M., Khorramdel, L., & Weitensfelder, L. (Eds.) „Wiener Autorenkollektiv Studienberatungstests“ (2012). *Self-Assessment: Theorie und Konzepte* [Self-assessment: Theory and concepts]. Lengerich: Pabst.

Kubinger, K.D., Frebort, M., Khorramdel, L. & Weitensfelder, L. (2013). Prinzipien und Verfahren der Self-Assessments vom „Wiener Autorenkollektiv Studienberatungstests“ [Principles and inventories of the self-assessments from the „Viennese Author Collective for Study Aptitude Testing”]. *Wirtschaftspsychologie, 15,* 17-24.

Kubinger, K. D., Haiden, A., Karolyi, M., & Maryschka, C. (2012). Diagnostik des Lernstils: Der Lerntest LAMBDA [Assessment of the learning style: The learning test LAMBDA]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment:Theorie und Konzepte* (pp. 71-82). Lengerich: Pabst.

Liu, J., Allspach, J. R., Feigenbaum, M., Oh, H.-J., & Burton, N. (2004). A Study of Fatigue Effects from the New SAT®. College Board Research Report No. 2004-5. New York: The College Board. Retrieved from http://www.ets.org/Media/Research/pdf/RR-04-46.pdf [Feb 20, 2016].

Ortner, T. M. (2008) Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment, 16,* 249-257.

Penk, C. (2015). *Effekte von Testteilnahmemotivation auf Testleistung im Kontext von Large-Scale-Assessments* [Effects of test participation motivation on test performance in the context of large scale assessments]. Unpublished doctoral thesis, Humboldt-Universität zu Berlin.

Reips, U.-D. (2002). Standards for Internet-Based Experimenting. *Experimental Psychology, 49,* 243-256.

Ryan, J. J., Glass, L. A., Hinds, R. M., & Brown, C. N. (2010) Administration Order Effects on the Test of Memory Malingering. *Applied Neuropsychology, 17,* 246-250.

Schünemann, A. L. (2013). *Testreihenfolgeneffekte bei Persönlichkeits- und Leistungsdiagnostik in realen Auswahlsituationen: Ein Experiment in der Personalauswahl* [Test order effects regarding personality and ability assessment in genuine selection situations: A personnel selection experiment]. Unpublished diploma thesis, University of Vienna.

Testkuratorium (1986). Beschreibung der einzelnen Kriterien für die Testbeurteilung [Description of of particular criteria for test evaluation]. *Diagnostica, 32,* 358-360.

Tulsky, D. S., & Zhu, J. (2000). Could Test Length or Order Affect Scores on Letter Number Sequencing of the WAIS-III and WMS-III? Ruling Out Effects of Fatigue. *The Clinical Neuropsychologist, 14,* 474-478.

Undeutsch, N. (2012). Schlussfolgerndes Denken figural: Der Färbige Matrizentest [Figural reasoning: The colourful matrices test]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment: Theorie und Konzepte* (pp. 143-152). Lengerich: Pabst.

Wagner-Menghin, M. M. (2004). *Der Lexikon-Wissen-Test (LEWITE)* [The lexicon knowledge test]. Mödling: Schuhfried.

Weitensfelder, L. (2012). Test zur Angewandten Raumvorstellung [Test for applied spatial abilities]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment: Theorie und Konzepte* (pp. 181-195). Lengerich: Pabst.

Weitensfelder, L., Frebort, M., Müller, C. E., & Mitschek, B. (2011). *Processing Characteristics and Reasonableness regarding self-assessments.* Paper presented at the 11th European Conference on Psychological Assessment, Riga, August 31st - September 3rd 2011.

Weitensfelder, L., Undeutsch, N., Khorramdel, L., & Useini, C. (2012). Intrinsische Studienmotivation: Interesse, Erwartungen und Selbstkonzept eigener Fähigkeiten [Intrinsic study motivation: interest, expectancies and self-concept of own abilities]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment: Theorie und Konzepte* (pp. 119-142). Lengerich: Pabst Science Publishers.

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment, 10*, 1-17.

Zhu, J., & Tulsky, D. S. (2000). Co-norming the WAIS-III and WMS-III: Is There a Test-Order Effect on IQ and Memory Scores? *The Clinical Neuropsychologist, 14,* 461-467.