Disentangling Setting and Mode Effects for Online Competence Assessment

Ulf Kroehne [a]

[a] Educational Quality and Evaluation, German Institute for International Educational Research (DIPF)

Timo Gnambs [b, c]

[b] LIfBi | Leibniz Institute for Educational Trajectories

[c] Johannes Kepler University Linz

Frank Goldhammer [a, d]

[a] Educational Quality and Evaluation, German Institute for International Educational Research (DIPF)

[d] Centre for International Student Assessment (ZIB)

To be published as: Kroehne, U., Gnambs, T., & Goldhammer, F. (2019). Disentangling setting and mode effects for online competence assessment. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Education as a lifelong process* (2nd ed.). Wiesbaden, Germany: Springer VS.

Abstract

Many large-scale competence assessments such as the National Educational Panel Study (NEPS) have introduced novel test designs to improve response rates and measurement precision. In particular, unstandardized online assessments (UOA) offer an economic approach to reach heterogeneous populations that otherwise would not participate in face-to-face assessments. Acknowledging the difference between delivery, mode, and test setting, this chapter extends the theoretical background for dealing with mode effects in NEPS competence assessments (Kroehne & Martens 2011) and discusses two specific facets of UOA: (a) the confounding of selection and setting effects and (b) the role of test-taking behavior as mediator variable. We present a strategy that allows the integration of results from UOA into the results from proctored computerized assessments and generalizes the idea of motivational filtering, known for the treatment of rapid guessing behavior in low-stakes assessment. We particularly emphasize the relationship between paradata and the investigation of test-taking behavior, and illustrate how a reference sample formed by competence assessments under standardized and supervised conditions can be used to increase the comparability of UOA in mixed-mode designs. The closing discussion reflects on the trade-off between data quality and the benefits of UOA.

Keywords: Education · Panel study · Online Testing · Computer-based competence test · Mode effects · Paradata · Test-taking Behavior

Trennung von Effekten des Settings und des Modus bei Onlinekompetenztests

Zusammenfassung

Viele großangelegte Assessmentprogramme wie das National Bildungspanel führen neue Testdesigns ein, um die Antwortraten und die Messgenauigkeit zu verbessern. Insbesondere bietet unstandardisiertes Online-Assessments (UOA) eine ökonomische Möglichkeit, um heterogene Bevölkerungsgruppen zu erreichen, die ansonsten nicht an direkten Testung teilnehmen würden. Unter Berücksichtigung des Unterschieds zwischen Testauslieferung, Testmodus und Testsetting erweitert dieses Kapitel den theoretischen Hintergrund für den Umgang mit Moduseffekten in der Kompetenztestung des Nationalen Bildungspanels (NEPS; Kroehne und Martens 2011) und diskutiert zwei spezifische Facetten von UOA: a) Die Konfundierung von Selektionseffekten und Effekten des Testsettings und b) die Rolle des Testbearbeitungsverhaltens als Mediatorvariable. Wir stellen eine Strategie vor, die die Integration von Ergebnissen aus UOA in Ergebnisse computerbasierter Kompetenztestung ermöglicht und welche die Idee des Motivationsfilterns verallgemeinert, das für die Behandlung von schnellem Rateverhalten in Low-Stakes-Assessments bekannt ist. Dabei wird insbesondere der Zusammenhang zwischen Paradaten und der Erforschung von Testbearbeitungsverhalten hervorgehoben. Es wird gezeigt, wie eine Referenzstichprobe mit Kompetenztestung unter standardisierten und überwachten Testbedingungen verwendet werden könnte, um die Vergleichbarkeit von UOA in Mixed-Mode-Designs zu verbessern. Die abschließende Diskussion reflektiert den aus dem Vorgehen resultierenden Kompromiss zwischen Datenqualität und den Vorteilen von UOA

.

Schlüsselwörter: Bildung · Panelstudie · Online Testung · Computerbasierte Kompetenztests · Mode Effects · Paradaten · Testbearbeitungsverhalten

Disentangling Setting and Mode Effects for Online Competence Assessment

10.1 Introduction

The National Educational Panel Study (NEPS) started with paper-based assessments but now uses different variants of technology-based assessment to measure the development of competencies across the life course (see chapter 4). The challenge of mode effects (see Kroehne & Martens 2011) in standardized testing conditions (e.g., paper-based vs. computer-based competence assessment embedded in computer-assisted interviews, CAPI) is met with cross-mode studies making use of random assignment of test takers to different modes. Experimental mode effect studies are designed to create valid comparisons regarding the mode while keeping other factors such as the testing conditions constant. This permits the investigation of mode differences regarding measurement invariance based on the assumption of random equivalent groups (see, e.g., Buerger et al. 2016), or invariant items (e.g., Heine et al. 2016).

This chapter extends the theoretical framework for the treatment of mode effects in NEPS competence tests administered under standardized and supervised conditions (Kroehne and Martens 2011) to also cover online testing. Thus, we present a proposal on how to integrate data collected in *online assessments* (i.e., educational tests embedded in computer-assisted web interviews, CAWI). Online assessments of educational tests can be characterized by many *U* words:[1] *u*nstandardized (concerning the test setting) and *u*nsupervised (concerning the absence of an interviewer or a test administrator). These two main characteristics emphasize that online assessments are typically answered using *u*ndefined hardware (e.g., any web-enabled device with any screen size and input method) and with *u*ser-selected software (e.g., the test takers' favorite browser can be used), accompanied by *u*nknown test-taking behavior and *u*nobserved selection and dropout processes. Moreover, these assessments are not only *u*nsupervised in the sense that no supervisor is present who offers at least limited support during the assessment, but also *u*nproctored, meaning that there is no monitoring of test security. Accordingly, online assessments of competencies represent unstandardized and unsupervised computer-based test scenarios that, hereafter, will be referred to as *unstandardized online assessments* (UOAs). Whereas NEPS routinely uses online surveys in mixed-mode designs, the applicability of this approach to the delivery of competence assessments, which are already administered in computer-based form in many waves and starting cohorts, is not yet well understood. Consequently, the first UOA was introduced to NEPS in 2013 as part of an experimental mixed-mode design.

UOA can reach a large number of test takers as a delivery in which the operational effort don’t rise proportionally to the number of administered tests. Beyond reaching more test takers, UOA also allows participation of panel members who are hard to assess with other test deliveries (and vice versa). For instance, students undertaking a semester abroad can be reached only in personal interviews or group testing sessions at their home universities with (a relatively) immense effort. Mixed-mode designs with UOA seem particularly attractive regarding the costs for competence tests that were already implemented as computer-based assessment using “web technologies” (e.g., HTML). However, in mixed-mode designs, the coherent construct measurement across different assessment conditions is frequently questionable.

From survey research it is known that the trade-off between benefits and costs accompanying mixed-mode designs requires comparability studies and studies that investigate hypotheses about the potential causes of differences between assessment conditions (e.g., Jäckle et al. 2010). Accordingly, up to now, the UOA of competences in NEPS has been incorporated into experimental designs with random samples as control groups that were tested under standardized and supervised conditions (e.g., embedded in CAPIs as mentioned above or administered in supervised group testing conditions in educational institutions such as schools or universities).

This chapter introduces a general strategy for dealing with mixed-mode competence assessments in panel studies. We describe requirements to achieve comparability in mixed-mode designs from a psychometric point of view (in terms of potential mode and delivery effects) and with respect to the validity of the assessment (in terms of threats to the validity of interpretations of the score obtained from tests administered in different settings). The goal of this discussion is to outline how to achieve competence scores that are comparable across different assessments in mixed-mode designs, particularly when measuring change over time.

Therefore, we start with a detailed description of the empirical phenomenon of UOA in comparison to other methods used for the administration of competence tests in NEPS (10.2), followed by a discussion on the role of test-taking behavior when comparing UOA to other standardized test administrations (section 10.3). Subsequently, in section 10.4, we describe the general framework in which paradata (e.g., Kreuter 2013) are used to incorporate differences in response processes between assessments, including a brief review of the existing literature on selected criteria for evaluating test-taking behavior. In the closing section (10.5), we summarize limitations of the current framework as well as possible generalizations that could also include mobile assessments.

The chapter goes beyond existing literature on mixed-mode measurements by focusing explicitly on educational tests (instead of surveys or questionnaires) and by describing a framework that uses standardized as well as supervised assessments as a reference to achieve comparability of UOA. This allows us to distinguish delivery and mode effects that can be corrected using bridge studies (or other linking approaches) from differences in test-taking behavior that cannot be corrected without making strong assumptions regarding the fit of the underlying measurement models of the educational tests (e.g., Wise & DeMars, 2006).

10.1.1 Preliminary Remarks

By describing the theoretical background and a strategy for dealing with test-taking behavior in UOA, this chapter does not aim to favor or suggest a specific test delivery method for future assessments in NEPS. For sure, it also cannot replace survey papers and accompanying psychometric analyses of competence data in the various scientific use files. Moreover, the strategy described in this chapter, and, in particular, the criteria mentioned for filtering cases with conspicuous test-taking behavior in UOAs, are not necessarily suitable for the UOAs in NEPS. This requires further research to reasonably weigh the pros and cons. Nonetheless, this chapter does aim to provide a framework as a starting point that can – if used – deal potentially and to some degree with the lack of standardization of UOA.

In light of ongoing research on mixed-mode assessments of competencies, we hope that this framework can serve as a starting point for a fruitful discussion on UOA and how to achieve comparable measurements across different testing conditions. In time, these suggestions might be developed into a standard for the treatment of unstandardized and unsupervised assessment.

10.2 Investigating Online Assessment

10.2.1 Defining Unstandardized Online Competence Tests

This section deals specifically with UOAs used to administer competence tests in NEPS. As Table 1 reveals, competence tests in NEPS are administered in different modes (paper-based, PBA, and computer-based, CBA), embedded in different test settings (personal interviews, group testing, or unknown). Standardized competence assessments so far have been conducted while an interviewer (personal interview) or a test administrator (group testing) was present, either in the household or in different institutions such as schools or universities, with the interviewer or test administrator delivering the competence tests to the test takers.

--- Insert Table 1 here ---

The crucial features of *online assessment* are neither the web-based *delivery* nor the computer-based testing per se.[2] Instead, the central defining characteristic of UOA is the *test setting* at unknown locations that differ from standardized assessments conducted in groups or embedded in individual interviews. This results in a potential *setting effect* (see Frein 2011).

Standardization is a central part of the definition of competence assessment (e.g., Kraus et al. 2010). The lack of (experimental) control over the test place and the absence of an interviewer or test administrator in UOA can introduce additional construct-irrelevant variability compared to standardized conditions. Whereas this setting effect can be seen as part of the ecological validity in the context of psychological experiments (Reips 2000), it might threaten the validity of competence assessments (e.g., Barry & Finney 2009).

UOA also differs from traditional paper-and-pencil tests in terms of the *mode* (CBA vs. PBA). The mode is understood as a combination of multiple properties of an assessment, such as the medium, the input device, the format (portrait vs. landscape), possible feedback on the number of missing items, and other properties (see Kroehne & Martens 2011). However, additional factors beyond the *mode* can affect the comparability of assessments and contribute to the necessity of treating UOA cautiously.

In the remaining part of this subsection, we elaborate on these additional factors in detail, starting with apparent differences between competence assessments under standardized and supervised conditions and UOA. This will be followed by emphasizing possible differences in setting-specific (self-)selection processes that result in either complete participation or dropping off from an UOA. Subsequently, we close this section by pointing out the theoretical relationship between selection and setting effects.

10.2.2 Delivery Mode Differences

UOA as defined above is understood as administering test items in a browser-based environment, using identical items and identical implementations as used for supervised computer-based testing.[3] However, UOA differs regarding the following five apparent features from computer-based assessments under standardized conditions.

First, the identity of test takers is typically either completely unknown, meaning there is no identity security (called *open mode*, Bartram 2005), or the test is made available to known test takers only (called *controlled mode*). Human supervision can be achieved to some extent in so-called *online proctored testing* (Rios & Liu 2017). In open mode and controlled mode, there is no guarantee that only the designated test taker answers the test. A third person, such as a more capable conspirator, can influence the answers gathered in UOA. Moreover, test takers might use additional materials that are either unauthorized or at least not available under standardized testing conditions (e.g., Bloemers et al. 2016). Note, although for low-stakes assessments, no apparent reason exists to fake results beside impression management, many tests takers will do so anyway, given the opportunity (Steger et al. in press). The apparent difference is that standardized, and in particular, supervised assessments are conducted in the so-called *managed mode* (Bartram 2005) in which human supervision has control over the test-taking environment.

Second, tests administered in UOA can be answered at different locations, including the private home, the test taker’s workplace, and any public site such as trains, cafés, or all other areas that either provide Internet access or allow the use of private devices to access web pages. The place chosen by the test taker to answer questions or items in an online assessment represents a proxy for different properties that change along with the location.

Noise, distraction, the presence of colleagues, family members, strangers, and other characteristics of the specific setting vary with the situation chosen mainly by the test taker.

Third, both hardware (e.g., tablet, notebook with touchpad, or desktop computer with mouse and keyboard) and software (e.g., web browser) used to access the test material in UOA are chosen by the test takers in UOA, resulting in an additional source of heterogeneity that is neither construct-related nor of interest because it does not represent any interindividual differences regarding the measured construct. A possible approach to reduce this heterogeneity is the formulation of restrictive inclusion criteria (i.e., requirements concerning the devices allowed or the browsers supported for a particular study). Consequently, online assessments might require prerequisites (such as a desktop computer with minimal display size and Internet access) that might either exclude some test takers from participation (International Test Commission 2006) or, at least, impose an additional burden on them.

Fourth, UOAs can be answered at self-selected time points. Whereas supervised tests administered in groups at, for instance, schools or universities are often scheduled in advance requiring a strict timing, assessments embedded in individual interviews (e.g., CAPI) in respondents’ private homes are typically less restrictive, but still typically require arrangements between the test taker and the interviewer. The self-selection of testing time in UOAs (i.e., the time of day chosen to start the assessment) might lead to data that are gathered at times convenient for the test takers. The apparent difference is that UOA can result in test administrations at times of day that are not observed in standardized assessment in managed mode. Because the individually chosen time of testing might reflect individual differences in unobserved traits, testing time might also relate indirectly to the measured ability (e.g., Könen et al. 2015). Thus, the time of assessment might affect the comparability of standardized and supervised computer-based assessment and UOA. However, it is not necessarily the case that UOA is unrestricted concerning the time of day for participation. If announced properly, online test administration could easily be restricted to an eligible time window (e.g., between 6 a.m. and 10 p.m.), that would be more comparable to standardized testing.

Fifth, the social situation during test taking differs between the different test administrations summarized in Table 1. Effects of the social situation are known for interview-administered surveys and questionnaires in which the answering process differs from self-administered instruments (e.g., Klausch et al. 2013a). Moreover, as shown, for instance, in a meta-analysis by Gnambs and Kaspar (2015), a mode effect exists for items and issues that are conventionally perceived as sensitive topics. Beyond other factors, this effect might also be influenced by the presence of other test takers, as is the case in group-based test sessions. Moreover, differences in how test takers are recruited (e.g., an invitation via e-mail or in a telephone interview) and differences in the level of human supervision of the test sessions (Bartram, 2005) are considered to create different levels of commitment contributing to the social situation during testing. As discussed by Maddox (2017) for the computer-based assessment embedded in the interviews conducted for the Programme for International Assessment of Adult Competencies (PIAAC), the household creates a specific testing situation that is influenced by many factors. Although we are typically not able to quantify the impact of the social situation on the assessment results, UOA and assessment in the presence of an interviewer are expected to show systematic differences on this dimension of the test setting.

10.2.2 Setting-Specific (Self-)Selection

It is known that UOA versus supervised and standardized computer-based testing (either in individual or group settings) could result in mode-specific *response rates*. Indeed, the assumption that different people reply in different modes underlies the general idea of mixed-mode surveys (Klausch et al. 2013b). Everything else equal,[4] different response rates are considered to be an outcome of features of test deliveries and test setting that lead to different hurdles for participating in the assessment that, in turn, represent the consequences of underlying and unobserved decision processes. The resulting net effects regarding response rates might turn out to be higher for online assessments when factors that increase the probability of responding (such as the freedom to choose location and time point) dominate over factors that decrease this probability (such as the prerequisites for participation, e.g., the availability of a specific hardware).

It should be emphasized again that the test *setting* for UOA differs from standardized and supervised assessments in multiple ways. Hence, the specific phenomenon of the test setting incorporates not only multiple decision processes that might result in dissimilar selection biases for starting the assessment but also in setting-specific processes for ending the assessment and differences while taking the assessment.

As shown in Table 2, the decision processes in an UOA can be structured into three stages: (a) processes that result in the decision to participate in an assessment (*starting*), (b) processes that determine how and when the assessment is completed (*ending*), and (c) processes that influence the way in which the assessment is answered (*taking*).

--- Insert Table 2 here ---

Note that Table 2 is not exhaustive: Depending on the design of a study, the first stage (starting) might require the consideration of refusal rates and general participation rates concerning nonresponse errors. For simplification, we restrict the discussion of the online-specific aspect of non-response-related processes to the assessment of panel members by assuming that the online competence assessment is not the first contact with panel members who have already participated in a previous wave. Hence, the three stages are considered as part of a panel design for a particular cohort.

The selectivity of participation in an online administered test is a phenomenon that requires incorporating *time* in two ways: a longitudinal perspective of participation in different waves[5] and a short-term perspective of decisions to persevere in test *taking* instead of *ending* the assessment after it has started.

Starting an online delivered test is associated with lower costs than agreeing to be visited by an interviewer or arranging for participation in a group testing session. However, once a test taker overcomes the initial threshold for a standardized test setting, the social pressure to complete the test, at least on the surface, is much higher compared to unsupervised online delivery. UOAs provide more information about the decision process by giving access to incomplete data resulting from test takers who would probably not have overcome the threshold to participate in other assessment deliveries. Hence, even if more test takers drop out in UOAs, the data quality is not necessarily worse, because either more or different test takers participate. However, test takers might not only drop out more often but also answer questions differently. In other words, the question answering process might differ in UOA (e.g., de Leeuw et al. 2011). As we shall describe in the following, this represents a confounding of selection and setting effects.

10.2.3 Confounding of Selection Effects and Setting Effects

The delivery and the mode can be *randomly assigned* to test takers, for instance, by inviting panel members to participate either in a standardized and supervised CBA embedded in an interviewer delivered CAPI or an UOA including a competence test. Random assignment and careful experimental designs allow, for instance, an unbiased interpretation of the effect of the assigned delivery on comparable outcome measures (e.g., Jäckle et al. 2010). This line of reasoning could be used to compare the number of started test administrations according to some liberal criterion (i.e., test takers who at least start to read the instructions for a computer-based administered competence test, either UOA or integrated into a standardized and supervised setting). However, concerning the comparison of the measured competencies, the interpretation is limited by the fact that, for instance, the dropout behavior cannot be randomly assigned. Consequently, *selection effec*ts and *setting effects* are confounded (Klausch et al. 2013a).[6] This confounding was described by, for instance, Vannieuwenhuyze et al. (2011) for mixed-mode designs in which different types of respondents choose different modes (i.e., self-selection of modes, labeled by the authors as *measurement effect*). This confounding is supported by empirical examples. For instance, Preckel and Thiemann (2003) found items of an online-administered high potential intelligence test to be easier compared to a paper-and-pencil version. These differences could be explained by self-selection, motivation, and dropout rates. However, the different delivery-specific response and completion rates result in a similar confounding even under randomization. Differences (or similarities) between the outcomes can be caused by either differences between the sample compositions (due to selectivity) or differences in the way the instrument works (due to the setting).

10.3 Test-Taking Behavior

Dropout from a started assessment is an example of a setting-specific test-taking behavior that might create incomparable assessments if not acknowledged appropriately. As mentioned above, other examples range from using material or tools not available under standardized and supervised conditions (e.g., calculator or dictionary), searching the internet for solution-relevant information, or getting help from others. All of these have been discussed for unstandardized online assessment in the context of *cheating* (e.g., Lievens & Burke 2011; Bloemers et al. 2016; for meta-analytic evidence, see also Steger et al., in press).

For the experimental comparison of UOA and CBA under standardized and supervised conditions, test-taking behavior becomes a *mediator*. The notion of *mediator variables* (from research on causal inference) emphasizes the limitations of random assignment of test takers to specific test-taking behavior(s). What can be assigned is the test delivery (e.g., web-based as for UOA), and this delivery is associated with a particular test setting. However, the resulting test-taking behavior, such as the dropout tendency, is neither defined deterministically by the random assignment nor under experimental control once the delivery is assigned. Instead, test-taking behavior is the result of usually unobserved processes that are facilitated differently in different settings.[7]

10.3.1 Setting-Specific Behavior as Mediator

Bosnjak and Tuten (2001) classify response behavior on the two dimensions "Number of Displayed Questions" and "Number of Questions Answered" into seven different segments in web-based surveys.[8] For instance, test takers showing a response pattern with a high number of displayed questions and a low number of answered questions were labeled as *lurkers*, referring to a phenomenon generally observed in online communication (see, e.g., Sun et al. 2014). Similarly, one might take the number of not reached items[9] into account as a measure of test-taking behavior that is related to speed and ability (e.g., Goldhammer, 2015). If there is a higher tendency to take tests with a higher speed level in UOA, the number of not reached items should be lower and, thus, reflect a setting effect.

Response times also allows defining dropout at the item level as the number of not answered items after the last answered item when the time limit for a domain has not been reached. Dropout behavior in online assessments might reflect lower levels of commitment to the test (e.g., Reips 2000). Accordingly, if the proportion of test takers with a lower commitment is higher in UOA, dropout is expected to occur more often as a setting-specific response behavior.

Response times can also be used to describe test-taking behavior for completed tests. In particular, fast responses are used to identify rapid-guessing behavior (Schnipke & Scrams 1997) that is related to test-taking engagement (Wise & Kong 2005). Although Rios and Liu

(2017) found no difference between proctored and unproctored online assessment, the presence of test administrators was found to affect test-taking engagement (Lau et al. 2009). Hence, rapid guessing is expected to differ between UOA and standardized, and, in particular, supervised testing.

The dropout tendency and rapid guessing behavior are examples for test-taking behaviors for which it could be hypothesized that they transmit the effects of the independent variable (test setting) to the outcome variables (item responses). After conceptualizing setting-specific behavior as a mediator that is triggered only by the setting, it becomes essential to formulate theoretical expectations regarding the appraisal of test-taking behavior. For instance, available theoretical considerations, such as the assumption about the existence of lurkers in online assessments (Bosnjak & Tuten 2001) or the link between response time and test-taking effort (Wise & Kong 2005), can be used to derive indicators of specific test-taking behaviors.

10.3.2 Criteria for Comparable Behavior

The methodology to evaluate *measurement invariance* across mode effects (e.g., PBA vs. CBA, administered under identical conditions) and setting effects (CBA vs. UOA) can be applied to noncognitive measures with multiitem scales (e.g., Hox et al. 2015; Pajkossy et al. 2015) and cognitive measures such as competence tests (e.g., Buerger et al. 2016). The investigation of measurement invariance requires either items that are not affected by mode and setting of the test administration or the assumption of (random) equivalent groups.

Comparability concerning test-taking behavior, as a prerequisite for both approaches, can be achieved by generalizing approaches developed for the treatment of rapid guessing behavior. *Motivation filtering*, used by Wise et al. (2004), might make it possible to increase the validity of test score interpretations (see also Wise et al. 2006). Such *filtering* on rapid guessing as test-taking behavior was found to be superior to filtering on self-reported effort

(Rios et al. 2014). The simple idea is to use only those cases from UOA that show a comparable test-taking behavior to the standardized and supervised condition. Test takers with unusual behavior that is not observed in the standardized and supervised condition could be filtered. Remaining selection effects can be adjusted in a second step. Phrasing this in causal inference terminology, filtering could be applied to establish *common support* regarding the values of the mediator between the different test settings. As soon as test-taking behaviors overlap between test settings, different techniques, such as matching or conditioning can be used to adjust for the remaining differences in observed variables.

By imposing the requirement that only cases from UOA are used that show a test-taking behavior comparable to standardized and supervised assessments, we create a trade-off between the benefits of online assessment (more liberal filtering) and the interpretability of competence assessment in terms of standardization (stricter filtering). Furthermore, this conceptualization assumes that the test-taking behavior observed in a standardized and supervised assessment represents the valid standard. This might not necessarily be the case, if, for instance, rapid guessing occurs in standardized and supervised assessments. Then, motivation filtering should be applied to both the standardized and the unstandardized assessment, because it is known from previous research that rapid guessing threatens the validity of assessment results (e.g., Wise & DeMars 2005). Hence, if possible, thresholds for acceptable behavior should be derived like those obtained with different methods for rapid-guessing behavior (e.g., Kong et al. 2007). If this is not possible, the standardized test administration can be used as reference sample in the context of mixed-mode assessments (Fricker 2005; Vannieuwenhuyze et al. 2011). This justifies the idea of filtering (instead of weighting), because it makes it possible to exclude particularly test-taking behavior that was not found at all under standardized conditions. Note that choosing standardized and supervised settings as the reference might, in fact, manifest the bias. However, the choice of standardized and supervised settings seems justifiable because NEPS uses this kind of setting for the majority of competence assessments (see for a similar perspective, e.g., Russell & Hubley 2017).

The filtered UOA sample and the sample from standardized and supervised testing could either be used directly for further analyses, or remaining differences in additional variables (beyond indicators for test-taking behavior) could be adjusted using weighting, matching, or regression-based approaches.

10.3.3 The Importance of Paradata

The theoretical perspective described above requires the integration of two phenomena for investigating setting effects and establishing comparability of competence assessments between UOA and computer-based testing in standardized and supervised conditions: First, UOA attracts different test takers (i.e., the initial selection) with heterogeneous devices, varying internet connectivity, test taking at different times of day, and so forth. Second, test-taking behavior can vary between settings resulting in both: (a) more dropout in UOA and (b) different response processes in UOA that reflect, for example, differences in motivation, distraction, and honesty.

Paradata defined in a broader sense (e.g., McClain et al. 2018) can provide valuable information to account for both sources of differences between standardized and unstandardized testing. Indeed, paradata can be a "*way of identifying behaviours that might be relevant to response processes related to the construct and validity*" (Russell & Hubley 2017, p. 243).

Access-related paradata, in the form of device information (e.g., information provided in the "user agent string," see Callegaro 2010) can provide insights into, for instance, the relationship between the device type and higher probabilities of ending an online administered competence test ahead of time before reaching the last item. Access-related paradata such as connection speed, screen size, and the time required for scrolling can also explain interindividual time differences in UOA (e.g., Couper & Peterson 2017).

Response-related paradata such as timestamps collected for each answer change, can help to identify rapid-guessing behavior by flagging unmotivated responses that are presented faster than *solution behavior* would require. Similarly, an overall measure of test speededness, such as the number of not-reached items or the total testing time can be derived from response-related paradata that might help to identify speed-related differences between test settings.

Finally, *process-related paradata*, which incorporate all gathered raw log events of an assessment platform (e.g., Kroehne et al. 2016), can be used to derive indicators from paradata for specific test-taking behavior, such as *short-term interruptions* (see 10.4.3).

Robling et al. (2010, p. 10) suggested, that "*as global descriptions of data collection method can obscure underlying mode features, comparative studies should describe these features more fully.*" Similarly, the collection of paradata should be implemented as completely as possible without negatively impacting on the collection of substantive data, because until now, no standard for the collection of paradata exists.

10.4 Framework for Integrating UOA

In this section, we present a possible framework for integrating UOA into standardized and supervised comptence assessments.

10.4.1 Reference Sample

In NEPS, test administrations under standardized and supervised conditions present the current standard. Therefore standardized and supervised computer-based define the reference against which UOAs are compared. Up to now, NEPS has used UOA only in combination with standardized and supervised test settings. The implemented designs combined random assignment of respondents to different test administrations, but allowed respondents to switch from the standardized and supervised assessment to UOA if they chose to (self-selection). Accordingly, data from randomly assigned respondents can be used as the empirical reference sample. These data are not affected by individual mode preferences, but still reflect mode-specific response rates (see section 10.2.2).

The randomly selected test takers from the empirical reference sample (tested under standardized and supervised conditions) could be used to derive cutoff values for indicators that represent typical test-taking behavior under the current NEPS standard.[10] Respondents in UOA who fall outside these cut-offs are suspected of employing setting-specific test-taking behavior. In particular, a reference sample would be crucial for criteria that were not investigated previously, such as the interruption of test-taking.

10.4.2 Potential Criteria

Two approaches can be adopted to identify appropriate criteria to compare test-taking behavior between UOA and the computer-based standardized and supervised testing. The *top-down* approach follows theoretical reasoning on, for example, motivation and engagement, speededness and time spent in the assessment, nonresponse and dropout, cheating and aberrant responses, as well as test takers’ attention, and uses this reasoning to derive indicators for test-taking behavior. The top-down perspective emphasizes the need for theoretical justifications of the criteria used to benchmark test-taking behavior. Moreover, the selection of criteria allows the targeting of specific concerns of domain experts regarding the validity of online assessments.

The *bottom-up* approach focuses on the available paradata for a given competence assessment and aims to find observable indicators that allow a comparison of test-taking behavior between individual test takers. This bottom-up approach is conducted specifically for each UOA, because the gathered paradata are highly specific for the platform used to implement the computer-based assessment instrument (e.g., the CBA ItemBuilder, Rölke

2012). This bottom-up perspective permits adjustment of the procedure to unexpected behavior such as cases showing hints of technical abnormalities.

In the following, we present an overview of potential indicators that might be used to filter online cases from UOA with test-taking behavior that would not occur under standardized supervised conditions.

Short interruptions: In NEPS competence assessments, test takers are instructed to work on the assessment without interruption for 60 minutes.[11] Although it is possible that respondents take unexpected breaks (e.g., using the bathroom), in line with the instructions given to test takers, we have no substantive reason to assume that periods of inactivity should occur more often in UOA as compared to standardized assessment (using the identical software platform). Therefore, aberrant test-taking behavior in UOA can be expected to result in more and longer periods without any logged interaction (Sendelbah et al. 2016). From the log data, time intervals without any activity can be identified for each test taker that allow the creation of a filter to exclude these cases. However, filtering requires an appropriate threshold to consider the interruptions for a given test taker unusual (e.g., the threshold should be substantively longer than the expected maximum reading time, and test takers who are slow but motivated must not be excluded). A similar approach has already been presented for online surveys (Beckers et al. 2011; Stieger & Reips 2010). However, the thresholds of 5 minutes and 4 minutes used by the authors to exclude cases seem arbitrary. More recently, Sendelbah et al. (2016) used standardized time measures to derive cutoffs by incorporating the distribution of the indicator into the definition of thresholds. As the aim is to filter test takers from the online sample who show interruptions that do not occur under the standardized condition, we prefer deriving the cutoff value from the distribution of the indicator in the reference sample (i.e., by taking the reference sample as the norm and deriving the thresholds empirically). The sensitivity of the filtering approach to different cutoff values needs to be investigated empirically.

Focus detection: Leaving the current page in the web browser, as indicated by a focus detection (Diedenhofen & Musch 2017) could be interpreted as an additional hint of aberrant test-taking behavior or respondent multitasking, or at least an interruption of the test session. Relative to a threshold, the number of interruptions (i.e., the number of defocusing events; Diedenhofen & Musch 2017) could be used to filter test takers with conspicuous behavior.

Technical issues: In case of technical issues, such as interrupted internet connectivity, paradata might be generated. One specific consequence of UOA administered in controlled mode is the registration of *re-logins*. Moreover, long-term interruptions during online testing might also indicate technical issues on the server side (Sinharay et al. 2014, 2015). If a substantial amount of cases is affected by technical issues, filtering could be considered to improve the validity of the competence assessment.

Test speededness: The number of not reached items is expected to be identical between settings if self-paced test-taking is comparable concerning the speed–ability compromise (Goldhammer 2015). However, the duration (time spent on the test) was found to be higher for an online assessment (compared to paper-and-pencil testing; Bayazit & Askar 2012). Even though time is typically not included in mode effect comparisons due to the lack of timestamps from paper-based assessment (see, for an exception, Dirk et al. 2017), there is some evidence that test speededness differs within standardized and supervised settings between CBA and PBA (Bodmann & Robinson 2004; Kroehne et al. 2018). If this result is replicated for UOA even after filtering for rapid guessing behavior, speededness could be considered as a potential mediator of setting effects.

Missing propensity: Beyond the number of not reached items, also the number of omitted responses (and the propensity to omit items, e.g., Köhler et al. 2014) should be comparable between UOA as well as standardized and supervised conditions. *Lurkers*, for instance, defined as test takers with an unexpectedly high amount of omitted responses (i.e., a striking test-taking behavior characterized by viewing but not answering most items), could be considered for filtering to achieve comparability.

The possibility of using these indicators is strengthened by the availability of a reference sample (see 10.4.1), because currently "*the links between observed behaviours or patterns and underlying processes are speculative, and have not been explored directly"* (Russell & Hubley 2017, p. 234).

Rapid guessing: For some selected indicators, such as solution behavior in relationship to test-taking engagement, robust theories (e.g., Wise & Kong 2005, Wise 2015, Guo et al. 2016) and sound evidence from previous research (e.g., Lee & Jia 2014, Finn 2015, Goldhammer et al. 2016, Liu et al. 2015, Rios et al. 2017) are available allowing the derivation of thresholds that can be used without the need for a reference sample. Thus, taking into account the mode- and setting-specific response time distribution and the proportion of correct responses conditional on response time to create item-level thresholds (e.g., Wise & Ma 2012) would make it possible to apply motivational filtering to both the UOA sample and the reference sample.

10.4.3 Creating Comparable Ability Estimates

Ability estimates can be derived using data gathered under standardized assessment conditions as well as data from UOA. Within each setting, specific characteristics of the test-taking behavior are possible, and one test setting is not necessarily superior to another. Accordingly, unfiltered data could be used independently for the subsamples created by the randomly assigned or self-selected test delivery (standardized vs. online). However, as soon as ability estimates are to be used interchangeably, effects of the mode and setting should be taken into account.

Within each setting, for instance, within group testing sessions at universities, random assignment of test takers to modes can justify the assumption of random equivalent groups (Buerger et al. 2016). As discussed in this chapter, the treatment of mode effects cannot be adapted directly to adjust for setting effects when test-taking behavior mediates the setting effect. In particular, when a test-taking behavior (such as short interruptions) is observed only in one setting, strong assumptions would be required (extrapolation).

In this chapter, we generalize the idea of motivation filtering (Wise et al. 2004) as a first step before a potential treatment of mode effects. Filtering in this first step is expected to be most effective if implemented as liberally as possible. After controlling for differences in test-taking behavior, remaining differences in the sample composition can be corrected if necessary, for instance, by using weighting or matching techniques.

Filtering regarding test-taking behavior and possibly the additional adjustment for the sample composition result in groups that can be assumed to be equal concerning their competence. Subsequently, measurement invariance can be investigated, and at least construct equivalence should be established.

Finally, remaining dissimilarities in the test-taking behavior within test settings, for instance, interindividual differences in the number of not reached items as a measure of test speededness, could be included in the background model when estimating person parameters – an approach recently implemented in PISA (see, e.g., Heine et al. 2016).

10.5 Discussion and Outlook

In this chapter, we discussed treating test-taking behavior as a mediator for the effect of test settings on the results of assessments. The idea of generalizing the filtering approach, known for motivation filtering in low-stakes assessments, was a response to two main challenges: concerns about the validity of online assessments (lurking, rapid guessing, inattentive responding, use of additional material) and the need for an argument for creating random equivalent groups as the prerequisite for dealing with psychometric differences between settings.

Altogether, the framework introduces a trade-off between the benefits of online assessment (that might result in more data, including more incomplete test administrations and test takers who are harder to reach with standardized assessments) and the restriction to cases with test-taking behavior that is also observed under standardized testing conditions.

As illustrated with selected examples, hints for different test-taking behaviors can be found in additional data about the processes by which the survey and test data were collected (paradata). Accordingly, as soon as paradata are used to exclude cases (i.e., filtering), procedures for cleaning and validating paradata would be required to ensure data quality. Moreover, to foster the reproducibility of analyses and results, strategies for disseminating the information used from paradata should be developed that balance between the effort to create scientific use files (e.g., including indicators derived from paradata) and the research potential (e.g., the possibility of investigating new indicators). Disseminating indicators requires established measures (such as time and sequence of questions) that are of general use for investigating test taker behavior. This applies not only for cognitive measures, but also for survey data, because it would allow, for instance, an investigation of rapid guessing for noncognitive measures (e.g., Johnston 2016) or straightlining as response behavior in questionnaires (e.g., Kim et al. 2018). Providing raw log data rests not only on the availability of resources to anonymize and document them, but also on the tools that can be used by substantive researchers to analyze these kinds of data (such as the PIAAC Log- Data Analyzer, Goldhammer et al. 2017). Given both prerequisites, providing access to raw log data might be desirable because it would particularly make it possible to investigate methodological research questions such as the effect of technical problems and re-logins (e.g., Sinharay et al. 2014) on online assessments.

Previous work on the treatment of mode effects for competence tests (see Kroehne & Martens 2011) has been extended here to incorporate online assessments that are conducted under different, unstandardized test settings. This extension was necessary even for studies that use identical computerizations of items used in CAPI and UOA. Further research will be necessary as soon as ability estimates from different computerizations of instruments are compared (see, e.g., Bennett 2003), for instance, across cohorts. The extension described in this chapter provides a framework for dealing with low-stakes UOA. This includes studies conducted for instrument development. As Barry and Finney (2009) showed by comparing UOA and different standardizations of classroom testing, standardized test conditions are superior even for test development.

A major limitation of the described strategy to deal with UOA is that it focuses only on the psychometric modeling of mode effects after treating the potential confounding due to setting-specific test-taking behavior with filtering. A valuable extension in further research might particularly be to address the measurement of setting-specific attitudes, privacy concerns, and the perceived level of supervision in standardized conditions.

Incorporating differences in test-taking behavior as they occur between assessments conducted in different settings is also relevant for assessments obtained on mobile devices (Huff 2015, Illingworth et al. 2015, King et al. 2015). This is another area of future research. However, when screen sizes and display sizes are small, identical layouts, as assumed for the comparison between online assessment and computer-based testing are no longer possible.

An additional area for future research relates to the choice of the reference condition. The core idea of considering test-taking behavior as a mediator for the comparison of assessments between settings can be applied with different choices of a reference condition. The suggestion to exclude cases with unexpected test-taking behavior by using cutoff values derived from a reference administration should be understood as a pragmatic approach that is justifiable, particularly when the sample size of the online administered tests is much larger compared to the sample size gathered under standardized conditions. Further research is needed to develop more sophisticated techniques that will also overcome the arbitrary selection of one of the possible test settings used as the reference to derive cutoff values. Because the reference test setting might be the result of setting-specific selection behaviors as well, measures of representativeness, such as r indicators (Schouten et al. 2009, Shlomo et al. 2012), could be used to balance selection effects concerning the derivation of cutoff values.

Finally, further research might study the person fit across modes, bridging the gap between the measurement model used to scale competence tests and the answering behavior of test takers (Glas & Meijer 2003; Goegebeur et al. 2010; Sinharay 2015).

References

Barry, C. L., & Finney, S. J. (2009). Does it matter how data are collected? A comparison of testing conditions and the implications for validity. *Research & Practice in Assessment*, *3*, 1–15.

Bartram, D. (2005). Testing on the internet: Issues, challenges and opportunities in the field of occupational assessment. In D. Bartram & R. K. Hambleton, (Eds.), *Computer-based testing and the internet* (pp. 13–37). Chichester, England: John Wiley & Sons.

Bayazit, A., & Aşkar, P. (2012). Performance and duration differences between online and paper–pencil tests. *Asia Pacific Education Review*, *13*, 219–226.

Beckers, T., Siegers, P., & Kuntz, A. (2011, March). *Speeders in online value research*. Paper presented at the GOR 11, Düsseldorf, Germany.

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning* (ETS-RM-03-05). Princeton, NJ: Educational Testing Service.

Bloemers, W., Oud, A., & van Dam, K. (2016). Cheating on unproctored internet intelligence tests: Strategies and effects. *Personnel Assessment and Decisions*, *2*, 21–29.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research*, *31*, 51–60.

Bosnjak, M., & Tuten, T. L. (2001). Classifying response behaviors in web-based surveys. *Journal of Computer-Mediated Communication*, *6*(3).

Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The transition to computer-based testing in large-scale assessments: Investigating (partial) measurement invariance between modes. *Psychological Test and Assessment Modeling*, *58*, 597–616.

Callegaro, M. (2010). Do you know which device your respondent has used to take your online survey? *Survey Practice*, *3*, 1–12.

Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review*, *35*, 357–377.

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology*, *106*, 639–650.

de Leeuw, E., Hox, J., & Scherpenzeel, A. (2011). Mode effect or question wording? Measurement error in mixed mode surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association* (pp. 5959–5967). Alexandria, VA: American Statistical Association.

Diedenhofen, B., & Musch, J. (2017). PageFocus: Using paradata to detect and prevent cheating on online achievement tests. *Behavior Research Methods*, *49*, 1444–1459.

Dillman, D. A. (2000). *Mail and internet surveys: The total design method.* New York, NY: Wiley.

Dirk, J., Kratzsch, G. K., Prindle, J. P., Kroehne, U., Goldhammer, F., & Schmiedek, F. (2017). Paper-based assessment of the effects of aging on response time: A diffusion model analysis. *Journal of Intelligence*, *5*, 12.

Finn, B. (2015). *Measuring motivation in low-stakes assessments.* Research Report No. RR-15-19. Princeton, NJ: Educational Testing Service.

Frein, S. T. (2011). Comparing in-class and out-of-class computer-based tests to traditional paper-and-pencil tests in introductory psychology courses. *Teaching of Psychology*, *38*, 282–287.

Fricker, S. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*, 370–392.

Glas, C. A., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*, 217–233.

Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, *47*, 1237–1259.

Goegebeur, Y., De Boeck, P., & Molenberghs, G. (2010). Person fit for test speededness: Normal curvatures, likelihood ratio tests and empirical Bayes estimates. *Methodology*, *6*, 3–16.

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*, 133–164.

Goldhammer, F., Lüdtke, O., Martens, T., & Christoph, G. (2016). *Test-taking engagement in PIAAC*. OECD Education Working Papers 133. Paris, France: OECD Publishing.

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education: Research, models and instruments* (pp. 407–425). Cham, Switzerland: Springer.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*, 173–183.

Heine, J.-H., Mang, J., Borchert, L., Gomolka, J., Kroehne, U., Goldhammer, F., & Sälzer, C. (2016). Kompetenzmessung in PISA 2015. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme, & O. Köller (Eds.), *PISA 2015 Eine Studie zwischen Kontinuität und Innovation*, (pp. 383–540). Münster, Germany: Waxmann.

Hox, J. J., De Leeuw, E. D., & Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology*, *6*, 1–11.

Huff, K. C. (2015). The comparison of mobile devices to computers for web-based assessments. *Computers in Human Behavior*, *49*, 208–212.

Illingworth, A. J., Morelli, N. A., Scott, J. C., & Boyd, S. L. (2015). Internet-based, unproctored assessments on mobile and non-mobile devices: Usage, measurement equivalence, and outcomes. *Journal of Business and Psychology*, *30*, 325–343.

International Test Commission (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing*, *6*, 143–171.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, *78*, 3–20.

Johnston, M. M. (2016). *Applying solution behavior thresholds to a noncognitive measure to identify rapid responders: An empirical investigation*. PhD Thesis, James Madison University, Harrisonburg, VA.

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2018). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, *29*, 208–220.

King, D. D., Ryan, A. M., Kantrowitz, T., Grelle, D., & Dainis, A. (2015). Mobile internet testing: An analysis of equivalence, individual differences, and reactions. *International Journal of Selection and Assessment*, *23*, 382–394.

Kitchin, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*, 1–10.

Klausch, T., Hox, J. J., & Schouten, B. (2013a). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research*, *42*, 227–263.

Klausch, T., Hox, J. J., & Schouten, B. (2013b). Assessing the mode-dependency of sample selectivity across the survey response process. *Discussion Paper 2013-03*. The Hague, Netherlands: Statistics Netherlands (Available from https://www.cbs.nl/-/media/imported/documents/2013/12/2013-03-x10-pub.pdf).

Köhler, C., Pohl, S., & Carstensen, C. H. (2014). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement*, *75*, 1–25.

Könen, T., Dirk, J., & Schmiedek, F. (2015). Cognitive benefits of last night's sleep: Daily variations in children's sleep behavior are related to working memory fluctuations. *Journal of Child Psychology and Psychiatry*, *56*, 171–182.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*, 606–619.

Kraus, R., Stricker, G., & Speyer, C. (Eds., 2010). *Online counseling: A handbook for mental health professionals*. *Practical resources for the mental health professional*. Boston, MA: Academic Press.

Kreuter, F. (Ed., 2013). *Improving surveys with paradata: Analytic uses of process information*. Hoboken, NJ: Wiley & Sons.

Kroehne, U., Hahnel, C., & Goldhammer, F. (2018, April). *Invariance of the response process between modes and gender in reading assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Kroehne, U. & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft*, *14*, 169–186.

Kroehne, U., Roelke, H., Kuger, S., Goldhammer, F., & Klieme, E. (2016, April). *Theoretical framework for log-data in technology-based assessments with empirical applications from PISA*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education*, *58*, 196–217.

Lee, Y.-H. & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, *2*, 8.

Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program: Unproctored internet testing. *Journal of Occupational and Organizational Psychology*, *84*, 817–824.

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*, 79–94.

Maddox, B. (2017). Talk and gesture as process data. *Measurement: Interdisciplinary Research and Perspectives*, *15*, 113–127.

McClain, C. A., Couper, M. P., Hupp, A. L., Keusch, F., Peterson, G., Piskorowski, A. D., & West, B. T. (2018). A typology of web survey paradata for assessing total survey error. *Social Science Computer Review*, Online First.

Pajkossy, P., Simor, P., Szendi, I., & Racsmány, M. (2015). Hungarian validation of the Penn State Worry Questionnaire (PSWQ): Method effects and comparison of paper-pencil versus online administration. *European Journal of Psychological Assessment*, *31*, 159–165.

Preckel, F., & Thiemann, H. (2003). Online-versus paper-pencil version of a high potential intelligence test. *Swiss Journal of Psychology*, *62*, 131–138.

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

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*, 74–104.

Rios, J. A., & Liu, O. L. (2017). Online proctored versus unproctored low-stakes internet test administration: Is there differential test-taking behavior and performance? *American Journal of Distance Education*, *31*, 226–241.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *161*, 69–82.

Robling, M. R., Ingledew, D. K., Greene, G., Sayers, A., Shaw, C., Sander, L., Russell, I. T., Williams, J. G., & Hood, K. (2010). Applying an extended theoretical framework for data collection mode to health services research. *BMC Health Services Research*, *10*, 180.

Rölke, H. (2012). The ItemBuilder: A graphical authoring system for complex item development. In T. Bastiaens & G. Marks (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 344 – 353). Chesapeake, VA: AACE. Retrieved from http://www.editlib.org/p/41614

Russell, L. B., & Hubley, A. M. (2017). Some thoughts on gathering response processes validity evidence in the context of online measurement and the digital revolution. In B. D. Zumbo & A. M. Hubley, (Eds.), *Understanding and investigating response processes in validation research* (pp. 229–249). Cham, Switzerland: Springer.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*, 101–113.

Sendelbah, A., Vehovar, V., Slavec, A., & Petrovčič, A. (2016). Investigating respondent multi-tasking in web surveys using paradata. *Computers in Human Behavior*, *55*, 777–787.

Shlomo, N., Skinner, C., & Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, *142*, 201–211.

Sinharay, S. (2015). Assessment of person fit for mixed-format tests. *Journal of Educational and Behavioral Statistics*, *40*, 343–365.

Sinharay, S., Wan, P., Choi, S. W., & Kim, D.-I. (2015). Assessing individual-level impact of interruptions during online testing. *Journal of Educational Measurement*, *52*, 80–105.

Sinharay, S., Wan, P., Whitaker, M., Kim, D.-I., Zhang, L., & Choi, S. W. (2014). Determining the overall impact of interruptions during online testing. *Journal of Educational Measurement*, *51*, 419–440.

Steger, D., Schroeders, U., & Gnambs, T. (in press). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment*. Manuscript accepted for publication.

Stieger, S., & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior*, *26*, 1488–1495.

Sun, N., Rau, P. P.-L., & Ma, L. (2014). Understanding lurkers in online communities: A literature review. *Computers in Human Behavior*, *38*, 110–117.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2011). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, *74*, 1027–1045.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*, 197–210.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*, 237–252.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, California.

Wise, S. L. and Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Wise, S. L., & Ma, L. (2012, May). *Setting response time thresholds for a CAT item pool: The normative threshold method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver.

Wise, V., Wise, S., & Bhola, D. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*, 65–83.

Footnotes

[1] The similarity to *big data*, characterized with *V*'s (see, for instance, Kitchin and McArdle 2016) has been chosen carefully.

[2] Computer-based assessments are used routinely in NEPS in standardized settings, and online delivered tests can also be administered in standardized settings (e.g., Csapó et al. 2014).

[3] Apparent differences between modes – such as different layouts, question and task designs, and so forth in the sense of Dillman (2000) – were avoided (*unified design*) as far as technically possible.

[4] To achieve a meaningful comparison of response rates between deliveries (UOA vs. CAPI), the assumption that *everything else is equal* is crucial when taking into account the complete process of recruitment and invitation to an assessment. Depending on the design of a particular wave, different assessment modes might be combined. A combination used in one particular wave in NEPS is the mixture of CAPI for one random subsample of the cohort and a combination of CATI and UOA for the remaining subsample. The mixture of CATI and UOA incorporates two selection processes: participating in the CATI first followed by the decision to participate in the UOA. Taking both together, the sample composition for the assessment part of interest (i.e., the competence test administered in the CAPI and UOA delivery) is the result of two different selection processes that might best be described as one measurement point (CAPI) versus two measurement points (CATI and UOA). For the resulting samples, the assumption of random equivalent groups seems hardly justifiable without additional verifications and, if necessary, subsequent adjustments.

[5] Response rates, given a sample member has responded in a previous wave, correspond to attrition rates (if the unit nonresponse is a final dropout) or temporary dropouts. In waves with competence assessments, temporary dropout is equivalent to test refusal).

[6] Note that this is true if the random assignment of respondents to the delivery mode cannot be conducted after the recruitment (Jäckle et al. 2010) that serves as the decision to participate in a particular wave in a panel study.

[7] Test-taking behavior can be studied experimentally by, for instance, using different instructional sets, as often done to determine the limits on fakability of personality scales (see for a meta-analysis, Viswesvaran & Ones, 1999). Similarly, mediator variables can become treatment variables. However, when the test setting (and not the test-taking behavior) is randomly assigned, the values of the mediator are only observed variables.

[8] Complete Responders, Unit Nonresponders, Answering Dropouts, Lurkers, Lurking Drop-Outs, Item Nonresponders, and Item Nonresponding Dropouts.

[9] Competence tests are administered with time limits for each domain. Due to the time limits, it is possible to distinguish between omitted responses (i.e., unanswered items that are followed by answered questions) and not reached items (i.e., unanswered items that are not followed by an answered question in a test part due to the time constraint).

[10] Using the empirical reference sample allows us to apply the approach even if no normative threshold exists or the appropriateness of thresholds is in doubt (e.g., outdated, derived for a different target population or different domain, etc.).

[11] “Für die ersten zwei Teile haben Sie jeweils 30 Minuten Zeit. Es ist nicht möglich, die Bearbeitung der Aufgaben zu unterbrechen und später fortzusetzen. Nehmen Sie sich deshalb bitte eineinhalb Stunden am Stück Zeit.” [For each of the first two parts, you have 30 minutes. It is not possible for you stop answering the tasks to take a break and continue later. So please reserve 1.5 hours time for the test.]

Table 1: Summary of test administrations used for competence tests in NEPS

Mode	**Test setting**	**Interviewer**	**Test place**	**Delivery**	**Standardized**
PBA	Personal interview	Yes	Household	Interviewer	Yes
PBA	Group testing	Yes	Institution	Test administrator	Yes
CBA	Personal interview	Yes	Household	Interviewer	Yes
CBA	Group testing	Yes	Institution	Test administrator	Yes
CBA	Online	No	Unknown	Web-based	No

Table 2: Examples for decision processes related to UOA in the three stages “starting,” “ending,” and “taking”

Stage	Examples
Starting	Coverage/proportion of the cohort that can participate Cost of participation/effort required for participation Perceived attractiveness of the assessment/expectancy and value …
Ending	Self-paced answering and the resulting number of not-reached items Short interruptions and the tendency to abandon the setting Test abortion/dropout (and costs regarding social desirability) …
Taking	Tendency to answer items or to omit responses (missing propensity) Compliance with instruction and directions given for the assessment Test-taking effort and motivation (tendency to show rapid guessing) ...