Z Erziehungswiss (2011) 14:169–186
DOI 10.1007/s11618-011-0185-4

ZfE

11 Computer-based competence tests in the national educational panel study: The challenge of mode effects

Ulf Kröhne · Thomas Martens

Abstract: Computerized competence tests promise a variety of advantages compared to paper–pencil delivered tests, for instance, increased test security, more information about test takers and the test-taking process, instant scoring, and immediate feedback. Moreover, new innovative item types can be administered to broaden the test content. Three benefits should be particularly emphasized for the assessment of cognitive competencies in the German National Educational Panel Study. First, reductions of test time can be obtained through the higher measurement efficiency of adaptive tests. Second, computerized testing is expected to enhance standardization and to increase test takers' interest in completing the test. Third, Internet-based assessment offers the opportunity to deliver tests to spatially distanced test takers. However, before we can exploit these opportunities, we have to study the equivalence between different test administrations in order to maintain comparability of test scores and to ensure the validity of score interpretations. In this chapter, we shall describe a theoretical framework of mode effects and discuss various properties of test administrations. We shall relate the resulting equivalence criteria to the specific settings of the National Educational Panel Study in which (a) the usage of computerized competence tests is being prepared for upcoming assessments, and (b) tests for different grades and age groups are being designed to assess competence development over the life span.

Keywords: Education · Panel study · Computer-based competence test · Adaptive testing · Mode effects

Computerbasierte Kompetenztests im Nationalen Bildungspanel: Herausforderung durch Mode Effects

Zusammenfassung: Im Vergleich zu Papier- und Bleistifttests versprechen computerisierte Kompetenztests eine Vielzahl von Vorteilen, beispielsweise eine erhöhte Testsicherheit, mehr Informationen über die Testteilnehmer, sofortiges Scoring und unmittelbares Feedback. Zudem

Dr. U. Kröhne (✉) · Dr. T. Martens
Department of Educational Quality and Evaluation,
German Institute of International Educational Research (DIPF),
60486 Frankfurt a. M., Germany
e-mail: kroehne@dipf.de

Dr. T. Martens
e-mail: martens@dipf.de

können neue, innovative Aufgabenformate angewendet werden, um die Testinhalte zu erweitern. Drei Vorteile sind für die Messung kognitiver Kompetenzen im Rahmen des Nationalen Bildungspanels besonders hervorzuheben: Erstens kann eine Reduktion der Testzeit durch die höhere Messeffizienz von adaptiven Tests erzielt werden. Zweitens ist zu erwarten, dass computerisiertes Testen die Standardisierung erhöht und das Interesse der Testteilnehmer an der Testdurchführung steigert. Drittens ermöglicht die internetbasierte Testdurchführung die Auslieferung von Tests an räumlich entfernte Testteilnehmer. Bevor diese Vorteile jedoch genutzt werden können, muss die Äquivalenz zwischen den verschiedenen Formen der Testadministration untersucht werden, um die Vergleichbarkeit der Testergebnisse und die Validität der Ergebnisinterpretationen sicherzustellen. In diesem Kapitel wird ein theoretisches Bezugssystem für Mode Effects beschrieben und spezifische Eigenschaften der verschiedenen Administrationsformen werden diskutiert. Darüber hinaus werden abgeleitete Äquivalenzkriterien im Hinblick auf die Gegebenheiten der Kompetenzdiagnostik im Nationalen Bildungspanel betrachtet unter denen a) die Nutzung computerisierter Kompetenztests für nachfolgende Testdurchführungen vorbereitet wird und unter denen b) die Kompetenzentwicklung über die Lebensspanne mit Tests für verschiedene Klassenstufen und Altersgruppen gemessen wird.

Schlüsselwörter: Bildung · Panelstudie · Computerbasierte Kompetenztests · Adaptives Testen · Mode Effects

11.1 Introduction

When studying development and change in the National Educational Panel Study (NEPS), measurement instruments such as competence tests should remain as comparable as possible between different measurement points. Domain-specific cognitive competencies (see Chap. 5, this volume) are being assessed with paper-based tests in the first cycle of the NEPS. Therefore, the appropriateness of computerized assessment for upcoming cycles has to be justified by cost-benefit considerations incorporating the effort to study mode effects and to investigate test equivalence. A variety of benefits of computerized tests might be relevant for the assessment of competencies in the NEPS: recording information about the test-taking process (e.g., item latency rates and response times), increasing test security and enabling instant scoring (Bugbee 1996), enhancing standardization of the test-taking process and permitting immediate feedback (Domino and Domino 2006), as well as increasing test takers' interest in completing the test (Pomplun et al. 2006).

In particular, substantial benefits for the NEPS are expected for adaptive tests, resulting in higher *measurement efficiency*, that is, in shorter tests with the same reliability or tests of the same length with higher reliabilities (e.g., Frey and Ehmke 2007). In this chapter, we focus on mode effects for items taken from existing NEPS competence tests that are being transferred to the computer and assembled to produce computer-based test forms. This means that we shall focus on computerized tests that are administered to achieve higher measurement efficiency compared to the currently implemented assessment procedure, while expecting to retain the same measurement validity and the same content as existing paper–pencil competence assessment (Green 1988). We shall discuss neither the computerized assessment of innovative items such as performance-based assessments of information and communication technologies literacy, which are expected to alter the tests' measurement validity and content, nor the measurement of constructs that can-

not be assessed with paper-based tests (e.g., electronic reading or typing on a computer keyboard).

11.1.1 Test delivery

For a theoretical framework of mode effects, the two terms paper–pencil tests (PPT) and computer-based tests (CBT) refer to the basic distinction between tests presented with a paper-based delivery and a computer-based delivery. Ideally, given that all other possible influences are constant, only the pure *medium of administration* (MOA) should differ for PPT and CBT deliveries. However, in practice, the simple distinction between the two test delivery strategies is only a convenient way to communicate a conglomerate of differences between the actual test administrations that might trigger *mode effects* and result in nonequivalent test administrations. A systematic consideration of the different properties of test administrations is necessary before we can discuss differences due to the MOA.

A huge amount of literature deals with mode effects for questionnaires and achievement tests. For instance, Russell et al. (2003) have reported studies indicating that using computers as a medium for assessment can have significant mode effects (see also Choi and Tinkler 2002). Other studies, in contrast, have found either no differences due to the MOA or generally only small effects that are claimed to be of no practical significance (e.g., Mazzeo and Harvey 1988). To explain these contradictory empirical findings on CBT and PPT equivalence, we shall now extend the theoretical framework to three further areas beyond the test delivery that might be influenced by the equivalence of two different test administrations.

11.1.2 Test assembly

Empirical comparisons between test deliveries are often performed for test forms with different measurement efficiencies (e.g., Kolen and Brennan 2004), because increased efficiency is one of the major advantages of computerized testing and accurate measurement is always preferable to inaccurate measurement (Bodmann and Robinson 2004). Nevertheless, measurement error is not always the primary focus in large-scale assessments (Bugbee 1996). For instance, Adams (2005) has clarified why (individual) reliability is not necessary for the estimation of *population parameters* for large-scale assessments like PISA. For other measurement contexts like the NEPS, an important concern is the measurement of individuals' competencies over time. For these *individual trajectories*, measurement error should be as low as possible for each member of the panel. Therefore, reduced testing time for the NEPS cannot be achieved with larger sample sizes and the theoretical framework of mode effects for the NEPS should include different test assembly strategies (also called test designs).

The most common test assembly strategy is linear *fixed testing (FIT)* for which a (sub)set of items is combined to a fixed form that is administered in the same order to each test taker. FITs are sometimes categorized either as *flat* or *peaked* (Mead and Drasgow 1993). Flat tests contain items of many difficulty levels, whereas peaked tests are constructed to measure only a small range of the competence with a high level of accuracy. To achieve high measurement accuracy, each test should be peaked around the

test taker's true abilities. When developing PPT, this is often approximated by matching a test's item difficulties with the expected competence distribution in the population of test takers. When assessing students' competencies, this might be implemented with the help of different (peaked) test forms for different school types. If no appropriate variable for the selection of a peaked test form is available, *branched testing* (Cleary et al. 1968), *flexi-level testing* (Lord 1980) and *multistage testing* (e.g., Mead 2006) are some of the less well-known approaches that could be considered in order to increase the measurement efficiency of NEPS competence assessment. Especially for the intended longitudinal assessment, ability estimates obtained in previous assessments might be used to select more peaked test forms in subsequent assessments in line with the test takers' predicted competence.

Nevertheless, the highest measurement efficiency will be obtained with *computer adaptive testing* (*CAT*), an assembly strategy that incorporates the test takers' responses when adapting the level of difficulty in the administered test. This is done by estimating the test takers' competence during test administration in order to permit a selection of items according to the expected gain in accuracy (e.g., maximization of the Fischer information)—thereby resulting in a peaked test form. CAT based on the selection of individual items (Weiss 1982) is sometimes distinguished from CAT based on the selection of a bundle of items (*testlets*, Wainer and Kiely 1987). Recently (one-dimensional) CAT has been extended to *multidimensional adaptive testing* (MAT, e.g., Segall 1996) to increase measurement efficiency for the assessment of multiple competencies by taking into account their correlations (Frey and Seitz 2010).

Again, empirical comparisons between different test assembly strategies are often confounded with the effect of different test deliveries (e.g., Wang and Kolen 2001). Nevertheless, more peaked tests (i.e. tailored item difficulties) can be implemented with a broad variety of strategies—either PPT or CBT, and nonequivalence due to different test assembly strategies is sometimes considered separately (see, for a comparison of CBT and CAT, Schaeffer et al. 1995). Different test assembly strategies are included in a theoretical framework of mode effects because of the following three differences: The comparability of the item difficulties for individual test takers (that might impact, for instance, on the test takers' motivation), the context in which items are used (that might, for instance, generate item position effects), and the comparability of the item content for individual test takers (that we shall discuss in more detail below).

Ideally, different test assembly strategies are performed according to an underlying framework (*test blueprint*) in order to obtain *content validity* of different test forms (i.e., to achieve comparable item content, e.g., for different test forms with different competence levels; Wainer 2000). For CAT and MAT, comparability of test content is often achieved by restricting the "on-the-fly" *automated test assembly* with the help of content-balancing algorithms. If content comparability cannot be achieved algorithmically, hybrid approaches based on manually assembled test forms might be indicated (e.g., computer-adaptive sequential testing, Luecht and Nungester 1998, or multiple-form structure tests, Armstrong et al. 2004). Content validity is less critical for PPT as long as each test taker answers every item in an instrument. Nevertheless, comparability of test forms with respect to item content needs to be considered for PPT as well, because multiple forms

of an instrument will be assembled for different age groups and might be developed for a specific group (e.g., specific test forms for different school types).

11.1.3 Test scoring

Although effects due to different scoring approaches will not be discussed in full length here, they are related to the potential nonequivalence of different test administrations (Kolen and Brennan 2004). Accordingly, scoring is included in the framework of mode effects for the NEPS. In general, CBT enables automatic scoring for a variety of selected-response formats (e.g., multiple-choice) and for some simple constructed-response formats (e.g., short text answers, text highlighting, see Sireci and Zenisky 2006). For other response modes (e.g., complex essay scoring), automatic scoring is still under research (e.g., Haberman and Sinharay 2010). Accordingly, not all paper-pencil items are applicable for CBT with automatic scoring, and content validity of tests (i.e., the equivalence of the item content for CBT and PPT) might change when items for which automatic scoring is not feasible with an available software are excluded for CBT. Moreover, even if automatic scoring is technically possible, human scoring of, for instance, hand-written versus typed versions of the same answer for a constructed-response item might differ (Bennett 2003).

Nevertheless, for the intended use of computerized testing for the NEPS in which instant feedback is not necessary, CBT can also be implemented when responses are only recorded by the computer and subsequently scored by human raters. Even for CAT, the inclusion of manually scored items is possible, and higher efficiency is expected when items are administered adaptively, although the recorded answers will not help to update the CAT's competence estimate (Frey and Seitz 2010).

11.1.4 Test setting

Pomplun et al. (2006) illustrated that the conventional distinction between *group testing* and *individual testing* might change for CBT in a group setting, when, for instance, computerized tests are administered through headphones (i.e., compared to oral administration by examiners). In addition, the elimination of a possible *experimenter bias* was recognized early as an additional side effect of CBT (Styles 1991), because the structure of social interactions with the test administrator is changed as a consequence of the alternative test delivery. Hence, although differences in test settings are often observed together with various test deliveries, we might distinguish test delivery and test setting for a theoretical framework of mode effects for the NEPS. Moreover, for traditional PPT, different test settings are usually realized, for instance, for different age groups (Pomplun 2007). With respect to Internet-based assessment, the additional variability of the non-standardized test settings needs to be considered for the theoretical framework.

We conclude that different test deliveries are only one possible source of nonequivalence in test administrations. Along with the conversion of tests from PPT to CBT, various changes might occur to the test assembly strategy, the test scoring applied, and the general test setting.

11.2 Sources of mode effects

We shall now disentangle various properties of test administrations and focus more specifically on separate characteristics that are likely to differ between different test deliveries. The presentation of the main findings reported repeatedly in the literature is organized in two sections: Based on a discussion of possible definitions of mode effects, we shall start with a review of test administration properties as sources of mode effects, and then complement this with a selection of important characteristics of test takers that might either mediate or moderate mode effects.

11.2.1 Definitions of mode effects

The issue of mode effects refers to the underlying idea that the test administration has a *causal effect on*, for example, estimated competence (i.e., the outcome). Accordingly, mode effects might be defined as the difference between the *latent* competencies of a test taker for two tests administered in different modes. Comparable conceptualizations of the outcome might be important, for instance, a definition of mode effects as the *observed score* difference between different test administrations, as well as a definition of mode effects as the difference between, for example, item characteristic curves (outcome on the item level). Various definitions (latent versus observed, test level versus item level) will lead to different empirical criteria for test equivalence, and these are typically approached with different statistical procedures. This may also help to explain why findings regarding test equivalence are controversial (Noyes and Garland 2008).

Although mode effects are usually estimated on the basis of a sample of test takers, a meaningful definition refers to the differences between two different test administrations for an individual test taker. Accordingly, diverse mode effects can be expected when test takers are sampled from different target populations. For the NEPS, this may restrict the validity of empirical mode effect studies to a particular age group and cohort, and the generalizability of specific findings should be considered in the theoretical framework of mode effects.

Mode effects can be studied *experimentally* using either *within-group* or *between-group designs*. Single group designs with only one distinct order of test administrations (unbalanced designs) suffer from their inability to disentangle order effects from mode effects. Random-equivalent group designs (i.e., between group designs) have less power to detect mode effects compared to balanced single group designs, but are robust against possibly asymmetric practice effects in the latter (Mazzeo and Harvey 1988). Without random assignment of different test administrations, quasi-experimental adjustment methods are necessary to analyze mode effects (Puhan et al. 2007), because, for instance, better students might be more familiar with computers and thus choose the computer-based test administration (Kingston 2009).

Finally, all different definitions for the term mode effect mentioned above are formulated as the comparison of a test administration $X=x$ with a different mode of test administration $X=x^*$. For the NEPS competence assessment the already implemented PPT test administration is used as a reference for the investigation of mode effects.

11.2.2 Properties of test administration

In the following, we shall try to disentangle different components of the test administration by discussing the elements of different test administrations, subsumed by the simple treatment indicator *X* used in the definition above. This review of *properties of test administrations* will be presented in order to illustrate that mode effects of test administration are expected not only for a comparison of PPT and CBT, but also for a comparison of two different CBTs (or even PPTs) of a test.

11.2.2.1 Medium of administration (MOA)

With respect to mode effects, the most important property of test administrations seems to be the presentation media for item stimuli and related questions or tasks. As noted by Mead and Drasgow (1993, p. 450), "*it is possible to use a computer as an 'electronic page turner' for a conventional test by presenting items on the computer's monitor rather than on a piece of paper*." However, an impact of monitor and presentation quality (e.g., monitor size and resolution) on the cross-media comparability has been reported. Screen size itself does not have to be related to the amount of information displayed on a computer screen (Bennett 2003), but screen resolution influences the size of texts and graphics, and it should be related to the amount of information presented on the screen (Bridgeman et al. 2003).

11.2.2.2 Item layout

The computerization of existing PPT items can differ in terms of item layout and graphical item design (e.g., font family, size, and color). Empirical comparisons of different MOAs might be confounded by layout effects, simply because computerized test administrations enforce layout adaptations (Zhang and Lau 2006). However, different item layouts could also be compared within each MOA. Based on a comprehensive review of the literature, Mazzeo and Harvey (1988) have concluded that tests requiring multiscreen, graphical, or complex displays are more likely to result in mode effects.

11.2.2.3 Response mode

According to Sireci and Zenisky (2006, p. 332), "*the format of a test item encompasses all aspects of the specific task an examinee is to complete*," and these can be separated into the presentation of the *stimulus* and the *response mode*. Obviously, different *response actions* are required to give the same response either for PPT or CBT (e.g., ticking answers for multiple-choice items versus making a text-based response of some length). Mead and Drasgow (1993) have emphasized the differences in using a pencil instead of pressing a key. Moreover, innovative response modes (e.g., drag and drop, hot spot, and point and click) as well as different stimulus presentations (e.g., interactive scenarios or the integration of multimedia elements like audio or video clips) are only available for computerized test delivery. Threlfall et al. (2007) compared different response modes across the two MOAs with respect to the concept of "*affordance*" typically used to describe human-com-

puter interfaces (Greeno 1998). They found the opportunities to explore problems and test solutions afforded by some response modes had a strong impact on the comparability of CBT and PPT.

11.2.2.4 Input device

Different input devices used for the implementation of CBT (e.g., mouse, touchpad, keyboard, touchscreen, light pens, digital ink, joysticks, trackballs, or microphone for speech recognition) require different *response actions* to answer the same items (Parshall et al. 2000). Furthermore, some response modes might require additional within-item navigation such as steering the pointer to a text entry field—particularly when multiple items are presented on a computer screen.

11.2.2.5 Multiple items per page

Differences in *presentation characteristics* such as the number of items onscreen versus the number of items on a printed page are sometimes discussed as sources for incomparability of delivery modes (Bennett 2003). For computerized test delivery, items are often administered with only one single item presented onscreen at a time, whereas paper-based tests present multiple items on a single printed page (Schwarz et al. 2003). However, multiple items per screen or single items per page are also technically feasible.

11.2.2.6 Within-item navigation (scrolling)

The need to scroll through, for example, a long text passage displayed on a computer screen is known to cause difficulties and thus result in mode effects—particularly for passage-linked questions (Mazzeo and Harvey 1988) that are likely to be more difficult when the text passage and the questions cannot be seen on the screen without scrolling. In the same way as page breaks between stimulus and response might alter the properties of items in PPT, mode effects caused by scrolling, and in particular horizontal scrolling, should be avoided whenever possible (Kingston 2009).

11.2.2.7 Within-test navigation

In PPT, students have at least some control over the order in which tasks are dealt with, the sequence in which items are answered, whether questions are deferred or not, and the way answers are reviewed and changed. In CBT, options for within-test navigation are not always available due to different implementations of the test environment. Therefore, we might expect mode effects for different CBT due to the following two properties: (a) *Item review*. To mimic the typical within-test navigation behavior of PPT, item review is often implemented in CBT with linear fixed tests (e.g., Schwarz et al. 2003) or at least in sections of the test (Pommerich 2004). Nevertheless, a number of operational CAT programs do not allow participants to review or change previous responses at all (e.g., because of an increase in testing time, or to avoid psychometric complications, Glas 2006). (b) *Omitting items*. For PPT it is possible to adapt the order in which items are answered, deferred,

omitted, or skipped. Mode effects have been found for CBT when omitting items was not possible (e.g., Lee et al. 1986), and, correspondingly, no mode effects were found for CBT that allow participants to defer, review, and change answers (e.g., Lunz and Bergstrom 1994). In general, test takers tend to omit more items under CBT than under PPT (e.g., Ito and Sykes 2004). Moreover, Pomplun et al. (2006) have found that the *usage* of the opportunity to omit items can be a *response style* and therefore be a differential factor for mode effects. Including review options in the implementation of CBT is expected to enhance the comparability of CBT and PPT test scores (cf., e.g., Spray et al. 1989).

11.2.2.8 Time and speed

Mead and Drasgow (1993) have reported that comparability was spoiled in most speeded tests. Several studies have shown that the time required for completing CBT differs from that required for PPT (cf. van de Vijver and Harsveld 1994). Moreover, different technical implementations of within-test navigations for CBTs may result in different testing times. For instance, CBTs with review options are likely to take more time than CBTs without review options (but not necessarily any longer than PPTs without restriction of within-test navigation, see Vispoel 2000). Moreover, different response modes of test administrations become more important when tests became more speeded (Pomplun et al. 2002). Different CBT versions of the same test might differ with respect to item-level time restrictions and visualizations of the remaining number of items or time. Additionally, whereas a PPT might require test takers to wait before moving on to the next section of a test (Bennett 2003), a CBT version might permit students to proceed whenever they are ready. For tests with reading passages, it is known that reading on a computer screen is not only more difficult (Bugbee 1996) but also slower (Pomplun et al. 2002) compared to PPT. Hence, CBT may be more speeded than PPT, and time limits need to be specified carefully for each test administration (Greaud and Green 1986) to reduce mode effects.

Various further properties of test administrations can be described and considered as potential sources of mode effects. For instance, CBTs might differ with respect to the given *instructions,* whether a specific *tutorial* is implemented to familiarize respondents with the software environment or not, and with respect to the availability and implementation of *help options* for test takers.

The sources of mode effects presented here exemplify that a distinction between CBT and PPT is not sufficient for a theoretical framework of mode effects. Instead, various properties of test administrations are necessary to describe the implemented tests completely, because "*mode of administration effects appear to be very complex, and likely depend on the particular implementation of the testing program*" (Kolen and Brennan 2004, p. 317). Although some of the discussed properties might be specific to CBT whereas others seem to be natural for PPT, tests within the same MOA can be nonequivalent when implemented differently (Thissen et al. 2007). Furthermore, the pure number of properties leads to the conclusion that mode effects cannot be predicted completely from previous research, which has focused mainly on the comparison of a conglomerate of different properties. Instead, empirical mode effect studies are necessary when changing characteristics of test administrations (Pommerich 2004).

11.2.3 Characteristics of test takers

We shall now briefly discuss the importance of test takers' characteristics in relation to mode effects. The inclusion of characteristics of test takers into the theoretical framework is especially important for the NEPS, because of their expected heterogeneity (within and between cohorts and age groups).

11.2.3.1 Moderation and mediation of mode effects

A specific property of a test administration might cause similar mode effects for all test takers (*direct effect*). Nevertheless, differences between PPT and CBT are sometimes found to be more related to test takers' characteristics than to properties of test administrations (Pomplun et al. 2006). As described by Wise et al. (1989), individual-difference variables might *moderate* the mode effects, and even small effects at the population level may have a substantial influence for some test takers. Moreover, mode effects of specific properties of test administrations might be mediated through test taker characteristics (*indirect effects*). From a practical point of view, the distinction between moderated and mediated mode effects is less strict. For instance, scrolling (i.e., within-item navigation) might have a direct mode effect, because the relevant information necessary to answer an item is less likely to be visible on the screen when scrolling is necessary. Scrolling might also have an additional indirect effect if possible frustrations in test takers triggered by the need to scroll reduce test motivation (Bennett 2003). Similarly, the possibility of navigating within a test or booklet may interfere with individual test-taking strategies. In particular, the mode effect of a review option has been found to be mediated by student level covariates (e.g., test anxiety, Vispoel 2000) and also to be moderated by test takers' competence (e.g., Vispoel et al. 2000).

11.2.3.2 Computer familiarity/computer experience

Individual differences in *computer-related skills* might influence mode effects when these skills are necessary for taking the test and answering the items. Accordingly, test takers with more computer experience, particularly more experience with the specific parts of the CBT, might perform differently compared to less experienced test takers. Examples for specific parts of the CBT include the method for selecting input controls (e.g., the use of a mouse or touchpad); the method for entering text (e.g., the possibility of entering mathematical formulas, Clariane and Wallace 2002); and the use of functions to review items, change answers, and navigate within the test. Computer-related skills might have direct effects (e.g., familiarity with using a mouse) or indirect effects (e.g., experienced students might feel more comfortable), but research is mixed on whether computer familiarity has a direct or indirect effect (Pomplun 2007). Inconsistent findings here might be explained by the different opportunities test takers have in advance to familiarize themselves with the CBT (Kingston 2009). Clearly, this emphasizes the importance of tutorials and help options. Computer familiarity might also be a moderator variable for the effects of different properties of the test administration. For instance, Zandvliet and Farragher (1997) found that students with less computer familiarity needed more time for the CBT. Finally,

computer experience might also explain differential mode effects found, for instance, with respect to test takers' s*ocioeconomic status* (MacCann 2006).

11.2.3.3 Further variables

Affective implications of CBT such as the effect of CAT on test takers' motivation due to the different test assembly strategy resulting in different *relative item difficulties* have been discussed for a long time (cf. Frey et al. 2009). *Test anxiety* and *computer anxiety* are also sometimes discussed as mediator variables (e.g., Wise et al. 1989). Moreover, with respect to the different response modes typing versus writing, Russell (1999) found that students with below-average keyboarding skills performed worse on CBT. Finally, it should be mentioned that mode effects are typically analyzed for moderator variables like test takers' gender, age, race/ethnicity, that is, for *demographic variables* that are also used routinely to analyze differential item functioning (DIF) during test development.

11.3 Consequences of mode effects

In general, CBT and PPT "*can be equivalent, but it is the responsibility of the test developer to show that they are*" (Bugbee 1996, p. 292). Hence, empirical mode effect studies focusing on the intended use of CBT are necessary. In preparation for these studies, which will be conducted on the basis of a selection of NEPS competence tests, we shall now review statistical criteria that are usually applied for comparisons of PPT and CBT. We shall structure the presentation as follows: We shall start with cross-mode correlations and the related criteria of dimensionality and validity. We shall then present a discussion of mean differences between test administrations and related criteria regarding the scale invariance of test administrations.

11.3.1 Dimensionality and validity

11.3.1.1 Cross-mode correlation

The correlation between PPT and CBT scores in within-subject designs should be as high as the test's reliabilities will allow, indicating that a competence is measured equivalently with two test administrations (Green et al. 1984). Latent correlations (or attenuation-corrected correlations) of 1.0 are expected for a perfect linear relationship between test administrations, and high cross-mode correlations are often interpreted as evidence for the desired equivalence (e.g., Gwaltney et al. 2008) or at least as failure to find evidence for *construct-irrelevant variance* (e.g., Gallagher et al. 2002).

11.3.1.2 Dimensionality

A more specific dimensionality criterion requires that PPT and CBT measure the same dimension to ensure *construct validity* on the item level (Wainer 2000). For the NEPS competence tests, this requirement implies that regardless of how the test is administered,

each item should at least fit the particular item response theory (IRT) model (i.e., that the relationship of each item to the measured construct is modeled appropriately). Obviously, the empirical analysis of this criterion of measurement invariance between test-administrations depends on the fit of the IRT model for the reference (PPT).

11.3.1.3 Content validity

Beyond perfect unidimensionality, items are often classified into content areas that may represent different but highly correlated dimensions (Wang and Kolen 2001). The comparability of tests with respect to content areas should be addressed because of the following three threads to content validity: First, item selection is the main mechanism for developing unidimensional tests (Green et al. 1984). Hence, one obvious strategy for dealing with single items revealing mode effects would be to exclude them from CBT as well. However, this might change the test content. Second, differences in test content can emerge when some PPT items cannot be computerized with the available test software, or when items that cannot be scored algorithmically are excluded from CBT. Third, different test assembly strategies will change the test content, if the assembling fails to take existing content areas into account. Therefore, an important prerequisite to ensure content validity (i.e., to avoid construct under representation) is the availability of a clear test specification (blueprint) that can be referred to, for example, when developing new items for CBT.

11.3.1.4 Structural relationships

When multiple constructs are considered, a second dimensionality criterion on the *test level* requires similar construct patterns. In other words, the *structural relationships* should not be influenced by properties of the test administration (Wainer 2000). Depending on the nature of the considered constructs, this criterion reflects the requirement of comparable construct or criterion validity. For within-group designs, multiple constructs can be analyzed with *multitrait-multimethod models* to quantify the variance due to different test administrations (cf., e.g., Wainer 2000).

Finally, a qualitative comparison of test administrations with respect to *construct-irrelevant test variance* might deliver valuable insights into test equivalence (see, for a nonpsychometric verbal protocol analysis, Kobrin and Young 2003).

11.3.2 Difficulty and reliability

11.3.2.1 Mean differences and score distributions

Scale invariance is often analyzed on the basis of mean differences (Bergstrom 1992), and cognitive tests often reveal overall mean shifts between test administrations at the test level (Mead and Drasgow 1993). Systematic effects of properties of the test administration on all items are removable, for instance, by transformations of the score scale (Hetter et al. 1994). Moreover, moderated mode effects (e.g., differences for test takers with specific characteristics) need to be investigated, and must be taken into account for possible transformations.

11.3.2.2 Item-by-mode interactions

Nonsystematic mode effects, that is, differential effects for only some items can be investigated by, for instance, comparing item parameter estimates obtained with PPT or CBT or by applying differential item functioning (DIF) analysis (e.g., Poggio et al. 2005). Nonsystematic item-by-mode interactions can be acknowledged by mode-specific calibrations if the items fit the IRT model under both test administrations (e.g., Choi and Tinkler 2002)—a strategy that is also sometimes suggested for DIF items as well (Thissen et al. 2007). Nevertheless, similar to the common test development practice, items with moderated mode effects should be excluded.

11.3.2.3 Equity and reliability

With respect to different test assembly strategies, *equity* is discussed as an equivalence criterion (Wang and Kolen 2001). First-order equity (i.e., equal expected scores conditional on the latent competence) might be violated when competence estimates are biased for one of the test administrations. Simulation studies might be useful to compare biases conditional on the competence level. Second-order equity (i.e., equal precision conditional on the latent competence) is expected to be violated when different test assembly strategies are implemented in order to increase measurement accuracy (Green et al. 1984). Furthermore, although different overall reliabilities are expected due to different test assembly strategies, internal consistency might be influenced by further properties of the test administration.

Further equivalence criteria (e.g., rank orders of individuals tested in alternative modes or equal probabilities of achieving passing scores) are sometimes discussed in the literature (see, for a summary, Wang and Kolen 2001).

11.4 Outlook

In this chapter, we addressed computerized testing based on the conversion of existing paper-pencil items tapping the cognitive competence domains to be implemented in upcoming cycles of the NEPS. From the various possible advantages of computerized testing, we focused particularly on the possibility of improving measurement efficiency (i.e., either lower measurement error or reduced test length) through more efficient test assembly strategies. In particular, shorter tests are important for the NEPS in order to reduce the strain on test takers and minimize panel mortality. We summarized that different test assembly strategies can be implemented for PPT and CBT, and that, for instance, measurement efficiency also might be increased by the incorporation of previous knowledge about test takers.

Different properties of test administrations that might lead to either direct or indirect mode effects were described, and characteristics of test takers were emphasized. Overall, the presented discussion highlighted the need for a careful implementation of computer-based assessment that takes account of various properties of tests such as within-item navigation, within-test navigation, and appropriate time limits along with various charac-

teristics of test takers such as necessary computer-related skills. Future research should focus on mode effects for *specific content areas* (e.g., for computerized reading tests), because some authors have suggested that score comparability might be test-specific (Ito and Sykes 2004).

Moreover, our theoretical consideration of mode effects clarifies the importance of a clear description and documentation of all properties of the test administration. This should be stored as metadata in the NEPS data warehouse (see Chap. 20, this volume).

We conclude with a look at future empirical mode effect studies that will relate the equivalence criteria presented here to the specific circumstances of the NEPS. CBT will either replace PPT or be used parallel to PPT, and the selection of equivalence criteria will depend on the intended use of the test instruments. In the NEPS, different instruments will be developed for the various age groups. When CBT is introduced in upcoming cycles, the parallel use of CBT and PPT for an instrument should be avoided to reduce the complexity of the required test equivalence. In more detail, this means that if CBT replaces a PPT for all test takers in a particular age group, cohort, and domain, only the following two main requirements need to be fulfilled: (a) cross sectional comparability of scores for the comparison of cohorts on the population level (interindividual) and (b) interchangeability of individual scores for the analysis of longitudinal trajectories and the computation of change scores (intraindividual).

Comparability of scores on the population level is necessary for the CBT and PPT version of a competence test addressing a particular domain and specific age group. As each panel member will take an instrument only once, interchangeability of individual scores is not necessary, that is, only marginal statistics of subpopulations need to be considered for the comparison of cohorts. This can be achieved by adjusting mean differences and the overall score distribution for potential mode effects (cf., e.g., Pomplun et al. 2002).

Interchangeability of scores on the *individual level* is required for two consecutive instruments of a specific domain in order to maintain comparability of intraindividual change scores over time. *Equating* is the statistical term used to describe the process of adjusting for differences among various test forms, so that scores on the forms can be used interchangeably (Kolen and Brennan 2004). Equating applies only if tests forms measure the same content with an equal degree of reliability. However, statistical approaches called *scaling to achieve comparability* or *calibration* (Linn 1993) can be used for tests of different reliability. In order to *measure change* in the NEPS, a similar statistical approach will be used to transform scores of consecutive age groups on a *common metric*. As mentioned in Chap. 5 in this volume, specific anchor item designs (i.e., with a sample of common anchor items used in two consecutive assessments) or additional linking studies (i.e., with a sample of test takers who are administered items at two consecutive assessments) will be implemented for this purpose. Depending on various factors such as the multidimensionality of the competence tests, different statistical methods are typically applied to generate robust *vertical scales*. Nevertheless, what all the different approaches have in common is that vertically scaled individual change scores will be influenced by additional statistical uncertainty, even when the change scores are based on two PPT test administrations. In line with this intended use of competence assessments for the intraindividual measurement of change, the equivalence between CBT and PPT (cross-modes)

needs to be comparable with the uncertainty of the link between consecutive assessments within one mode.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31,* 162–172.

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement, 28,* 147–164.

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning*. Princeton: Educational Testing Service.

Bergstrom, B. (1992, April). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16,* 191–205.

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.

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education, 28,* 282–299.

Choi, S. W., & Tinkler, T. (2002, April). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Clariane, R., & Wallace, P. (2002). Paper-based versus computer-bases assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33,* 593–602.

Cleary, A. T., Linn, R. L., & Rock, D. A. (1968). An exploratory study of programmed tests. *Educational and Psychological Measurement, 28,* 345–360.

Domino, G., & Domino, M. L. (2006). *Psychological testing: An introduction*. Cambridge: Cambridge University Press.

Frey, A., & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Messung von Bildungsstandards in Mathematik. In M. Prenzel, I. Gogolin, & H.-H. Krüger (Eds.), *Kompetenzdiagnostik* (Zeitschrift für Erziehungswissenschaft: Sonderheft 8, pp. 169–184). Wiesbaden: VS Verlag für Sozialwissenschaften.

Frey, A, & Seitz, N.-N. (2010). Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz. *Zeitschrift für Pädagogik, Beiheft 56,* 40–51.

Frey, A., Hartig, J., & Moosbruuger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. *Diagnostica, 55,* 20–28.

Gallagher, A., Bennett, R. E., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: Detecting construct-irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment, 8,* 27–41.

Glas, C. A. W. (2006). *Violations of ignorability in computerized adaptive testing*. (Computerized Testing Report 04-04). Newtown: Law School Admission Council.

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10,* 23–34.

Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 77–86). Hillsdale: Erlbaum.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347–360.

Greeno, J. G. (1998). The situativity of knowing, learning and research. *American Psychologist, 53,* 5–26.

Gwaltney, C. J., Shields, A. L., & Shiffman, S. (2008). Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. *Value in Health, 11,* 322–333.

Haberman, S. J., & Sinharay, S. (2010). The application of the cumulative logistic regression model to automated essay scoring. *Journal of Educational and Behavioral Statistics, 35,* 586–602.

Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1994). A comparison of item calibration media in computerized adaptive testing. *Applied Psychological Measurement, 18,* 197–204.

Ito, K., & Sykes, R. C. (2004, April). *Comparability of scores from norm-referenced, paper-and-pencil, and web-based linear tests for 4–12*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22,* 22–37.

Kobrin, J. L., & Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. *Applied Measurement in Education, 16,* 115–140.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices*. New York: Springer.

Lee, J., Moreno, K. E., & Sympson, J. B. (1986). The effects of mode of test administration on test performance. *Educational and Psychological Measurement, 46,* 467–473.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83–102.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillside: Erlbaum.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35,* 229–249.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement, 31,* 251–263.

MacCann, R. (2006). The equivalence of online and traditional testing for different subpopulations and item types. *British Journal of Educational Technology, 37,* 79–81.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional versions of educational and psychological tests: A review of the literature* (Research Report CBR 87-8, ETS RR 88-21). Princeton: Educational Testing Service.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19,* 185–187.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449–458.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*, 1352–1375.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. der van Linden & C. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129–148). Boston: Kluwer.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment*, *3*. http://www.jtla.org. Accessed 11 Nov 2009.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment*, *2*. http://www.jtla.org. Accessed 4 March 2010.

Pomplun, M. (2007). A bifactor analysis for a mode-of-administration effect. *Applied Measurement in Education, 20,* 137–152.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62,* 337–354.

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment, 11,* 127–149.

Puhan, P., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning, and Assessment, 6*. http://www.jtla.org. Accessed 19 Nov 2009.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7,* 20.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: a look back into the future. *Assessment in Education, 10,* 278–293.

Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computer-adaptive GRE General Test* (Research Rep. No. 95-20). Princeton: Educational Testing Service.

Schwarz, R. D., Rich, C., & Podrabsky, T. (2003, April). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331–354.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–348). Hillsdale: Erlbaum.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carl, J. E. (1989). Effect of medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26,* 261–271.

Styles, I. (1991). Clinical assessment and computerized testing. *International Journal of Man-Machine Studies, 35,* 133–155.

Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C. H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research, 16,* 109–119.

Threlfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics, 66,* 335–348.

van de Vijver, F. J. R., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computer versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79,* 852–859.

Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement, 60,* 371–384.

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement, 37,* 21–38.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A Primer*. Mahwah: Erlbaum .

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38,* 19–49.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.

Wise, S. L., Barnes, L. B., Harvey, A. L., & Plake, B. S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. *Applied Measurement in Education, 2,* 235–241.

Zandvliet, D., & Farragher, P. (1997). A comparison of computer administered and written tests. *Journal of Research on Computers in Education, 29,* 423–438.

Zhang, L., & Lau, C. A. (2006, April). *A comparison study of testing mode using multiple-choice and constructed-response items—Lessons learned from a pilot study*. Paper presented at the AERA annual conference, San Francisco.