Journal of Educational Psychology
2014, Vol. 106, No. 3, 639–650

© 2014 American Psychological Association
0022-0663/14/$12.00 DOI: 10.1037/a0035756

Computer-Based Assessment of School Readiness and Early Reasoning

Benő Csapó, Gyöngyvér Molnár, and József Nagy
University of Szeged

This study explores the potential of using online tests for the assessment of school readiness and for monitoring early reasoning. Four tests of a face-to-face-administered school readiness test battery (speech sound discrimination, relational reasoning, counting and basic numeracy, and deductive reasoning) and a paper-and-pencil inductive reasoning test were transferred to an online platform and administered at the beginning of school to samples of first-grade children (the sample sizes were between 364 and 435). Results of the original and the computerized tests were analyzed to explore (a) whether the new scales were identical to the original ones; (b) how the change of media influenced the reliability of the tests; and (c) whether the migration into a new medium affected gender differences. Analyses indicated that measurement invariance held in a strict sense in the case of the inductive reasoning test (the migration did not change the general look of the test or the item types) and only partially for the speech sound discrimination test (neither the item type nor the scoring principle was changed). Measurement invariance did not hold for the 3 remaining tests. In 3 tests—speech sound discrimination, relational reasoning, and deductive reasoning—the online versions demonstrated improved reliability. Only certain items of the numeracy test could be assessed on computer, and the reliability of the shortened test decreased. No differences were found between the 2 versions of the inductive reasoning test. Gender differences were explored for the speech sound discrimination test, and latent analyses indicated that measurement invariance did not hold. Girls' performance was somewhat better, similarly to former face-to-face assessments, where girls performed slightly better than boys. These results encourage further research on the extension of computer-based assessment to early childhood education.

Keywords: computer-based assessment, online testing, school readiness, inductive reasoning, early childhood assessment

A large number of studies have highlighted the importance of smooth preschool-to-school transition and the successful first years of schooling from different perspectives. Research has paid increasing attention to identifying the conditions of a successful start in schooling. Among these efforts, creating instruments for assessing school readiness and monitoring development at the beginning of schooling play an important role. A broad range of instruments, including observation protocols, tests, and test batteries, are available, which can be used to assess different aspects of general cognitive development as well as specific precursors of skills learners are expected to master at school. However, many instruments that have been proven valid and reliable under research or pilot conditions turn out to be too complicated to use regularly in schools. Sometimes they are not sufficiently precise if not used under standardized conditions or if not administered by specially trained teachers. In many cases, the time and human resources required to administer and score the tests prevent their frequent use. Technology-based assessment may solve these problems, but administering computerized tests to young children before or at the initial stage of formal schooling may raise a number of questions concerning the validity of results obtained through technology-based assessment in young children.

In this article, we explore the possibilities of online testing at the beginning of formal education by comparing traditional and digitized versions of five tests. Four of them are tests from the DIFER (DIagnosztikus FEjlődésvizsgáló Rendszer–diagnostic system for assessing development) school readiness test battery, an instrument with a long developmental history (Nagy, 1980, 1987; Nagy, Józsa, Vidákovich, & Fazekasné Fenyvesi 2004a, 2004b). The fifth is an inductive reasoning test prepared to measure learners' general mental ability. These five instruments measure different psychological attributes, and their computerized versions require different technological solutions. This variety of instruments offers a number of possibilities to analyze the prospects for and limitations of technology-based assessment around the time of the kindergarten-school transition. As we focus on the applicability of technology, we deal only in brief with the general functions of school readiness tests and other instruments used for monitoring children's development during the first school years.

Assessment of School Readiness and Early Development

Mastery of basic literacy and numeracy skills is the main goal of the first school years; therefore, school readiness tests are often composed of tasks that measure precursors to speaking skills, vocabulary, early reading, writing, counting, computing, reasoning (comprehending relations and inferential processes), and the elements of behavior and social skills (attention, following instruc-

This article was published Online First February 17, 2014.

Benő Csapó, MTA-SZTE Research Group on the Development of Competencies, University of Szeged, Szeged, Hungary; Gyöngyvér Molnár and József Nagy, Institute of Education, University of Szeged, Szeged, Hungary.

Correspondence concerning this article should be addressed to Benő Csapó, MTA-SZTE Research Group on the Development of Competencies, University of Szeged, Petőfi sgt. 30-34, Szeged, Hungary. E-mail: csapo@edpsy.u-szeged.hu

tions, and collaborating) that are necessary for working in classroom settings (Konold & Pianta, 2005). Longitudinal research indicates that early (preschool as well as first grade) mathematical and reading skills represent strong predictors of later achievement (Duncan et al., 2007; Hair, Halle, Terry-Humen, Lavelle, & Calkins, 2006; Magnuson, Ruhm, & Waldfogel, 2007; Merrell & Tymms, 2010). For example, Tymms, Jones, Albone, and Henderson (2009) reported correlations ranged from .65 to .80 between kindergarten assessments and mathematics and reading achievements measured in the first and fifth grades.

A number of instruments have been developed to monitor early development (see C. E. Snow & Van Hemel, 2008), but only a few of them are used in regular educational practice due to theoretical and practical constraints. Among the theoretical problems often cited are the difficulties of defining the concept of school readiness and properly determining the purpose of testing. For instance, readiness may mean either readiness to learn or readiness to perform in a school setting (Carlton & Winsler, 1999). If readiness testing is focused on predicting school performance, then it will be the slow developers or low-performing children who are most in need of the developmental influences provided by school that are prevented from entering it (Shepard, 1997). To overcome these difficulties, a more complex conception of school readiness is proposed, a concept that also takes into account children's cognitive, emotional, and social development (Blair, 2002). These issues are less crucial if school readiness tests are used as diagnostic tools and identification of deficiencies is followed by treatment.

Early tests must take into account that the children assessed may not be able to read. Thus, these tests are usually individually administered with stimuli presented and instructions read by test administrators, who also record the answers. This limits the standardization of testing conditions and leaves the process open to subjective influences and interpretations of test takers' responses. Research on the quality of school readiness testing indicates that assessments made by teachers are often biased (as they are less strict with the children) when their conclusions are compared with results from objective assessment instruments (Mashburn & Henry, 2004). Despite these constraints, a number of school readiness assessments are based on the direct observation of children (e.g., the Early Development Instrument; see Guhn, Janus, & Hertzman, 2007). The Performance Indicators in Primary Schools (PIPS) tests are used to monitor children's development in the early years of primary school. Its Baseline Assessment (PIPS BLA) measures early reading; mathematics; phonological awareness; and personal, social, and emotional development on a 5-point scale (Merrell & Bailey, 2012). This instrument was used in a large-scale international study (iPIPS) to compare children's early development in English-speaking countries.

Although the majority of studies on school readiness assessment have focused on the cognitive domain, recent research identified several further factors, which play a crucial role in kindergarten-school transition and later development, such as self-concept, peer status, classroom contexts, and parenting (Bossaert, Doumen, Buyse, & Verschueren, 2011; McWayne, Cheung, Wright, & Hahs-Vaughn, 2012). Although there are still a number of open questions related to certain details of the content of school readiness assessment and the ways their data may be used there is a consensus that the availability of appropriate and easy-to-use measurement instruments is crucial to helping children to begin school successfully and to identify those who are in need of additional support (K. L. Snow, 2006).

Computer-Based Assessments and the Context of the Present Study

In educational practice, there may be different forces and interests driving the search for better solutions and applications of technology to replace traditional (face-to-face and paper-and-pencil) forms of assessment. The main factors motivating the use of technology are improving the assessment of already established assessment domains (Csapó, Ainley, Bennett, Latour, & Law, 2012) and measuring constructs that would be impossible or difficult to measure without the means of technology (e.g., Complex Problem Solving; see Greiff, Wüstenberg, & Funke, 2012; Greiff, Wüstenberg, Holt, Goldhammer, & Funke, 2013; Greiff at al., 2013).

Computer-based (CB) and paper-and-pencil (PP) test comparability studies were among the most extensively researched questions over the last two decades. Because of several advantages, CB assessment delivery has been gradually replacing traditional PP delivery as it permits the tailoring of tests to the individual characteristics of learners (e.g., adaptive testing), automated scoring (including promising developments in children's speech recognition) and immediate feedback, the inclusion of innovative item formats (e.g., multimedia elements, simulation, and dynamic items), precise control over the presentation of test stimuli, and reduced costs of test administration (see Price et al., 2009). One of the regular large-scale assessment programs, the Programme for International Student Assessment (PISA), is also gradually shifting from PP to CB assessments. In PISA (2006, 2009, 2012; see Organisation for Economic Co-Operation and Development, 2010, 2011, 2014), CB assessments were offered as international options or took place in one of the innovative domains; in 2015, the major domains (reading, mathematics, and science) will be assessed with computerized instruments.

A number of studies have been conducted in different knowledge and competence domains using a variety of educational tests to examine whether test delivery mode affects children's performance (Clariana & Wallace, 2002; Kingston, 2009; Wang, Jiao, Young, Brooks, & Olson, 2008). The differences between PP and CB test performance in terms of validity and reliability, advantages and disadvantages, and the effects of background variables (gender, race/ethnicity, and technology-related factors, such as computer familiarity; Csapó, Molnár, & Tóth, 2009; Gallagher, Bridgeman, & Cahalan, 2000) have been widely studied and well documented. Most of the recent media effect studies have indicated that PP and CB testing are comparable and that students prefer CB tests to traditional PP testing. Although the research results are inconsistent to some extent, comparability problems are likely to decrease over time as computers become more broadly accessible at schools (Way, Davis, & Fitzpatrick, 2006). Even though there is a lively debate over comparative studies, less attention has been paid to the effects of different delivery modes on different subgroups of the samples, and only a few studies have focused on testing very young learners in a technology-based environment (Carson, Gillon, & Boustead, 2011; Choi & Tinkler, 2002).

The most widely studied subgroup differences are those between girls and boys. Gender differences are routinely analyzed in large-scale assessment programs and are especially relevant in CB testing. The results of previous studies have revealed that the new media slightly changed the pattern of differences compared with the traditional PP assessments. In large-scale international PP assessments, the overall pattern is that boys and girls perform alike or boys do slightly better than girls in mathematics and science, whereas girls perform better than boys in reading. Boys usually perform better on the information-communication technology (ICT) literacy and computer familiarity test, and their better ICT skills may affect the results of CB tests. For instance, in the PISA (2006) study, science was tested in three countries using computers as well, and gender differences varied across the three participating countries on the Computer-Based Assessment of Science. However, boys outperformed girls on average (Organisation for Economic Co-Operation and Development [OECD], 2010). In the PISA (2009) survey, electronic reading was an innovative assessment domain. On the Electronic Reading Assessment, girls outperformed boys on average, and this pattern was the same across all OECD countries (OECD, 2011). Horne (2007) reported similar results in reading and spelling tests. In general, no gender differences were found on the computerized versions, whereas girls outperformed boys on the paper versions of the tests. A study compared the achievement of fifth-grade (11-year-old) primary-school children in inductive reasoning measured by PP and CB in a larger representative sample in Hungary and indicated no achievement differences between boys and girls in PP or CB test results (Csapó et al., 2009). In the context of early testing, analyzing gender differences may be essential for making existing instruments equally usable for boys and girls.

One of the main tasks of current developmental efforts is to migrate well-established face-to-face or PP tests to the new technology. However, whereas a switch to CB delivery is accompanied by some obvious improvements in efficiency, cost-effectiveness, and precision, further research is required to determine potential changes in reliability, ecological validity, applicability, and possible biases when migrating testing to the new medium. Most previous mode effect studies compared PP and CB delivery modes only. In the present study, we explore the differences between individual face-to-face and online testing as well.

The Development of the DIFER Test Battery and the Inductive Reasoning Test

The development of the school readiness test battery, which is at the center of this study, started back in the 1970s, when the first large-scale assessment of young learners in Hungary explored a group of skills necessary for a successful start of schooling. The results of this work (Nagy, 1980) formed the foundations for developing an extensive instrument, the PREFER test battery, administered face to face (FF) by teachers and covering the most essential competencies needed to begin school successfully (Nagy, 1987). After using it in educational practice for more than a decade, it was revised and renewed as the DIFER test battery (Nagy et al., 2004b). It has been used in several large-scale assessments to establish its reliability and predictive validity (Nagy et al., 2004a). Five DIFER tests were used in a longitudinal program, where they were the first instruments administered to a sample followed for 10 years (Csapó, 2007; Józsa, 2004). Strong correlations were found between the DIFER test and later school achievement. For example, the results of the counting and basic numeracy DIFER test correlated at .60 with a counting test administered at the end of second grade, and the correlation remained .49 (with a mathematical reasoning test) at the end of fifth grade and .48 at the end of eighth grade, with the mathematics test administered within the framework of the National Assessment of Basic Competencies (Csapó, 2013)

In educational practice, the DIFER can be used as a diagnostic instrument. Children are assessed regularly over time, and a record of their development is kept in a booklet. The development of those who lag behind may be stimulated by special purpose exercises. The DIFER is designed so that its administration does not require specific expertise; it can be administered by kindergarten and primary-school teachers. A major drawback of the test battery is that it must be administered face to face and individually. This is especially problematic for primary schools, where it is difficult to fit testing sessions into teachers' and learners' schedules. Another issue is the objectivity of the test administration as teachers may read the instructions to children in slightly different ways, and the scoring of the responses may also vary. An online delivery of prerecorded voice instructions (with texts read by trained speakers) and automated scoring may solve these problems. As two out of the seven DIFER tests (social skills and writing) cannot be immediately digitized, the remaining five (speech sound discrimination, relational reasoning, deductive reasoning, inferential reasoning, and counting skills) were transferred to an online platform in this study. The characteristics of inferential reasoning do not differ much from those of deductive reasoning; thus, we omitted inferential reasoning from the analyses presented in this article.

To increase the variety of the instruments, a PP inductive reasoning test and its digitized version were added to the four remaining DIFER instruments (speech sound discrimination, relational reasoning, counting and basic numeracy, deductive reasoning). The development of the inductive reasoning tests began in the early 1990s (Csapó, 1997). Several PP inductive reasoning tests have been in regular use for almost two decades both for mapping the development of inductive reasoning itself (Csapó, 2007) and for measuring inductive reasoning performance as an indicator of the developmental level of higher order thinking skills (Csapó & Nikolov, 2009). Later on, a second inductive reasoning test was constructed for the early grades, based on Klauer's model of inductive reasoning (Klauer, 1989), and has been used in experiments to assess the effect of training (Molnár, 2011). Using item response theory (IRT) analyses, both tests were equated so that their results were represented on the same scale (common-person methods were used; Molnár & Csapó, 2011). Finally, computerized versions were created for both tests, and the effects of delivery mode were studied by comparing the PP and CB versions (Csapó et al., 2009).

Research Questions

In this study, we explore the possibilities of the application of online assessment in regular educational practice at the beginning of schooling. For this purpose, we apply FF and PP tests that already have established psychometric characteristics; we transfer them to the new media and compare them by answering three research questions.

1. Does the medium of delivery influence the results, or can the results of the tests be represented on the same scale (i.e., testing of measurement invariance)?

2. If the tests differ between the two media, what influences these differences (i.e., psychometric properties)?

3. Does changing the mode of administration affect gender differences (i.e., latent mean differences between boys and girls)?

Method

A number of constraints have to be taken into account when carrying out comparative assessments with children entering school using an emerging technology.

1. Although online technology has had a relatively short developmental history, it has been extensively piloted with schoolchildren of different age groups, but not yet with preschool children.

2. To ensure comparability and to prevent the impact of schooling, a very short period is available for testing; schoolchildren may only be assessed at the very beginning of the first school year. (All assessments reported here took place during the first weeks of the first school year.)

3. As ecological validity is a main concern of the research, all assessment occurred in real school settings using the available infrastructure.

These conditions were equally taken into account when the study was designed, data sources selected, and procedures planned.

Participants

Data for two DIFER tests (relational reasoning test and counting and basic numeracy test) were drawn from an assessment in which all Hungarian children of school-entering age were assessed with the DIFER tests. A subsample was randomly selected for further detailed analyses; we used these data in this study. Two further DIFER tests (speech sound discrimination and deductive reasoning) were administered to different samples representatively drawn from the school-entering population. The PP inductive reasoning test was administered to a further representative sample. In each case, school classes formed the units of selection. The sample sizes and the attributes of the tests administered to the samples are summarized in Table 1.

The digitized versions of all five tests were administered to different samples due to organizational issues. These samples were randomly drawn from first-grade children in Hungarian primary schools. The online version of the speech sound discrimination test was administered to the same sample as its FF version. In this case, the order of modes was randomized, and there was a 2-week interval between the two testing sessions.

Instruments

The study is based on five tests that measure different skills essential for later learning. These key skills include (a) speech sound discrimination, a prerequisite of successful reading; (b) the ability to understand the meaning of words that denote relations; (c) number concept and basic counting skills; and basic (d) deductive and (e) inductive reasoning skills, all of which are prerequisites to learning to read and to studying mathematics and science.

An FF or PP version existed for each test, as described in the theoretical part of the present article. For the present study, we constructed electronic versions of the existing instruments, basically by migrating the items to the new platform. The viability and success of the migration depended on the content of the assessment and the item type. In the process of test digitization, one of the central aims was to preserve as many features of the items as possible in order to make the two delivery modes comparable. The paper and screen layouts were identical or as similar as possible.

Two out of the seven DIFER tests could not be implemented in the new medium. The FF social skills test is based on an observation of the children's behavior. This test proved to have high predictive validity in a longitudinal study, but it could not be realized on a computer. The writing test examined fine hand movement (fine motor skills), which is a precondition of learning handwriting. It was not possible to implement this with the available technology. The other FF DIFER tests were converted into CB formats, although some items had to be omitted and the open-ended items were reformulated and converted into multiple-choice items to allow automated scoring. Only items implemented in both media were used in the comparative analyses.

Speech sound discrimination test. This test includes 60 items that measure the perception of phonemic contrasts. The test reveals whether children have good hearing and are able to differentiate some critical pairs of phonemes, for instance /v/ - /f/ and /b/ - /p/ (e.g., in the pairs of Hungarian words *vonal-fonal* and *bont-pont*).

In the first part of the original FF version of the test, the administrator read two sentences; each one contained one of the words from the pairs and showed the matching picture depicting the object referred to in the sentence. Then the administrator read only one of the words. The children indicated their answer by pointing to the picture that matched the word the administrator had

Table 1
The Samples in the Study and the Attributes of the Tests (Number of Items, Reliability)

Test	Sample sizes		Number of items	Cronbach's α	
	N (FF or PP)	*N* (CB)		FF or PP	CB
Speech sound discrimination[a]	(FF) 364	364	60	.887	.938
Relational reasoning	(FF) 1,892	426	24	.796	.844
Counting and basic numeracy	(FF) 1,895	435	13	.812	.770
Deductive reasoning	(FF) 424	402	32	.743	.831
Inductive reasoning	(PP) 952	377	37	.855	.856

Note. FF = face-to-face; PP = paper and pencil; CB = computer-based.
[a] The FF and CB tests were administered to the same sample.

read. Finally, the test administrator scored and logged the answers on a scoring sheet. In CB mode, the same pictures were presented on the screen, and instructions were given online by a prerecorded voice. Children used headsets and heard the same voice of a trained speaker. They had to indicate their answer by using the mouse and clicking on the correct picture. An analogous English-language example could be the following: "This is a sheep (showing a picture of a sheep). This is a ship (showing a picture of a ship). Now I will only say one word. Point/click at the picture, which depicts it."

The second part of the test focused on children's phoneme perception in fluent speech and on the correct pronunciation of a word depicted by a picture. Finally, the third and fourth subtests contained pairs of real or pseudowords or words that differed in one phoneme. Test takers had to decide whether the two words in each pair matched or not.

Relational reasoning test. Understanding words that denote relations between different objects, attributes, or processes is a precondition of school learning. The DIFER contains four equivalent versions of relational reasoning tests both in FF and in CB mode, each containing 24 items. As their structure was identical, we use only one version in this analysis. In each test, there were eight relation words tied to space (e.g., *inside*, *between*), four relation words encoding quantity (e.g., *odd*, *few*), four relation words referring to actions (e.g., *step on*, *step in*), four relation words related to time (e.g., *earlier*, *later*), and, finally, four different relational expressions encoding physical measures (e.g., "the shortest," "the same length"). Figure 1 shows a sample item from the CB test; the picture was the same in FF mode as well.

With FF administration, the instructions were given and the test administrator scored the answers. Children had to supply their answers by pointing to the matching picture(s). In the CB environment, instructions were given online; students had to provide their answers by using the mouse and clicking on the matching picture(s).

Counting and basic numeracy skills test. The original test constructed for FF administration consisted of items that measure the understanding of the meaning of numbers, number relations, and basic mathematical thinking. Some items were based on oral counting, and, as the online platform is not yet able to handle oral responses, a number of items on the original numeracy test were omitted from the CB version. Only items that test recognition of quantities, numbers, and representations of numbers were kept. Figure 2 illustrates items on the CB version of the test.

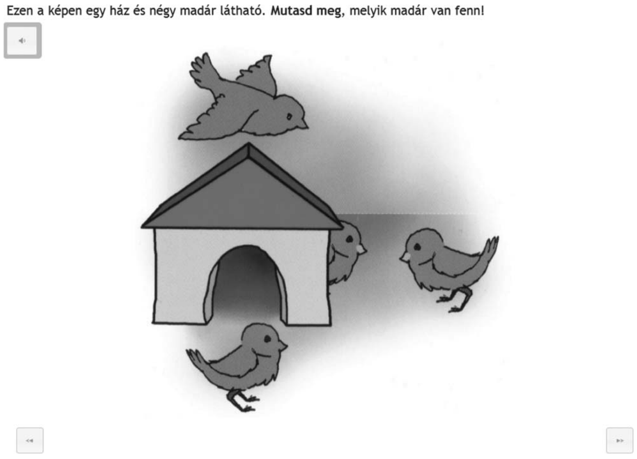

Ezen a képen egy ház és négy madár látható. **Mutasd meg**, melyik madár van fenn!

Figure 1. Sample item from the computer-based version of the relational reasoning test. [In this picture, you can see a house and four birds. Click on the bird that is higher up than the others.]

The FF and CB data collection proceeded the same way as in the sound discrimination test described before.

Deductive reasoning test. Deductive reasoning was measured with 32 open-ended, contextually embedded tasks in FF mode and with 32 multiple-choice tasks in CB mode to make automated scoring practicable and to allow immediate feedback after testing in the latter case. Each task began with two premises (statements), and children had to reach and formulate a logical conclusion. The context of the situations presented to them may have been familiar from everyday life, so it would have been possible for them to use real-world knowledge to formulate their conclusions.

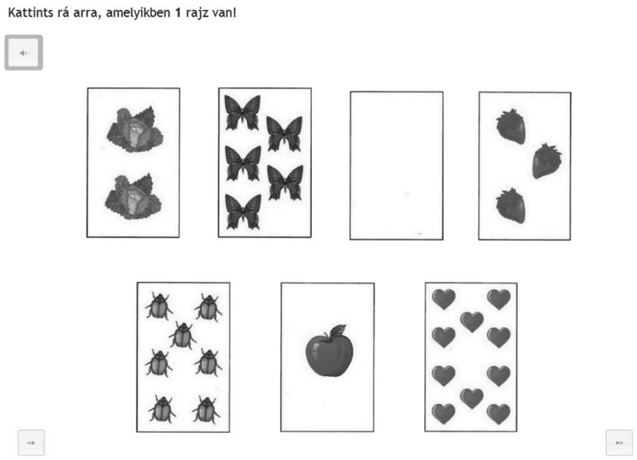
Inductive reasoning test. The structure of the inductive reasoning test was based on Klauer's (1989) definition of inductive reasoning. Klauer defined inductive reasoning as discovering regularities by detecting similarities, dissimilarities, or a combination of both, with respect to attributes or relations to or between objects. This involved six classes in total (generalization, discrimination, cross-classification, recognizing relations, discriminating relations, and system formation). The test consisted of 37 figural, nonverbal items belonging to the six subclasses of inductive reasoning described above. Figure 3 illustrates the items on the inductive reasoning test; the same pictures were used both in PP and CB modes.

During the digitization of the test, all features of the items were preserved to make the two versions as similar as possible. For example, in the PP multiple-choice items, children had to circle or underline the letter or the picture, whereas in the CB format, they had to click on the same letter or picture to indicate their answer (see Figure 3).

Procedure

Two traditional delivery methods, FF and PP testing modes, were used. During FF administration, children were tested individually. The instructions for the items were read by test administrators, most of whom were the children's homeroom teachers, and children's answers were recorded on a scoring sheet. The PP version of the inductive reasoning test was taken in the children's regular classroom under the supervision of their class teachers. The scoring sheets of all tests were collected after the testing session, data were centrally processed, and no feedback was provided to the children.

The online data collection was carried out via the *eDia* (Electronic Diagnostic Assessment) platform through the Internet. Testing took place in the computer labs at the participating schools, using the available computers and browsers installed. A session lasted approximately 20–45 min, depending on the test. The items were automatically scored, and children received immediate feedback (percent of correct answers) at the end of the testing session. The *eDia* platform allows the use of proxy servers.



Kattints rá arra, amelyikben 1 rajz van!

Figure 2. Sample item from the computer-based version of the counting and basic numeracy skills test. [Click on the card with a drawing of only one thing on it.]

Statistical Analyses

In this article, we analyze the differences between paper-based, FF, and online tests. We not only examine whether tests presented in different modes are equivalent, but we also show where they are different, as one of the aims of this study was to support the design of better online instruments. To reach this goal, we applied several analyses, including computations based on classical test theory, confirmatory factor analyses within structural equation modeling (SEM; Bollen, 1989), to test the underlying measurement model and to test measurement invariance, and IRT. In this section, we only discuss the theoretical background of how SEM analyses were applied in the present study.

Providing a meaningful interpretation of test scores and ensuring the comparability and validity of FF and CB test results is only possible if the structure of the construct does not change across delivery modes (Byrne & Stewart, 2006). That is, measurement invariance must be analyzed to examine whether test results are affected by the test medium and to ensure that the same constructs are being assessed in each group. If measurement invariance is sufficiently met, and, thus, structural stability exists, between-group differences can be interpreted as true and not as psychometric differences in latent ability (Greiff et al., 2013).

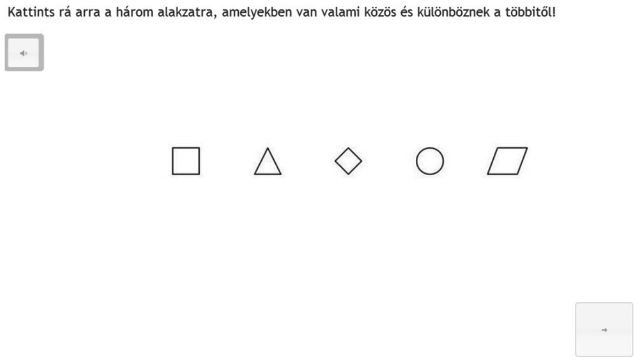
A number of approaches, statistical methods, and concepts are available to test measurement equivalence (Schroeders & Wilhelm, 2011). State-of-the-art methods share a common feature: The definition of the measurement model is provided through a comparison of the latent structure for several groups in a single model. The most prominent methods are those used to detect differential item functioning within the IRT approach (Raju, Laffitte, & Byrne, 2002) and multigroup confirmatory factor analysis (MGCFA) (Bollen & Curran 2006; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000) within the SEM framework.

In the present study, a between-subject design was used to test invariance by means of MGCFA. Weighted least squares, mean- and variance-adjusted (WLSMV) estimation was applied, and THETA parameterization was used because all items were scored dichotomously (Muthén & Muthén, 2010). All measurement models were computed with Mplus. Goodness of fit to the sample data was evaluated on the basis of multiple criteria. Different fit indices have been developed (Wu, Li, & Zumbo, 2007), and numerous cutoff criteria, such as the Tucker–Lewis Index (TLI), comparative fit index (CFI) $\geq$ 0.90 or 0.95, and root-mean-square error of approximation (RMSEA) $\leq$ 0.06 or 0.08, have been proposed to assist in determining model fit (see Byrne & Stewart, 2006; Fan & Sivo, 2005; Vandenberg & Lance, 2000). In this study, an absolute fit index (the RMSEA), a relative fit index (the TLI), and an incremental, normed fit index (the CFI) were used to evaluate model fit. Nested model comparisons were conducted using a special chi-square difference test for the WLSMV estimator (Muthén & Muthén, 2010).

Testing for measurement invariance (Muthén & Muthén, 2010; Vandenberg & Lance, 2000) with categorical data involves a fixed sequence of model comparisons, testing different levels of invariance by comparing measurement models from the least to the most restrictive model by using MGCFAs. Measurement invariance exists if restrictions of model parameters in one model do not generate a substantially worse model fit in comparison to an unrestricted model. The procedure for testing measurement invariance is explained thoroughly by Byrne and Stewart (2006).

Configural invariance investigates whether the basic model structure is invariant across groups (Byrne, 2008), that is, whether children in the CB and FF environments conceptualize the construct in the same way (Milfont & Fischer, 2010) and thus use the same conceptual framework to answer the test items (Vandenberg & Lance, 2000; Wu, Li, & Zumbo, 2007). Configural invariance indicates that the same item is an indicator of the same latent factor in each group, but factor loadings may differ across groups. When we tested configural invariance with categorical outcomes, thresholds and factor loadings were not constrained across groups, factor means were fixed at 0 in all groups, and residual variances were fixed at 1 in all groups. The next step is to test weak factorial invariance, that is, to test cross-group equality in the loadings. However, testing it for categorical data is not recommended (Muthén & Muthén, 2010), and thus weak factorial invariance was not tested (see, e.g., Greiff et al., 2013; Schroeders & Wilhelm, 2011).

Strong invariance, as the subsequent step in testing measurement invariance, indicates that the variances for latent variables and the covariances between the latent variables are equal between



Kattints rá arra a három alakzatra, amelyekben van valami közös és különböznek a többitől!

Figure 3. Sample item from the computer-based version of the inductive reasoning test. [Click on the three shapes that have one thing in common that the other two do not.]

CB and FF modes; that is, cross-group equality exists in the loadings and intercepts. When this was tested, thresholds and factor loadings were constrained so that they would be equal across groups, and residual variances were fixed at 1 and factor means at 0 in the FF group, whereas there were no constraints specified in the CB group (Muthén & Muthén, 2010). If strong factorial invariance did not hold according to the modification indices, partial strong invariance was tested. Strong factorial invariance is the level at which latent mean comparisons can be conducted (Byrne & Stewart, 2006).

Finally, strict factorial invariance indicates whether the CB and FF groups have the same item residual variances (Byrne, 2008). It requires cross-group equality in the loadings, intercepts, and residual variances. Therefore, in addition to the restrictions applied in strong factorial invariance, all residual variances were fixed at one in all groups, even though strict factorial invariance is not a prerequisite for media comparisons of latent factor means and variances.

Results

The presentation of the results is organized according to the research questions. First, we examine measurement invariance for each of the five tests (Research Question 1). Second, as we see that in some cases the scales did not remain identical, we study the direction of the changes (Research Question 2). Finally, we examine whether changing the testing media has the same impact on boys and girls (Research Question 3).

Research Question 1: Examining the Media Effect Through Analyses of Measurement Invariance

The measurement invariance analyses were performed as described in the Method section. The results are summarized in Table 2.

For the speech sound discrimination test, both the FF and the CB versions were administered to the same sample, and a multivariate, single-level approach made it possible to test measurement invariance in a single-group analysis. First, we tested the confirmatory factor analysis at each of the two points in time to be sure the model fit well in both modes. Examination of the modification indices suggested that model fit would be significantly improved by changing the original model. According to the results of the LaGrange multiplier test, we needed to delete some 16 items from the analyses because of ceiling effects. The remaining 44 items fit the data in both modalities well (FF: RMSEA = .019, CFI = .943, TLI = .940; CB: RMSEA = .035, CFI = .915, TLI = .910). The strong factorial invariance model did not fit well and resulted in a significant decrease in fit relative to the configural invariance model. The examination of the modification indices suggested that model fit would be significantly improved by allowing the intercept for one item to differ between data collections and adding residual covariances between two items to the CB model. Partial strong invariance did hold. The observed differences in item means between PP and CB testing was due to factor mean differences (for one item, children in FF mode were expected to have higher item response) and a residual covariance in CB mode (two items proved to be correlated). Finally, we tested partial strict invariance, resulting in a significant decrease in fit relative to the configural model; that is, the relations of the items to the latent factor of speech sound discrimination were not equivalent in PP and CB modes in a strict sense. However, strict factorial invariance is not a prerequisite for latent mean comparisons; in this case, partial strong factorial invariance is sufficient to compare latent means.

The results regarding the strong factorial invariance model for relational reasoning indicated a significant decrease in fit relative to the configural invariance model. The modification indices suggested a freeing of the intercept for two items between PP and CB

Table 2
Goodness-of-Fit Indices for Measurement Invariance of the Tests

Test	Model	χ^2	*df*	CFI	TLI	RMSEA	$\Delta\chi^{2a}$	Δdf^{a}	*p*
Speech sound discrimination	(1)	2905.7	2628	.908	.906	.017			
	(2)	2950.2	2663	.905	.904	.017	60.0	35	<.05
	(2.1)	2935.9	2661	.909	.908	.017	44.8	33	>.05
	(3)	3231.4	2700	.824	.824	.024	287.0	72	<.05
Relational reasoning	(1)	389.9	97	.971	.961	.051			
	(2)	490.5	108	.962	.954	.055	89.2	11	<.01
	(2.1)	396.0	105	.971	.964	.049	27.4	8	<.01
	(3)	1027.1	119	.910	.900	.081	463.9	22	<.01
Counting and numeracy	(1)	46.2	17	.996	.997	.038			
	(2)	243.4	22	.984	.978	.093	128.2	5	<.01
	(2.1)	157.0	20	.990	.985	.076	70.8	3	<.01
	(3)	268.9	27	.983	.981	.087	169.6	10	<.01
Deductive reasoning	(1)	185.8	136	.980	.973	.030			
	(2)	293.1	146	.942	.927	.049	75.9	10	<.01
	(2.1)	264.4	142	.951	.938	.046	56.4	6	<.01
	(3)	348.1	166	.928	.921	.051	34.0	14	<.01
Inductive reasoning	(1)	1791.2	908	.929	.923	.037			
	(2)	1828.4	930	.924	.919	.038	42.0	22	>.01
	(2.1)	1806.4	931	.926	.921	.038	15	23	>.05
	(3)	1868.1	962	.921	.916	.039	55.0	32	>.05

Note. Model: (1) = configural invariance; (2) = strong factorial invariance; (2.1) = partial strong factorial invariance; (3) = strict factorial invariance. CFI = comparative fit index; TLI = Tucker–Lewis Index; RMSEA = root-mean-square error of approximation.
[a] $\Delta\chi^2$ and Δdf were estimated with the Difference Test procedure (DIFFTEST) in Mplus. When using weighted least squares, mean- and variance-adjusted estimation, χ^2 differences between models cannot be compared by subtracting χ^2 and *df* (Muthén & Muthén, 2010).

testing. The partial strong invariance model in which the intercept for these two items was allowed to differ between PP and CB groups fit better than the strong factorial invariance model, but still significantly worse than the configural invariance model. This means no (partial) measurement invariance could be established for the relational reasoning test.

For counting and basic numeracy, strong factorial invariance model resulted in a significant decrease in fit relative to the configural invariance model as well. According to the LaGrange multiplier test, the intercept for Items 6 and 7 between PP group and CB group were freed. The partial strong invariance model fit better than strong factorial invariance model, but still significantly worse than configural invariance model. Thus, no measurement invariance could be established in this case either.

The result for the invariance testing of deductive reasoning indicated a decrease in model fit for all levels of invariance; thus, there was no measurement or partial measurement invariance between the FF and CB testing modes for deductive reasoning. These data suggest that CB (online administration) does not measure exactly the same construct as FF delivery does. To this end, mean differences between FF and CB groups could not be interpreted as true differences in the underlying deductive reasoning construct; this could also be due to psychometric issues. One of the possible reasons for this is that information was more standardized in the CB environment; achievements in the CB environment were independent of the teacher's attitude and judgment during data collection. This was not the case in the FF mode.

The result for the invariance testing of inductive reasoning represented no loss in model fit; even imposing the most restrictive constraints did not lead to deterioration in model fit. The model of strong factorial invariance did not show a decrease in model fit compared with the model of configural invariance. Strict factorial invariance could also be established; that is, even residual variances proved to be equal across delivery media in a strict sense.

Research Question 2: Differences Between the Tests Delivered in Different Media

As shown in the previous section, depending on the content of the assessment and the item types, there are differences between the tests in respect of how the measurement scales changed when the items were transferred to the online platform. In this section, we compare the reliability of the tests delivered by the two media, examine the impact of the media on performance, and have a closer look at the item level differences by comparing the difficulties of the items in the two media.

Reliability of the tests. The internal consistencies of the tests were examined by computing Cronbach's alpha for each test. The DIFER tests were previously also administered to participants in the Hungarian Educational Longitudinal Program ($5,000 > n > 6,000$), and the reliability indices of those two digitized in this study were high (relational reasoning: .726; counting and basic numeracy: .915), the relational reasoning test showing the lowest value (Csapó, 2007; Józsa, 2004). In the present study, the reliability indices were slightly different but in general also good both in FF/PP and CB modes. They ranged from .743 to .887 in the FF/PP mode and from .770 to .938 in the CB mode. Generally, the reliability indices of the CB tests proved to be somewhat higher than those of the FF/PP test versions (see Table 1).

The reliability value was already high for the FF administration of speech sound discrimination (.887), and it improved further (to .938), being the highest within this set of tests. This improvement may be attributed to the standardized voice stimuli. There were slight improvements in Cronbach's alpha for relational reasoning and deductive reasoning. A major drop of reliability was observed for the counting and basic numeracy test. Although several items were dropped from the FF version because they required oral responses and it was not possible to implement this in the CB version, the reduced FF test consisting of 13 items still had a relatively high reliability (.813). The PP version of the inductive reasoning test was digitized without major changes. This was reflected in the reliabilities as Cronbach's alphas of the two versions did not differ.

The impact of the assessment media on performance. A meaningful interpretation of differences in test scores is only possible if the structure of the construct measured does not change across test media (Byrne & Stewart, 2006). That is, latent (and manifest) mean comparison can only be interpreted meaningfully if at least strong factorial invariance is established (Brown, 2006). According to the measurement invariance analyses, testing for latent mean differences was only meaningful in the case of speech sound discrimination and inductive reasoning tests, where partial strong and strict measurement invariance held, respectively.

Latent mean comparisons were conducted by constraining the item intercepts of the observed variables equal and setting the latent factor means for the FF (or PP) group as reference group to zero (Byrne & Stewart, 2006). We also calculated performance on tests in percentages, summarizing the results in Table 3. We report

Table 3
Test-Level Achievement Differences Between Traditional (FF or PP) and CB Modes

	FF or PP (%)		CB (%)			Latent		
Test	*M*	*SD*	*M*	*SD*	*d*	*M*	*SE*	*p*
Speech sound discrimination	91.35	9.88	82.61	16.80	.59	−.77	.13	<.01
Relational reasoning[a]	80.86	15.44	78.03	17.86	.17	−.67	.11	<.01
Counting and basic numeracy[a]	87.35	18.46	88.90	14.84	−.09	.19	.07	*ns*
Deductive reasoning[a]	70.69	14.34	63.59	20.87	.39	−.58	.06	<.01
Inductive reasoning	47.52	19.04	45.99	18.11	.08	−.10	.06	*ns*

Note. *d* is the difference between the means in standard deviation units (Cohen's *d*). Latent mean for the FF group was set to zero. FF = face-to-face; PP = paper and pencil; CB = computer based.
[a] Measurement invariance does not hold.

the means also for those tests for which measurement invariance could not be established.

The data indicate that performance on the speech sound discrimination test was significantly lower in the CB tests than in the FF test: it fell from a high (91.4%) to a still high but significantly lower level (82.6%). A similar drop was found in the case of deductive reasoning and a modest drop for relational reasoning. No significant decreases in mean differences were observed for the counting and basic numeracy test or the inductive reasoning test. The latter represents a PP-CB transition, and no significant difference was found between the two versions.

Item-level differences. A further way of analyzing the characteristics of CB testing is to have a look at the results at the item level. To do this, we computed the item difficulties for each item in both media. To illustrate the possibilities for this type of analysis, we present the results of the speech sound discrimination, deductive reasoning, and inductive reasoning tests.

For the speech sound discrimination test, we once again took advantage of the two test versions being administered to the same sample and computed the item parameters on the basis of IRT scaling. We considered the entire item pool of the two versions of the test as items of a single test and calculated the item parameters. According to the analysis of the content, items that contained nonsense words proved to be most affected by FF administration. A possible reason for this difference may be that in the case of meaningless words, teachers helped children to find the correct answer. The same effect was observed for words where the lack of context could in part have been replaced by teachers' helpful behavior, whereas the impact of FF administration was less apparent for words in complete sentences (only in the case of this latter test was a significant correlation found between the item difficulty parameters: $r = .550$, $p < .01$). These results also suggest that items presented by teachers lose their objectivity in some cases, especially if stimuli are taken out of their usual context.

For the deductive reasoning tests, a much higher correlation was found between the item difficulties in the two media ($r = .750$, $p < .01$). The inductive reasoning test showed the most "regular" picture, in agreement with the previous observations. The items correlated to a very high degree ($r = .948$, $p < .01$), indicating that the PP and CB tests measure inductive reasoning skills very similarly, not only at the overall test level but also at the level of items as well. The inductive reasoning test is the only one where one of the versions was taken on paper. The same can be observed here as what we have already shown concerning the reliability and the difference between the means: The two versions of the test behave very much alike, so they can essentially be considered identical.

Research Question 3: Gender Differences

As previous studies have indicated, gender differences are influenced by the medium of testing. In general, boys perform somewhat better if they are assessed via technology-based instruments. In the previous FF versions of the DIFER tests, girls performed somewhat better on the speech sound discrimination test, whereas boys performed better on the counting and numeracy test (Józsa, 2004).

To examine how transferring the DIFER tests and the inductive reasoning test to the online platform affected gender differences at this very young age, we carried out several analyses. First, we performed invariance analyses with regard to gender.

The model used to test configural invariance of speech sound discrimination for boys and girls fit well. The model of strong factorial invariance did not show a decrease in model fit based on the stricter perspective (nonsignificant chi-square difference test; cf. Table 4) compared with the model of configural invariance. Finally, strict factorial invariance could not be established. As strong factorial invariance held (see Meredith, 1993), mean differences could be interpreted as true differences in the construct being measured between girls and boys (Byrne & Stewart, 2006).

In the computerized versions of the tests examined here, a significant gender difference was only found for the speech sound discrimination test: Girls performed somewhat better than boys (in %, girls: $M = 86.37$, $SD = 12.76$; boys: $M = 80.40$, $SD = 18.03$. $t = -3.94$, $p < .001$). No significant gender differences were found for the other tests.

Discussion

Measurement Invariance

We found that measurement invariance held in two out of the five cases (if we consider the practical perspective and the strong invariance model sufficient) and, in a strict sense, only in one case, for the inductive reasoning test. This test was originally a PP test, and the migration of the items took place so that neither the general look of the test nor the item types were changed.

Measurement invariance held only partially for the speech sound discrimination test. This was originally an individually administered FF test, and the same pictures were presented to the children in both modes. The item type and the scoring principle did not change either. This finding indicates that under certain conditions, even an FF test can be transferred to the online platform with acceptable results in terms of measurement invariance. This may also hold in part for the relational reasoning test.

Measurement invariance did not hold for the three remaining tests. These results indicate that equivalent scales may only be constructed if the migration of the items does not change the item types and only changes the testing context moderately. If the migration influences the objectivity of scoring, the tests administered in the two media are not exactly identical. This happened on

Table 4
Goodness-of-Fit Indices for Measurement Invariance of the Speech Sound Discrimination Test for Boys and Girls

Model	χ^2	*df*	CFI	TLI	RMSEA	$\Delta\chi^{2a}$	Δdf^a	*p*
(1)	2858.7	2536	.932	.929	.024			
(2)	2887.0	2580	.935	.934	.024	44.8	44	>.05
(3)	2920.2	2588	.930	.928	.024	80.2	52	<.05

Note. Model: (1) = configural invariance; (2) = strong factorial invariance; (3) = strict factorial invariance. CFI = comparative fit index; TLI = Tucker–Lewis Index; RMSEA = root-mean-square error of approximation.
[a] $\Delta\chi^2$ and Δdf were estimated with the Difference Test procedure (DIFFTEST) in Mplus. When using weighted least squares, mean- and variance-adjusted estimation, χ^2 differences between models cannot be compared by subtracting χ^2 and *df* (Muthén & Muthén, 2010).

the deductive reasoning test as the open-ended items were converted to multiple-choice items.

From an applied perspective, the scale differences may only cause problems if there is a need to compare performance within the two media. If the purpose of transferring the test is to construct a new, more applicable instrument, the lack of measurement invariance does not cause a problem. As the aim of the present study was to establish the development of a new instrument, analyzing further details of the differences where the measurement invariance does not hold may help to construct better tests. If the changes may be attributed to the improvement of the instrument, the lack of invariance may even be favorable, but in this case, the new instruments can only be used in practice after careful piloting and validation processes.

The Impact of Media on Reliability and Achievement Scores

The study has shown that the reliability of assessments may be improved by transferring individually administered FF instruments to an online platform. Having a look at the reliability coefficients of the five instruments in two modes, we may observe that the reliability increased in those cases when CB assessment provided more standardized conditions compared with FF administration.

The results indicated that performance was lower on the CB tests than on the FF or PP tests in most cases. This suggests that teachers tend to give higher scores to children than the scores they receive when their responses are automatically (and objectively) scored. Another explanation for the differences might be that children had difficulty handling the computerized tests, and this lowered their performance. However, as no significant mean differences were found for the counting and basic numeracy test or the inductive reasoning test, low computer familiarity is not a sufficient explanation for achievement differences. In fact, a more realistic explanation may be that teachers were more tolerant in accepting children’s responses.

The Impact of Media on Gender Differences

Due to the limitation of the available data, gender differences were explored by latent analyses only for the speech sound discrimination test. The results indicated that measurement invariance did hold; thus, FF and CB testing was the same for boys and girls, and latent means could be compared. Transferring the test to the new medium affected their performance similarly. A comparison of the raw scores indicated that girls’ performance was somewhat better, similarly to former FF assessments, where girls usually performed slightly better than boys. Together, these data may indicate that using the new medium will not cause a major bias in future applications.

Limitations of the Present Study

Due to the context of the study (using a system that is still under development), there were smaller sample sizes available compared with previous large-scale assessments. The availability of computers at school and the time first-graders were available to work on the computers limited the possible complexity of the study. The analyses are also constrained by the unavailability of additional background variables. In this phase of the research, we have no data concerning the possibilities of using the same instrument for repeated testing to monitor development. (The previous FF version of the DIFER is routinely used for this purpose without problems.) There are no data on the predictive validity of the CB instruments either. However, as indicated earlier, the FF version was administered in a longitudinal study and proved to be a good predictor of later school performance.

A further limitation of the present study is that the original DIFER tests were designed to assess children in the kindergarten-to-school transition period. Thus, students who have already started school tend to perform close to ceiling on these tests, and their data are not ideal for analyzing the characteristics of the tests. This deficiency may be rectified by extending future investigations to kindergarten populations, although the use of computers with that age group calls for further feasibility studies.

Conclusions and Further Prospects for Online Assessment of School Readiness

This study shows the potential and limitations of transferring school readiness tests to the new assessment medium of computers. In the digitization process, we have lost a strong and important test with high reliability and predictive validity as the social skills test cannot be replicated in the new medium while remaining close to its original form. We have also lost the writing test, but a closely related construct (fine hand movement) can easily be measured using emerging technologies. In fact, an alternative construct (handling keyboard and mouse) can also be easily measured by computer. The relevant technology is at hand for research purposes, and it will probably also be widely available in schools. We have also lost a large number of relevant items on the counting and basic numeracy skills test, as it was not possible to capture children’s oral responses, and we have paid for this loss with a drop in reliability. Further efforts are therefore needed to develop a suitable CB counting skills test.

The study has demonstrated the applicability of technology-based assessment in regular school practice at the earliest possible point of schooling in a number of highly relevant competency domains. These assessments can be carried out practically anytime, at very low cost, and with almost no extra teacher time. However, devising and using such instruments require further research in at least three dimensions: (a) making constructs currently assessed with traditional instruments measurable using computer technology and assessing new constructs that are especially well suited to CB assessment; (b) enhancing the online assessment technology with functionalities that are already in use in other areas of information technology (e.g., speech recognition and detection of emotions); and (c) exploring ways of integrating frequent early assessment into educational processes.

A number of technological solutions that can be built into the online assessment system to enhance its capabilities already exist elsewhere. Interaction, simulation, and manipulation of objects on screen, and new types of stimuli, such as video and animation, are currently in use in some assessments. It is also possible to time the stimuli and control the presentation of information in other ways. Measuring response time, logging keystrokes, and mouse movement can also routinely be used, although further studies are required to explore how these methods may contribute to solving the real problems of early assessment. One example of a real problem, where an existing technological solution may be essential, emerges from the present study: Voice recognition technology is needed to make the counting test deliverable online.

Further research is needed to explore the educational applicability of online assessment. Ecological validity is an issue that requires careful consideration. Examining predictive validity is crucial for tests that assess the preconditions of further learning and are used to identify early indicators of later problems. An examination of the online tests discussed here has already started as an extension of the present study. An exploration of the effects of repeated testing has also begun, but the accumulation of a sufficient quantity of data for analyses will take time in both cases. As learners' assessment results can easily be stored in the online assessment system, it is possible to gather not only overall performance data but also information collected on behavioral processes. In addition, the growing information base facilitates adequate monitoring of learners' development.

In the past decade, the issue of early development has been approached not only by researchers in numerous fields of study in education and psychology but also by those in other social sciences, such as sociology and economics. Results from these comprehensive studies have indicated that numerous problems that arise later are rooted in difficulties in the first school years. Research has also shown that these difficulties may be overcome with adequate intervention and that investing in such programs produces high returns. An important component of the well-prepared and well-timed intervention is proper diagnosis. To this end, CB assessment of the basic skills may be one of the best means to diagnose problems and monitor development.

References

Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist, 57,* 111–127. doi:10.1037/0003-066X.57.2.111

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.

Bossaert, G., Doumen, S., Buyse, E., & Verschueren, K. (2011). Predicting children's academic achievement after the transition to first grade: A two-year longitudinal study. *Journal of Applied Developmental Psychology, 32,* 47–57. doi:10.1016/j.appdev.2010.12.002

Brown, T. (2006). CFA with equality constraints, multiple groups, and mean structures. In T. Brown (Ed.), *Confirmatory factor analysis for applied research* (pp. 236–319). New York, NY: Guilford Press.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20,* 872–882.

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13,* 287–321. doi:10.1207/s15328007sem1302_7

Carlton, M. P., & Winsler, A. (1999). School readiness: The need for a paradigm shift. *School Psychology Review, 28,* 338–352.

Carson, K., Gillon, G., & Boustead, T. (2011). Computer-administrated versus paper-based assessment of school-entry phonological awareness ability. *Asia Pacific Journal of Speech, Language and Hearing, 14,* 85–101.

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper and computer based assessment in a K-12 setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33,* 593–602. doi:10.1111/1467-8535.00294

Csapó, B. (1997). The development of inductive reasoning: Cross-sectional assessments in educational context. *International Journal of Behavioral Development, 20,* 609–626. doi:10.1080/016502597385081

Csapó, B. (2007). Hosszmetszeti felmérések iskolai kontextusban - az első átfogó magyar iskolai longitudinális kutatási program elméleti és módszertani keretei [Longitudinal assessments in school context – theoretical and methodological frameworks of the first large-scale school-related longitudinal program in Hungary]. *Magyar Pedagógia, 107,* 321–355.

Csapó, B. (2013, May). *The predictive validity of school readiness assessment: Results from an eight-year longitudinal study*. Poster presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Csapó, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2012). Technological issues of computer-based assessment of 21st century skills. In B. McGaw, P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). New York, NY: Springer. doi:10.1007/978-94-007-2324-5_4

Csapó, B., Molnár, G., & Tóth, K. R. (2009). Comparing paper-and-pencil and online assessment of reasoning skills: A pilot study for introducing TAO in large-scale assessment in Hungary. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 113–118). Luxemburg, Belgium: Office for Official Publications of the European Communities.

Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences, 19,* 209–218. doi:10.1016/j.lindif.2009.01.002

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428–1446. doi:10.1037/0012-1649.43.6.1428

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12,* 343–367. doi:10.1207/s15328007sem1203_1

Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). *The effect of computer-based tests on racial/ethnic, gender, and language groups* (GRE Board Professional Report No. 96-21P). Princeton, NJ: Education Testing Service.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement, 36,* 189–213. doi:10.1177/0146621612439620

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of complex problem solving: Concept, implementation, and application. *Educational Technology Research and Development, 61,* 407–421. doi:10.1007/s11423-013-9301-x

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond *g*: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105,* 364–379. doi:10.1037/a0031856

Guhn, M., Janus, M., & Hertzman, C. (2007). The Early Development Instrument: Translating school readiness assessment into community actions and policy planning. *Early Education & Development, 18,* 369–374. doi:10.1080/10409280701610622

Hair, E., Halle, T., Terry-Humen, E., Lavelle, B., & Calkins, J. (2006). Children's school readiness in the ECLS-K: Predictions to academic, health, and social outcomes in first grade. *Early Childhood Research Quarterly, 21,* 431–454. doi:10.1016/j.ecresq.2006.09.005

Horne, J. (2007). Gender differences in computerised and conventional educational tests. *Journal of Computer Assisted Learning, 23,* 47–55. doi:10.1111/j.1365-2729.2007.00198.x

Józsa, K. (2004). Az első osztályos tanulók elemi alapkészségeinek fejlettsége Egy longitudinális kutatás első mérési pontja [Developmental level of first-grade students' basic skills. The first measurement point of a longitudinal research program]. *Iskolakultúra, 14,* 3–16.

Kingston, N. M. (2008). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22,* 22–37. doi:10.1080/08957340802558326

Klauer, K. J. (1989). *Denktraining für Kinder I* [Training of thinking for children]. Göttingen, Germany: Hogrefe.

Konold, T. R., & Pianta, R. C. (2005). Empirically-derived, person-oriented patterns of school readiness in typically-developing children: Description and prediction to first-grade achievement. *Applied Developmental Science, 9,* 174–187. doi:10.1207/s1532480xads0904_1

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly, 22,* 18–38. doi:10.1016/j.ecresq.2006.10.002

Mashburn, A. J., & Henry, G. T. (2004). Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. *Educational Measurement: Issues and Practice, 23,* 16–30. doi:10.1111/j.1745-3992.2004.tb00165.x

McWayne, C. M., Cheung, K., Wright, L. E. G., & Hahs-Vaughn, D. L. (2012). Patterns of school readiness among head start children: Meaningful within-group variability during the transition to kindergarten. *Journal of Educational Psychology, 104,* 862–878. doi:10.1037/a0028884

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543. doi:10.1007/BF02294825

Merrell, C., & Bailey, K. (2012). Predicting achievement in the early years: How influential is personal, social and emotional development? *Online Educational Research Journal*. Retrieved from http://www.oerj.org/View?action=viewPaper&paper=55

Merrell, C., & Tymms, P. (2011). Changes in children's cognitive development at the start of school in England 2001–2008. *Oxford Review of Education, 37,* 333–345. doi:10.1080/03054985.2010.527731

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3,* 111–121.

Molnár, G. (2011). Playful fostering of 6- to 8-year-old students' inductive reasoning. *Thinking Skills and Creativity, 6,* 91–99. doi:10.1016/j.tsc.2011.05.002

Molnár, G., & Csapó, B. (2011). Az 1–11 évfolyamot átfogó induktív gondolkodás kompetenciaskála készítése a valószínűségi tesztelmélet alkalmazásával [Constructing inductive reasoning competency scales for years 1–11 using IRT models]. *Magyar Pedagógia, 111,* 127–140.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Nagy, J. (1980). *5–6 éves gyermekeink iskolakészültsége* [School readiness among 5- to 6-year-old children]. Budapest, Hungary: Akadémiai Kiadó.

Nagy, J. (1987). *Prefer: Preventív fejlettségvizsgáló rendszer 4–7 éves gyermekek számára* [A test battery for assessment of 4- to 7-year-old children's school entry competencies]. Budapest, Hungary: Akadémiai Kiadó.

Nagy, J., Józsa, K., Vidákovich, T., & Fazekasé Fenyvesi, M., (2004a). *Az elemi alapképességek fejlődése 4–8 éves életkorban. Az eredményes iskolakezdés hét kritikus alapkészségének országos helyzetképe és a pedagógiai tanulságok* [The development of elementary skills between the ages of 4 and 8. A national overview of the seven basic skills needed for academic success and their pedagogic consequences]. Szeged, Hungary: Mozaik Kiadó.

Nagy, J., Józsa, K., Vidákovich, T., & Fazekasné Fenyvesi, M. (2004b). *Diagnosztikus fejlődésvizsgáló és kritériumorientált fejlesztő rendszer 4–8 évesek számára: DIFER programcsomag* [Diagnostic assessment and criterion-oriented development system for 4- to 8-year-olds: The DIFER package]. Szeged, Hungary: Mozaik Kiadó.

Organizsation for Economic Co-Operation and Development. (2010). *PISA computer-based assessment of student skills in science*. Paris, France: Author.

Organisation for Economic Co-Operation and Development. (2011). *PISA 2009 results: Students on line: Digital technologies and performance (Volume VI)*. Paris, France: Author.

Organisation for Economic Co-Operation and Development. (2014). *Skills for life: Student performance in problem solving*. Paris, France: Author.

Price, P., Tepperman, J., Iseli, M., Duong, T., Black, M., Wang, S., . . . Alwan, A. (2009). Assessment of emerging reading skills in young native speakers and language learners. *Speech Communication, 51,* 968–984. doi:10.1016/j.specom.2009.05.001

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529. doi:10.1037/0021-9010.87.3.517

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71,* 849–869. doi:10.1177/0013164410391468

Shepard, L. A. (1997). Children not ready to learn? The invalidity of school readiness testing. *Psychology in the Schools, 34,* 85–97. doi:10.1002/(SICI)1520-6807(199704)34:2<85::AID-PITS2>3.0.CO;2-R

Snow, C. E., & Van Hemel, S. B. (Eds.). (2008). *Early childhood assessment: Why, what, and how*. Washington, DC: National Academies Press.

Snow, K. L. (2006). Measuring school readiness: Conceptual and practical considerations. *Early Education and Development, 17,* 7–41. doi:10.1207/s15566935eed1701_2

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25,* 78–107. doi:10.1086/209528

Tymms, P., Jones, P., Albone, S., & Henderson, B. (2009). The first seven years at school. *Educational Assessment, Evaluation and Accountability, 21,* 67–80. doi:10.1007/s11092-008-9066-7

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70. doi:10.1177/109442810031002

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68,* 5–24.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of Texas assessment of knowledge and skills*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*(3). Retrieved from http://pareonline.net/pdf/v12n3.pdf

Received April 18, 2013
Revision received December 9, 2013
Accepted December 27, 2013 ■