VALIDITY OF COGNITIVE ABILITY TESTS – COMPARISON OF COMPUTERIZED ADAPTIVE TESTING WITH PAPER AND PENCIL AND COMPUTER-BASED FORMS OF ADMINISTRATIONS

Peter ŽITNÝ[1], Peter HALAMA[1], Martin JELÍNEK[2], Petr KVĚTON[2]

[1]Department of Psychology, Faculty of Philosophy and Arts, University of Trnava
Hornopotočná 23, 918 43 Trnava, Slovak Republic
E-mail: peterzitny@gmail.com; peter.halama@savba.sk

[2]Institute of Psychology, The Academy of Sciences of the Czech Republic
Veveří 97, 602 00 Brno, Czech Republic
E-mail: jelinek@psu.cas.cz; kveton@psu.cas.cz

Abstract: The study analyzes and compares the validity of computerized adaptive testing, paper and pencil and computer-based forms of cognitive abilities tests. The research was conducted on a sample of 803 secondary school students (567 paper and pencil, 236 computer-based/computerized adaptive administration; 363 males, 440 females), their mean age was 16.8 years (SD = 1.33). The test set consisted of the Test of Intellect Potential and the Vienna Matrices Test. Overall results showed that the validity of CAT was reasonably comparable across administration modes. Consistent with previous research, CAT selecting only a small number of items gave results which, in terms of validity, were only marginally different from the results of traditional administration. CAT simulated administration of the TIP was roughly 55% and VMT 54% more economical than the traditional version. These results indicate that CAT is a useful way of improving methodology of psychological testing.

Key words: item response theory, computerized adaptive testing, paper and pencil, computer-based, criterion and construct validity, efficiency

Computers have played an integral role in scoring psychological tests virtually since the first electronic computers were developed in the mid-20th century. Over the past several decades, the use of computers has broadened and they have served as a useful tool in the area of psychological testing (for an overview, see Květon, Klimusová, 2002). Today, many psychological tests have computerized versions, but present developments in the area of psychological assessment place emphasis on methodological improvements and the importance of increasing effectiveness (Butcher, Perry, Hahn, 2004). To achieve both precision and efficiency in assessments, computerized adaptive testing (CAT) has been suggested (Wainer, 2000). This assessment tool involves the use of a computer to administer items to respondents and allows respondent's levels of function to be estimated as precisely as desired (i.e., to reach a preset reliability level). The scores for the com-

This research was sponsored by the Grant Agency under the SR VEGA grant no. 1/0228/10, and the Grant Agency under the ČR grant no. 406/09/P284.

puter-based (CB) or computerized adaptive (CA) version of a test that is also a paper-pencil (PP) may unintentionally differ from that of the paper format. If so, the scores from one format would not be comparable to scores from another. Also, the construct being measured might be affected by the change in testing format (e.g., Hol, Vorst, Mellenbergh, 2007; Květon et al., 2007; but cf. Roper, Ben-Porath, Butcher, 1995). In any of these situations, an examinee might receive different scores, depending on the administration mode under which he or she was tested. This is an important consideration, as many instruments are using PP, CB, and CA versions of a test interchangeably and treating the scores as comparable to one another.

Conventional tests usually require that all examinees respond to all items in the test. Also with computerized tests, examinees respond to all items in the test on the computer, but computer-based version of the test has general advantages over traditional paper-and-pencil testing, such as reduced costs for many elements of the testing lifecycle or new advanced and flexible item types, etc. (Mead, Drasgow, 1993). Comparability of CB and PP administration appears to be a localized issue, and depends on the test being examined (cf. Květon et al., 2007). But neither computer-based nor paper-pencil versions of the test are an efficient way of testing, because examinees of low abilities may find it frustrating to attempt all items. Examinees of high abilities may find it boring to go through all the items which seem too easy for them. Adaptive administration by computer can take advantage of this dynamic medium which stands in contrast to the static scheme of computer-based and paper-and-pencil format of instruments, because computerized adaptive testing (CAT) involves the dynamic selection of items to match the performance of a test taker during test administration. Unlike conventional tests in which all the examinees are provided with the same questions, adaptive tests provide different test item sets for each examinee based on that person's estimated ability (or trait) level (see Wang, Kolen, 2001). CAT requires a testing algorithm to select the questions from a pool of calibrated items ("item bank") and control the evaluation process. The testing algorithm is defined as a set of rules specifying the questions to be answered by the examinee and their order of presentation (Thissen, Mislevy, 2000). The testing algorithm of a CAT involves three main components: 1) item selection procedure, 2) ability estimation methodology, and 3) termination criteria (see generally Van der Linden, Glas, 2002).

CAT is an evolutionary step toward future testing methodologies because it consists of an optimally informative set of items given a particular person. Exams based on CAT can achieve at least as good a precision as a paper-pencil test, using considerably fewer items than traditional tests (Embretson, Reise, 2000). In contrast to a static short test, a computerized adaptive test has the advantage of decreasing respondent's (Simms, Clark, 2005) and administrator's burden with little measurement precision loss. That is, a person's ability can be measured precisely with relatively few items (e.g., Halama, 2005; Jelínek, Květon, Denglerová, 2006; Weiss, 2004).

Computerized adaptive testing (CAT) is one of the important applications made possible by item response theory (IRT) methodology (Jelínek, Květon, Vobořil, 2011b;

Weiss, 1985). IRT, also known as latent trait theory, is model-based measurement in which trait level estimates depend on both persons' responses and on the properties of the items that were administered (Embretson, Reise, 2000; Lord, 1980; Urbánek, Šimeček, 2001). IRT satisfies challenges of adaptive testing through: 1) characterizing item variations in a useful way, 2) equating individual scores from different items on a common scale, and 3) determining efficient rules for item selection (Weiss, 1982). These properties are derived from depicting the interaction between an examinee and an item in terms of item parameters and person ability parameter, which are independent of each other (Kingsbury, Houser, 1993). That is, the item parameter estimates, such as difficulty, discrimination, and guessing, are independent of the particular examinees' ability levels used in item calibration (Wainer, Mislevy, 2000). Likewise, the person ability parameter estimates are independent of the particular items administered to examinees (Hambleton, Jones, 1993). Therefore, once an IRT model is fit to the data, each item's characteristics can be fully specified by parameters. In addition, the item and ability scores are reported on the same scale even when different individuals answer different items on the test (cf. Hahn et al., 2006; Hambleton, 2000). In order to take full advantage of the IRT framework, an item bank must be built in which the data fits the IRT model and satisfies the required assumptions (see also Wainer, 2000). In general, the assumptions of IRT models include unidimensionality and local independence. Moreover, the item parameters should be invariant for all respondents (see generally Hambleton, Swaminathan, Rogers, 1991). IRT-based CAT has provided efficient and effective measurement solutions (Jelínek, Květon, Vobořil, 2011a; Weiss, 2004).

Present developments in the area of psychological assessment place emphasis on methodological improvements and the importance of increasing effectiveness. Computerized adaptive testing (CAT) algorithms based on item response theory (IRT) offer attractive opportunities for simultaneously optimizing both measurement precision and efficiency. While converting paper-pencil instruments to computer-based and/or computerized adaptive provides the opportunity for desirable innovations, this conversion process brings with it new challenges in testing. Although computerized adaptive testing can be expected to improve reliability and measurement precision, the increased reliability does not necessarily translate into substantially greater validity. In fact, there is always a danger when changing item content or format that the new test may be measuring a slightly different ability, which may not relate to, or predict outcomes as well as the old test. CAT is administered by computer, and CAT research therefore fits well into the research effort addressing the psychometric comparability of PP, CB, and CA. Two meta-analyses have been reported on this subject: the first one studied potential administration mode effects of CB and CA administration of dichotomous ability items (Mead, Drasgow, 1993), and the second one studied potential administration mode effects of CB administration of the Minnesota Multiphasic Personality Inventory, which also consists of dichotomous items (Finger, Ones, 1999). Both studies showed that CB and CA administration does not greatly affect psychometric quality. In addition,

Mead and Drasgow (1993) concluded that there was no additional effect of adaptive administration. Research that compares CAT and conventional tests also demonstrates substantial similarity between scores from the two procedures (see Cudeck, 1985). Recent theoretical analysis (Žitný, 2011) of 15 research studies from field of ability testing, clinical psychology, personality testing and health care designed to explore the reliability, utility (in terms of item savings) and validity (in terms of correlations with existing tools) of CAT leads to overall findings that CAT provides an effective means of gaining an optimal amount of information needed to answer an assessment question, while keeping time and/or number of items required to obtain that information at a minimum. The fact that the CAT score correlated highly with the score from the full item bank (range r = 0.83 – 0.99) and moderately with the established measures (range r = 0.58 – 0.83) provides the evidence for reliability, validity and comparability of adaptive tools (Žitný, 2011).

RESEARCH GOAL

The goal of the present study was twofold: first, the authors studied the criterion and construct validity of CAT using real-data simulation to compare computerized adaptive testing with paper-and-pencil and computer-based form of two cognitive abilities tests: the Test of Intellect Potential (TIP) and the Vienna Matrices Test (VMT). Also, we dealt with the question, whether the construct validity patterns were comparable across administration modes in terms of gender and residence.

Second, the authors studied item savings (number of items needed to administer) of adaptive versus full-scale PP and CB administration of the TIP and VMT.

RESEARCH DESIGN

The basic design of the study included three data types for comparison, real (PP and CB administrations of TIP and VMT) and simulated (CA administration of TIP and VMT). Thus, the tests used for this study varied only in the test administration mode. In real-data (Post Hoc) simulation studies, a CAT procedure is applied to item response data of items that were administered to participants using fixed length conventional paper-pencil or conventional computerized tests. For simulation of computerized adaptive testing in our study, a real data set was used that contained responses from computer-based version of TIP and VMT. This paper examined the relatively simple item-level adaptive testing format, though many different formats exist (for an overview, see Wainer, 2000). Often, researchers employ simulated data in their CAT systems analyses (Mills, Stocking, 1996); some researchers may be able to utilize their existing PP test data in a post-hoc CAT simulation (e.g., Wang, Pan, Harris, 1999; Weiss, 2005).

To analyze the criterion and construct validity we computed reciprocal correlation between scores of TIP and VMT tests, and correlations between both test scores and the student's school achievement in the subjects of Slovak language, foreign language, and mathematics. By comparing differences across administration modes (PP, CB, CA), we calculated Fischer *z*-transformations to compare this correlation between test scores and school achievement; and effect sizes (Cohen's d) using formulas based on the t-tests for independent samples comparing

the means of test scores within the scope of groups with different administration mode in terms of gender and residence.

Sample

The research was conducted on a Slovak sample of 803 secondary school students (567 paper-pencil and 236 computer based administration of TIP and VMT) enrolled in a "gymnasium" (479 participants, 59.7%) and other secondary schools (324 participants) from all of Slovakia: 363 of them were male (45.2%) and 440 female, their mean age was 16.8 years (SD = 1.33, range 14-21). In the Slovak Republic, "gymnasium" is a type of school providing secondary general education (non-vocational) and prepares students for higher education, comparable to British grammar schools or sixth form colleges and U.S. college preparatory high schools.

We calibrated IRT item parameters on a sample of 567 participants using data from paper-pencil version of TIP and VMT. Calibration sample characteristics (gender, type of school) are shown in Table 1.

To assess the similarity of the criterion and construct validity patterns across three modes of administration (paper-pencil, computer-based, computerized adaptive), 236 participants studying at a "gymnasium" completed the computer-based version of the TIP and VMT.

Participants from the calibration sample of paper-pencil administration, who attended other secondary schools than "gymnasium", were excluded from further analysis of validity and efficiency. This rule yielded a sample size of 243 participants studying at a "gymnasium" to whom tests were administered using the paper-pencil form of tests.

The final design of the study included three data types for analysis of validity and efficiency: real paper-pencil administration (243 participants studying at a "gymnasium"), computer-based administration (236 participants studying at a "gymnasium"), and post-hoc simulated computerized adaptive administration derived by "re-administering" computer-based data (236 participants studying at a "gymnasium") of TIP and VMT tests. Sample characteristics (gender, residence) of paper-pencil and computer-based/computerized adaptive administrations are shown in Tables 2 and 3.

Table 1. Calibration sample characteristics (gender, type of school)

N = 567	Males	Females	Total
Gymnasium	113	130	243
Secondary school	166	158	324
Total	279	288	567

Table 2. Sample characteristics (gender, residence) of paper-pencil administration

N = 243	Males	Females	Total
City	88	77	165
Village	25	53	78
Total	113	130	243

Table 3. Sample characteristics (gender, residence) of CB/CA administration

N = 236	Males	Females	Total
City	48	80	128
Village	36	72	108
Total	84	152	236

Measures

The test set consisted of cognitive abilities tests: the *Test of Intellect Potential* (TIP) and the *Vienna Matrices Test* (VMT). The Test of Intellect Potential (TIP) is a non-verbal method for identification of general intellect abilities via deduction of relations. By its construction, it tries to capture predominantly fluid intelligence, independent of education, and that is why it is close to culture-fair tests (Říčan, 1971). The test consists of 29 items (12-minute time limit) and it can be applied to persons aged 13 years and older. The TIP has two parallel forms, A/B and the B form was used here. The Vienna Matrices Test (VMT), which is similar to Raven's Standard Progressive Matrices, is a non-verbal assessment of the general intelligence based on deductive thinking, and is thus mainly independent of cultural and social backgrounds (Vonkomer, 1992). The test consists of 24 items (25-minute time limit). The items resemble classical matrices, but they are based on explicit construction rules. It can be applied to persons aged 14 years and older.

The traditional paper-pencil version of TIP and VMT contained items presented in standard-booklet order. The TIP-CB and VMT-CB, the computer-based version of the TIP and VMT, was presented over a computer in the same way as the paper-pencil version. Once the examiner started the computer-based administration, the examinee was asked to select a response by clicking with the mouse. The computer automatically saved each response.

To assess the similarity of the criterion and construct validity patterns across three modes of administration (paper-pencil, computer-based, computerized adaptive), participants were also asked about their age, gender, residence and school achievement in subjects of Slovak language, foreign language and mathematics expressed by the final grades at the end of the first semester of the school year; the marks go from 1 (best) to 5 (worst).

For simulation of computerized adaptive testing in this study, the CAT simulation program, CATO – Computerized Adaptive Testing *optimized* (Květon et al., 2008), was used. The CATO program is a user-friendly and understandable application for building, administration and simulation of adaptive tests. In the current stage of development the software is capable of working with dichotomously-scored items and it has one-, two-, and three-parameter IRT models implemented. In CATO, item responses can be input from an external file or generated internally on the basis of item parameters provided by users. The program allows users to choose among methods of setting initial, approaches to item selection, trait estimators, CAT stopping criteria (sufficient precision

of estimation, screening, maximum test-length), and other CAT parameters. In addition, CAT simulation results can be saved easily and used for further study (ordered set of items, answers, trait estimation a respective error). For simulation of computerized adaptive testing in this study, a real data set was used that contained responses from the computer-based version of TIP and VMT. The CAT simulation in this study started with a random selection of three test items (CATO program has only this option). After the simulated participant responded, the program used the response to estimate theta and then searched the remaining items for the single item that provided the most psychometric information at that current trait estimate. The identified item was then administered, followed by *expected a posteriori* (EAP) theta estimation and assessment of the termination rule, which means that the standard error of the trait estimate drops below 0.50. We decided for this stopping rule because this is a common option and is sufficient for our purpose. However, we realize that the nuances of a precision based (and hence a variable test length) stopping rule can be a matter of discussion. This cycle of item selection, theta estimation, and termination rule assessment was repeated until the termination rule described above was satisfied; once met, the adaptive theta estimate, standard error, and the number of items administered were recorded.

Calibration Process

For simulation of computerized adaptive testing, we calibrated IRT item parameters on a sample of 567 participants (279 male, 288 female) using data from the paper-pencil version of TIP and VMT. The size of this calibration sample was adequate for this IRT calibration (see Embretson, Reise, 2000). We estimated IRT item parameters for each of the test using Multilog 7.0.3 (Thissen, Chen, Bock, 2003). We chose to estimate the Three-Parameter Logistic model (3PL) which uses an item response theory model that specifies the probability of a correct response to a dichotomously scored multiple choice item as a logistic distribution. The 3PL extends the 2PL by introducing a guessing parameter. Items vary in terms of their discrimination, difficulty, and probability of guessing a correct response. We assessed scale unidimensionality, underlying most IRT models, by fitting a one-factor model to the items within each test using TESTFact 4 (Wood et al., 2003), software that conducts factor analyses on matrices of tetrachoric correlations. The results support the unidimensionality of tests and provide evidence of their appropriateness for IRT modeling.

RESULTS

Criterion and Construct Validity

To assess the similarity of the criterion validity patterns across modes (PP, CB, CA) we computed correlations between the school achievement and test scores of TIP and VMT. Results of this analysis can be found in Table 4 to 6.

As can be seen, Fischer *z*-transformations revealed a significant difference only between the correlations for students' school achievement in mathematics for the paper and pencil ($r_s = -0.33$) and the computerized adaptive ($r_s = -0.15$; $z = 2.08$, $p = 0.038$) administration of VMT. No other significant differences between the correlations across modes (PP,

Table 4. Spearman criterion validity correlations (PP, CB) and Fischer *z*-transformations

Subjects – school achievement	TIP-PP (r_s)	TIP-CB (r_s)	Fischer *z*	Sig.
Slovak language	- 0.31	- 0.19	1.39	0.165
Foreign language	- 0.22	- 0.21	0.11	0.912
Mathematics	- 0.33	- 0.21	1.41	0.159
	VMT-PP (r_s)	VMT-CB (r_s)		
Slovak language	- 0.27	- 0.15	1.37	0.171
Foreign language	- 0.27	- 0.13	1.59	0.112
Mathematics	- 0.33	- 0.17	1.86	0.063

Note: All correlations are significant ($p \leq 0.05$); PP = paper and pencil (*n*s = 243); CB = computer-based (*n*s = 236)

Table 5. Spearman criterion validity correlations (CB, CA) and Fischer *z*-transformations

Subjects – school achievement	TIP-CB (r_s)	TIP-CA (r_s)	Fischer *z*	Sig.
Slovak language	- 0.19	- 0.19	0.00	1.000
Foreign language	- 0.21	- 0.20	0.11	0.912
Mathematics	- 0.21	- 0.23	- 0.23	0.818
	VMT-CB (r_s)	VMT-CA (r_s)		
Slovak language	- 0.15	- 0.14	0.11	0.912
Foreign language	- 0.13	- 0.12	0.11	0.912
Mathematics	- 0.17	- 0.15	0.22	0.826

Note: All correlations are significant ($p \leq 0.05$); CB = computer-based (*n*s = 236); CA = computerized adaptive (*n*s = 236)

Table 6. Spearman criterion validity correlations (PP, CA) and Fischer *z*-transformations

Subjects – school achievement	TIP-PP (r_s)	TIP-CA (r_s)	Fischer *z*	Sig.
Slovak language	- 0.31	- 0.19	1.39	0.165
Foreign language	- 0.22	- 0.20	0.23	0.818
Mathematics	- 0.33	- 0.23	1.18	0.238
	VMT-PP (r_s)	VMT-CA (r_s)		
Slovak language	- 0.27	- 0.14	1.48	0.139
Foreign language	- 0.27	- 0.12	1.70	0.089
Mathematics	- 0.33	- 0.15	2.08	0.038

Note: All correlations are significant ($p \leq 0.05$); PP = paper and pencil (*n*s = 243); CA = computerized adaptive (*n*s = 236)

CB, CA) for students' school achievement were found.

To analyze the construct validity by comparing differences across administration modes (PP, CB, CA), we transformed raw scores (PP, CB) and the IRT-based score to z-scores metric to correct comparison of different metrics (e.g., raw scores compared with thetas); and subsequently we calculated effect sizes (Cohen's *d*) using formulas based on the *t* statistic resulting from t-tests for independent samples comparing the means within the paper and pencil groups, computerized groups and computerized adaptive groups in terms of gender (Table 7) and residence (Table 8). Comparing differences across administration modes in terms of gender and residence revealed largely comparable effect sizes across modes of administration.

In the next step, a construct validity analysis was performed. As shown in Table 9 below, the construct validity patterns were reasonably comparable across administration modes: Fischer *z*-transformations revealed that there were no significant differences between the correlations across modes (PP, CB, CA) for VMT and TIP.

Table 7. Comparing differences across administration modes in terms of gender

Gender	males M (SD)	females M (SD)	Sig.	*d*
TIP-PP	0.066 (1.04)	0.064 (0.88)	0.985	0.00
TIP-CB	- 0.082 (1.02)	- 0.058 (1.06)	0.868	- 0.02
TIP-CA	- 0.078 (1.00)	0.043 (1.00)	0.376	- 0.12
VMT-PP	- 0.030 (0.96)	0.190 (0.88)	0.063	- 0.24
VMT-CB	- 0.253 (1.10)	- 0.001 (1.05)	0.083	- 0.23
VMT-CA	- 0.138 (1.00)	0.076 (1.00)	0.115	- 0.21

Note: (*n*s = males/females); PP = paper and pencil (*n*s = 113/130); CB = computer-based (*n*s = 84/152); CA = computerized adaptive (*n*s = 84/152)

Table 8. Comparing differences across administration modes in terms of residence

Residence	city M (SD)	village M (SD)	Sig.	*d*
TIP-PP	0.105 (0.98)	- 0.019 (0.90)	0.346	0.13
TIP-CB	0.010 (1.04)	- 0.158 (1.05)	0.220	0.16
TIP-CA	0.063 (1.03)	- 0.075 (0.97)	0.289	0.14
VMT-PP	0.143 (0.91)	- 0.030 (0.95)	0.173	0.19
VMT-CB	0.055 (1.04)	- 0.263 (1.08)	0.022	0.30
VMT-CA	0.089 (0.98)	- 0.106 (1.02)	0.135	0.20

Note: (*n*s = city/village); PP = paper and pencil (*n*s = 165/78); CB = computer-based (*n*s = 128/108); CA = computerized adaptive (*n*s = 128/108

Table 9. Spearman construct validity correlations and Fischer *z*-transformations

TIP – VMT	r_s	r_s	Fischer *z*	Sig.
PP, CB	0.50	0.59	- 1.40	0.162
CB, CA	0.59	0.53	0.94	0.347
PP, CA	0.50	0.53	- 0.44	0.660

Note: All correlations are significant ($p \leq 0.01$); PP = paper and pencil (*n*s = 243); CB = computer-based (*n*s = 236); CA = computerized adaptive (*n*s = 236)

Table 10. Efficiency of adaptive versus full-scale administration of the TIP and VMT

N = 236	Full-item bank number	Number of items selected for CAT simulated administration			
		Median (*n*s)	Mode (*n*s)	Maximum (*n*s)	Minimum (*n*s)
TIP	29	13 (14)	8 (28)	29 (11)	5 (1)
VMT	24	11 (62)	11 (62)	24 (12)	4 (7)

*Note: n*s = number of subjects (respondents)

Efficiency Analysis

The TIP-CA and VMT-CA yielded significant item savings compared with the PP version and full-scale administration on the computer. As can be seen, CAT simulated administration of the TIP was roughly 55% (*Median* = 13) and VMT 54% (*Median* = 11) faster than the traditional version (Table 10). Inspection of the item administration data revealed that the full-item bank was administered adaptively only to 4.7% (11 of 236, TIP) and 5.1% (12 of 236, VMT) of the participants.

DISCUSSION

Computerized adaptive testing (CAT) algorithms based on item response theory (IRT) offer attractive opportunities for simultaneously optimizing both measurement precision and efficiency (Žitný, 2011). Demonstrating that computerized adaptive testing can save item and time cost is necessary, but not sufficient, to establish the utility of this procedure. Therefore, the primary objective of this study was to extend the computerized adaptive literature by comparing real-data simulation of cognitive abilities tests TIP and VMT with paper-and-pencil and computer-based forms of administration. While many tests have been converted from PP to CB and/or CA format to take advantage of the benefits offered by the computer and adaptive technology, it is often the case that the PP version is not replaced, but more frequently both modes are maintained. Thus, equivalence must be established before scores from the computer-based and/or computerized adaptive test form can be used interchangeably with those from the paper-

based test. Recent data provide evidence that traditional measures administered by computer do not differ substantially from standard PP administrations (e.g., Vispoel, Boo, Bleiler, 2001; Williams, McCord, 2006; etc.). However, CAT includes features not typically present in standard or computerized tests, such as differences in item selection and presentation order across participants (Meijer, Nering, 1999). The purpose of the present study was to evaluate validity and efficiency of IRT-based CAT using real-data simulation to compare computerized adaptive testing with paper-and-pencil and computer-based forms of cognitive abilities tests. CAT was simulated by using the existing item responses as if they had been collected adaptively. To assess the similarity of the criterion validity patterns across modes (PP, CB, and CA) for students' school achievement in their subjects of Slovak language, foreign language and mathematics, we computed correlations between the TIP and VMT. Overall, we can conclude, that the criterion validity patterns were reasonably comparable across administration modes and were consistent with findings in other studies (e.g., Wang, Kolen, 2001). The adaptive algorithm resulted in a significant difference only between the correlations for students' school achievement in mathematics for the paper and pencil and the computerized adaptive administration of VMT. No other significant differences between the correlations across modes were found. These findings also suggest that the adaptive algorithm resulted in a small reduction in the strength of the correlations across PP and CB modes for students' school achievement in almost all of the subjects. This finding may possibly be explained by the fact that validity reduction is likely due to the adaptive administration specifically. Thus, this finding represents potential risk associated with the CAT, and perhaps adaptive cognitive testing more generally, that deserves consideration in future CAT projects of this variety, which will involve the administration of real tests to live examinees. To analyze the construct validity by comparing differences across administration modes (PP, CB, CA), we calculated the effect sizes comparing the means within the paper and pencil groups, computerized groups and computerized adaptive groups in terms of gender and residence. Results showed that differences in terms of gender and residence were reasonably comparable across administration modes. Also, we are dealing with the question, whether the construct validity patterns were comparable across administration modes. In a similar manner, there were no significant differences between the correlations across modes (PP, CB, CA) for VMT and TIP (cf. Schaeffer et al., 1998; Schaeffer et al., 1993; Schaeffer et al., 1995). Furthermore, we conducted efficiency analysis (number of items needed to administer) of adaptive versus full-scale administration of the TIP and VMT. The TIP-CA and VMT-CA yielded significant item savings compared with the PP version and full-scale administration on the computer. CAT simulated administration of the TIP was roughly 55% and VMT 54% more economical than the traditional version. Inspection of the item administration data revealed that the full-item bank was administered adaptively only to 4.7% (TIP) and 5.1% (VMT) of the participants. These item savings are consistent with the results of previous research (Becker et al., 2008; Fliege et al., 2009; Hart et al., 2006; etc.). Moreover, these item savings are greater than those typically found for non-IRT CAT applica-

tions (see Handel, Ben-Porath, Watt, 1999). Consistent with previous research (Jelínek, Květon, Vobořil, 2011a; Žitný, 2011), CAT selecting only a small number of items gave results which, in terms of validity, were only marginally different from the results of traditional paper-pencil version. Results showed no essential administration mode effects. This finding indicates that CAT is a very fruitful way of improving methodology and the efficiency of psychological testing.

As with any research study, there are limitations concerning the degree to which the findings can be generalized. It is assumed that the results of this study cannot be generalized to other cognitive and non-cognitive measures; rather, they lend support to the idea that comparability studies must be conducted for each test. The results of the present study suggest that CA and CB versions of TIP and VMT are a largely comparable to the traditional paper-pencil form. The PP, CB and CA administrations yielded to reasonably comparable criterion and construct validity. The adaptive algorithm resulted in significant item savings rather than the CB and PP versions of TIP and VMT. Although our results and those of previous simulation studies have been impressive, it is an open question whether the same findings would be obtained with live participants. Thus, additional Live-CAT studies (administration of real tests to live examinees) are needed to confirm this pattern of findings.

Received October 19, 2011

REFERENCES

BECKER, J., FLIEGE, H., KOCALEVENT, R.-D., BJORNER, J.B., ROSE, M., WALTER, O.B., KLAPP, B.F., 2008, Functioning and validity of A Computerized Adaptive Test to measure anxiety (A-CAT). *Depression and Anxiety*, 25, 12, E182-E194.

BUTCHER, J.N., PERRY, J., HAHN, J., 2004, Computers in clinical assessment: Historical developments, present status, and future challenges. *Journal of Clinical Psychology*, 60, 3, 331-345.

CUDECK, R., 1985, A structural comparison of conventional and adaptive versions of the ASVAB. *Multivariate Behavioral Research*, 20, 3, 305-322.

EMBRETSON, S.E., REISE, S.P., 2000, *Item response theory for psychologists* (Multivariate Applications Book Series). Mahwah, NJ: Lawrence Erlbaum Associates.

FINGER, M.S., ONES, D.S., 1999, Psychometric equivalence of the computer and booklet forms of the MMPI: A meta-analysis. *Psychological Assessment*, 11, 1, 58-66.

FLIEGE, H., BECKER, J., WALTER, O.B., ROSE, M., BJORNER, J.B., KLAPP, B.F., 2009, Evaluation of a Computer-Adaptive Test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research*, 18, 1, 23-36.

HAHN, E.A., CELLA, D., BODE, R.K., GERSHON, R., LAI, J.-S., 2006, Item banks and their potential applications to health status assessment in diverse populations. *Medical Care*, 44, 11, 189-197.

HALAMA, P., 2005, Adaptívne testovanie pomocou počítača: Aplikácia teórie odpovede na položku v diagnostike inteligencie [Computerized adaptive testing: Application of item response theory in intelligence testing]. *Psychológia a Patopsychológia Dieťaťa*, 40, 3, 252-266.

HAMBLETON, R.K., 2000, Emergence of item response modeling in instrument development and data analysis. *Medical Care*, 38, 9, 60-65.

HAMBLETON, R.K., JONES, R.W., 1993, Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12, 3, 253-262.

HAMBLETON, R.K., SWAMINATHAN, H., ROGERS, H.J., 1991, *Fundamentals of item response theory* (Measurement methods for the social science). Newbury Park, CA: Sage Publications.

HANDEL, R.W., BEN-PORATH, Y.S., WATT, M., 1999, Computerized adaptive assessment with the MMPI-2 in a clinical setting. *Psychological Assessment*, 11, 3, 369-380.

HART, D.L., MIODUSKI, J.E., WERNEKE, M.W., STRATFORD, P.W., 2006, Simulated computerized adaptive test for patients with lumbar spine impairments was efficient and produced valid measures of function. *Journal of Clinical Epidemiology*, 59, 9, 947-956.

HOL, A.M., VORST, H.C.M., MELLENBERGH, G.J., 2007, Computerized adaptive testing for polytomous motivation items: Administration mode effects and a comparison with short forms. *Applied Psychological Measurement*, 31, 5, 412-429.

JELÍNEK, M., KVĚTON, P., DENGLEROVÁ, D., 2006, Adaptivní testování - základní pojmy a principy [Adaptive testing - basic concepts and principles]. *Československá Psychologie*, 50, 2, 163-173.

JELÍNEK, M., KVĚTON, P., VOBOŘIL, D., 2011a, Adaptivní administrace NEO PI-R: Výhody a omezení [Adaptive administration of NEO PI-R: Limits and benefits]. *Československá Psychologie*, 55, 1, 69-81.

JELÍNEK, M., KVĚTON, P., VOBOŘIL, D., 2011b, *Testování v psychologii - Teorie odpovědi na položku a počítačové adaptivní testování* [Testing in Psychology: Item response theory and computerized adaptive testing]. Praha: Grada Publishing.

KINGSBURY, G.G., HOUSER, R.L., 1993, Assessing the utility of item response models: Computerized Adaptive Testing. *Educational Measurement: Issues and Practice*, 12, 1, 21-27.

KVĚTON, P., JELÍNEK, M., DENGLEROVÁ, D., VOBOŘIL, D., 2008, Software pro adaptivní testování: CAT v praxi [Software for adaptive testing: CAT in action]. *Československá Psychologie*, 52, 2, 145-154.

KVĚTON, P., JELÍNEK, M., VOBOŘIL, D., KLIMUSOVÁ, H., 2007, Computer-based tests: The impact of test design and problem of equivalency. *Computers in Human Behavior*, 23, 1, 32-51.

KVĚTON, P., KLIMUSOVÁ, H., 2002, Metodologické aspekty počítačové administrace psychodiagnostických metod [Methodological aspects of computer administration of psychodiagnostics methods]. *Československá Psychologie*, 46, 3, 251-264.

LORD, F.M., 1980, *Applications of item response theory to practical testing problems*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

MEAD, A.D., DRASGOW, F., 1993, Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114, 3, 449-458.

MEIJER, R.R., NERING, M.L., 1999, Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23, 3, 187-194.

MILLS, C.N., STOCKING, M.L., 1996, Practical issues in Large-Scale Computerized Adaptive Testing. *Applied Measurement in Education*, 9, 4, 287-304.

ROPER, B.L., BEN-PORATH, Y.S., BUTCHER, J.N., 1995, Comparability and validity of Computerized Adaptive Testing with the MMPI-2. *Journal of Personality Assessment*, 65, 2, 358.

ŘÍČAN, P., 1971, *Test Intelektového Potenciálu - TIP* [Test of Intellect Potential - TIP]. Bratislava: Psychodiagnostické a didaktické testy.

SCHAEFFER, G.A., BRIDGEMAN, B., GOLUB-SMITH, M.L., LEWIS, C., POTENZA, M.T., STEFFEN, M., 1998, *Comparability of paper-and-pencil and Computer Adaptive Test scores on the GRE general test*. Princeton, NJ: Educational Testing Service (Research Report No: RR-98-38).

SCHAEFFER, G.A., REESE, C.M., STEFFEN, M., MCKINLEY, R.L., MILLS, C.N., 1993, *Field test of a computer-based GRE general test*. Princeton, NJ: Educational Testing Service (Research Report No: RR-93-07).

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*. Princeton, NJ: Educational Testing Service (Research Report No: RR-95-20).

SIMMS, L.J., CLARK, L.A., 2005, Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment*, 17, 1, 28-43.

THISSEN, D., CHEN, W., BOCK, R.D., 2003, *MULTILOG 7 for Windows: Multiple-category item analysis and test scoring using item response theory*. Lincolnwood, IL: Scientific Software International, Inc. [Computer software].

THISSEN, D., MISLEVY, R.J., 2000, Testing algorithms. In: H. Wainer (Ed.), *Computerized adaptive testing: A Primer* (pp. 101-133). Mahwah, NJ: Lawrence Erlbaum Associates.

URBÁNEK, T., ŠIMEČEK, M., 2001, Teorie odpovědi na položku [Item response theory]. *Československá Psychologie*, 45, 5, 428-440.

VAN DER LINDEN, W.J., GLAS, C.A.W., 2002, *Computerized adaptive testing: Theory and practice*. New York: Kluwer Academic Publishers.

VISPOEL, W.P., BOO, J., BLEILER, T., 2001, Computerized and paper-and-pencil versions of the

Rosenberg Self-Esteem Scale: A comparison of psychometric features and respondent preferences. *Educational and Psychological Measurement*, 61, 3, 461-474.

VONKOMER, J., 1992, *Viedenský Matricový Test - VMT* [Vienna Matrices Test - VMT]. Bratislava: Psychodiagnostika.

WAINER, H., 2000, *Computerized adaptive testing: A Primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

WAINER, H., MISLEVY, R.J., 2000, Item response theory, item calibration, and proficiency estimation. In: H. Wainer (Ed.), *Computerized adaptive testing: A Primer* (pp. 61-100). Mahwah, NJ: Lawrence Erlbaum Associates.

WANG, T.Y., KOLEN, M.J., 2001, Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement*, 38, 1, 19-49.

WANG, X.B., PAN, W., HARRIS, V., 1999, *Computerized adaptive testing simulations using real test taker responses*. Newtown, PA: Law School Admission Council.

WEISS, D.J., 1982, Improving measurement quality and efficiency with adaptive theory. *Applied Psychological Measurement*, 6, 4, 473-492.

WEISS, D.J., 1985, Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53, 6, 774-789.

WEISS, D.J., 2004, Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 37, 2, 70-84.

WEISS, D.J., 2005, *Manual for POSTSIM: Post-hoc simulation of computerized adaptive testing. Version 2.0*. St. Paul, MN: Assessment Systems Corporation.

WILLIAMS, J.E., MCCORD, D.M., 2006, Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. *Computers in Human Behavior*, 22, 5, 791-800.

WOOD, R., WILSON, D., GIBBONS, R., SCHILLING, S., MURAKI, E., BOCK, D., 2003, *TESTFACT: Test scoring, item statistics, and item factor analysis*. Lincolnwood, IL: Scientific Software International, Inc. [Computer software].

ŽITNÝ, P., 2011, Presnosť, validita a efektívnosť počítačového adaptívneho testovania [Computerized adaptive testing: Precision, validity and efficiency]. *Československá Psychologie*, 55, 2, 167-179.

VALIDITA TESTOV KOGNITÍVNYCH SCHOPNOSTÍ – POROVNANIE POČÍTAČOVÉHO ADAPTÍVNEHO TESTOVANIA S ADMINISTRÁCIOU FORMOU PAPIER-CERUZKA A CEZ POČÍTAČ

P. Žitný, P. Halama, M. Jelínek, P. Květon

Súhrn: Štúdia analyzuje a porovnáva validitu administrácie testov kognitívnych schopností prostredníctvom počítačového adaptívneho testovania s administráciou formou papier-ceruzka a cez počítač. Výskum bol realizovaný na súbore 803 študentov stredných škôl (567 vyplnilo testy formou papier-ceruzka, 236 cez počítač/simulácia CAT; 363 mužov, 440 žien), ich priemerný vek bol 16,8 rokov (SD = 1,33). Testová batéria pozostávala z Testu intelektového potenciálu a Viedenského matricového testu. Celkovo sa z výsledkov ukázalo, že validita CAT bola adekvátne porovnateľná cez jednotlivé formy administrácie. V súlade s predchádzajúcim výskumom, CAT používa len malé množstvo položiek dávajúc výsledky, ktoré, pokiaľ ide o validitu, sú len nepatrne odlišné od výsledkov tradičnej administrácie. Simulovaná CAT administrácia TIP bola zhruba o 55% a VMT o 54% úspornejšia ako tradičné verzie. Tieto výsledky naznačujú, že CAT je užitočný spôsob, ako zlepšiť metodológiu psychologického testovania.