

Item Response Times in Computerized Adaptive Testing

Lutz F. Hornke*

Aachen Technical University, Germany

Computerized adaptive testing (CAT) provides a score as well as item response times. However, the investigation of what additional meaning may result from response time information is of particular interest. Data from 5,912 young men on a computerized adaptive test were available. Earlier studies indicate longer response times for wrong responses. This was replicated in larger settings. However, average item response time for wrong and right responses do not show any differential interpretations of score, nor do they correlate differently with several proficiency tests. Discussion is made as to whether or not response times should be interpreted on the same proficiency dimension as the CAT measured trait or on other dimensions. Since the early 1930's response times have been considered as indicators for personality traits which should be differentiated from scores. This idea is discussed and pro and contra arguments are offered. Recent modeling approaches are also presented. The question remains whether additional diagnostic information is to be gained from CAT with detailed and programmed test-taking protocols.

Key words: computerized adaptive testing, response times, differential predictability

It is often advocated that the mere score of a psychological test does not convey enough information about a test taker given the costly testing situation. For this reason young psychologists are advised to observe their test takers noting strategies, individual idiosyncrasies, or other behavioral patterns in order to improve their interpretations. However, with the advent of computerized testing (Green, 1970, Hornke, 1976) the machine could be programmed to take over some of these detailed observations. Item and test response times are possible observations that can be used for diagnostic purposes to yield additional differential information..

If two test takers work on the same power test, get the same score, and differ in their working time, wouldn't it be wise to hire the person who worked more quickly? In this case it is not the test scores that stipulate the

* Prof. Dr. Lutz F.Hornke, Department of Psychology, Aachen Technical University, D-53056, Aachen (Germany). Tel.+49-241-806013, Fax+49-241-8888.31.
Email:LFHornke@psycho.rwth-aachen.de.

decision but the work speed. In other instances two applicants may differ slightly, the lower ranking person making no errors in the same time that the other makes 20. Whom to hire now?

This central issue was addressed by Margaret Kennedy (1930): "There is a popular theory that some people are of a slow, stolid type and other of a quick, nervous type. The slow type is supposed to plod along persistently with great care for details and accuracy. The quick type, ..., works in a more slap-dash fashion, has little regard for details, and is inclined to be inaccurate. These types are considered to be the result of temperament, not of difference in intelligence" (p. 286). Her literature review of publications earlier than 1930 revealed that this topic had been studied quite often showing that speed-intelligence score correlations ranged from approximately .30 to .80. It seems warranted to ask whether or not it is possible to extract more diagnostic information by considering the amount of mental work and the speed in which it is accomplished: Intelligence and personality diagnosis from applying just one measure.

Discussion about achievement level and time to achieve is raised again (Carroll, 1993, Iseler, 1970). Psychometric interpretations are different from cognitive ones. The former compare across individuals whereas the latter are interested in processes within individuals. At present little is known about the possible diagnostic surplus of item response latencies, testing times in power test environments, and the like. It seems that response time is used as trait indicator as such, if the (cognitive) task is very simple. Models for this case are presented by Samejima (1983) and Scheiblechner (1985) among others. Schnipke & Scrams (1999), in a lucid report, quote Samejima on this issue "... that response times for more complicated tasks [as in the matrices items used below] would require more complicated modeling approaches because the response time will have a less straightforward relationship to the cognitive process of interest. ... For such [simple] tasks, all test takers could probably correctly respond to each item given sufficient time, so errors are likely to be caused by time urgency rather than item difficulty as defined by IRT" (p. 5). They report further that Tatsuoka & Tatsuoka (1980) modeled response times to classify examinees according to their item solving strategy. It is Thissen (1983), they argue, who integrated item accuracy and response time in his model for the first time. More complex problem items can be handled by this IRT-model. Their analysis of models by Verhelst, Verstralen & Jansen (1997) and Roskam (1997) comes closest to what was envisioned above: The examinee's momentary ability is a mix of his mental ability and the time he devotes to the item (series). A right answer is given if the examinee has sufficient ability and stays on working on the problem until he finds the solution. An incorrect answer becomes more likely when the

examinee shifts away from the problem and gradually ceases to work on it, despite his sufficient ability.

It is only recently that some researchers address this problem at the item level. Schnipke & Scrams (1997) discuss “solution behavior” and “rapid guessing behavior” which they propose to have assessed separately. At an item level two aspects can be differentiated: “item not reached” as an aspect of pure speediness and “rapid guessing behavior” (s. Fig 8, p. 226). The latter does not yield information about a test taker’s mental ability. It just represents the odd tendency to gain scores by little or no mental work. Test takers “may skim items briefly for keywords, but they do not thoroughly read the item. Consequently, item characteristics, such as difficulty, length, and content may have little effect on response times” (Schnipke & Scrams, 1997, p. 214). However, it was concluded that “...[I]deally on a CAT, time limits would be relaxed, and rapid-guessing behavior would not be an issue” (p. 230).

The following data analysis attempts to reveal what might be gained from test scores plus response latencies. That there is much to discover was shown by Beckmann (2000), Rammsayer (2000), Beckmann, Guthke & Vahle (1998) and Hornke (1994, 1997).

METHOD

To investigate the relationship of response times and test scores in due detail, data from a larger study with computerized adaptive testing of general mental ability is used. Test takers took an adaptive matrices test (see Hornke, 1976, Hornke & Habon, 1986). Response time to any item was limited to generous 180 seconds and there was no fixed testing time. However, with a maximum of 30 items, in theory, testing time was limited to 90 minutes. Items were drawn according to individual performance from a large pool of 456 items. All items had been precalibrated by means of the two parameter logistic test model (Lord & Novick, 1968). Both item parameters, difficulty and discrimination, were used during testing at each step to find the item from the pool which would provide the most information. Testing stopped when a standard error of measurement (SEM) of .38 or less was reached. This is equivalent to a reliability of .85 in any case and at any score level. All items were administered on a CRT screen in a Windows 3.1 environment. Test takers were n=5912 drafted young men for prescreening for the German military services (Hornke, 1999a).

RESULTS

A first glance at general descriptive results reveal quite interesting trends. It is noteworthy that this kind of computerized adaptive testing works very well with as few as five items and that 90 percent of the population had to work on 10 or less items (see Figure 1). On the average some 7.5 items were presented. From all 456 items available in the item pool the adaptive algorithm used only 93 items based on their contributing information. Contrary to what Revuelta & Ponsoda (1998) suggested, there was no exposure rate control mechanism employed. In contrast to their study only 20 percent of the items from the item pool were used. This indicates that there are sufficient parallel items in the pool and the maximum information algorithm pulls out the locally best suited items. This is fully in line with earlier simulation studies (Hornke, 1999a, b, c).

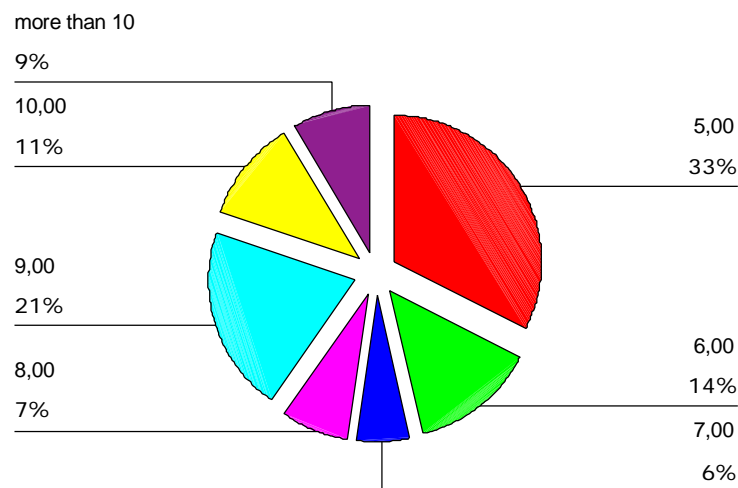


Figure 1. Number of items used in adaptive testing until reaching an SEM of .38 and less (overall percentages are given in bold face)

Secondly, the descriptive statistics reveal quite a satisfactory picture (see Table 1). As expected from the previous norming study the overall mean of Theta-score-estimates was nearly zero with a standard deviation of nearly one. Also the computed standard error of measurement (SEM=.362) undershoots, on average, the preset level of .38 or an equivalent of a reliability of .85 at all score levels.

Table 1. Computerized adaptive test data

Descriptive Statistics			
	AM	SD	N
THETA	,068	1,061	5912
SEM	,362	,022	5912
N used	7,536	2,724	5912
AM(right)	54,467	25,849	5908
AM(wrong)	82,460	42,309	5911
Total Testing Time	633,252	350,334	5912
Time for Instruction	384,036	138,520	5912

Note. THETA represents the maximum likelihood estimated person score of the adaptive test; SEM is the standard error of measurement; N used gives the number of items used until reaching the stopping criterion SEM; AM(right) and AM(wrong) represent times in seconds used to get items right or wrong, respectively.

Thirdly, testing time is roughly two thirds of the total test time, or a bit less than twice the instruction time. Test takers may feel that considerable effort is required for instruction with regard to the short testing time.

Last, and most striking, previous tendencies for wrong and right response times were replicated. As Hornke (1994, 1997) and Rammsayer (2000) found, wrong responses required more time than right responses. In a previous study the median odds were $98[\text{wrong}]/78[\text{right}]=1.25$ for students (see Hornke, 1997) whereas for a general population this turns out to be $75/51=1.47$. However, there was a slight tendency for longer response times to widen the gap between wrong and right response times. At all percentiles the relation held that wrong responses required more time (see figure 2).

Inspection of the corresponding time distributions reveals that response times for wrong responses are more spread than those for right responses (see figure 3). As expected both distributions are mildly left skewed.

Total testing times and Theta score estimates correlated at .65 in a previous study (Hornke, 1997) and in this study they yield an impressive r of .65 (see Table 2). If adaptation is disregarded, i.e. the varying number of items test takers are required to answer, and controls for the number of items used, then a correlation of .50 remains. The same correlation between AM(wrong) and AM(right) holds if Theta is controlled for. This indicates that the total response times represent something not substantially correlated

with proficiency. In addition it is surprising that time spent on instruction is not substantially related to overall proficiency.

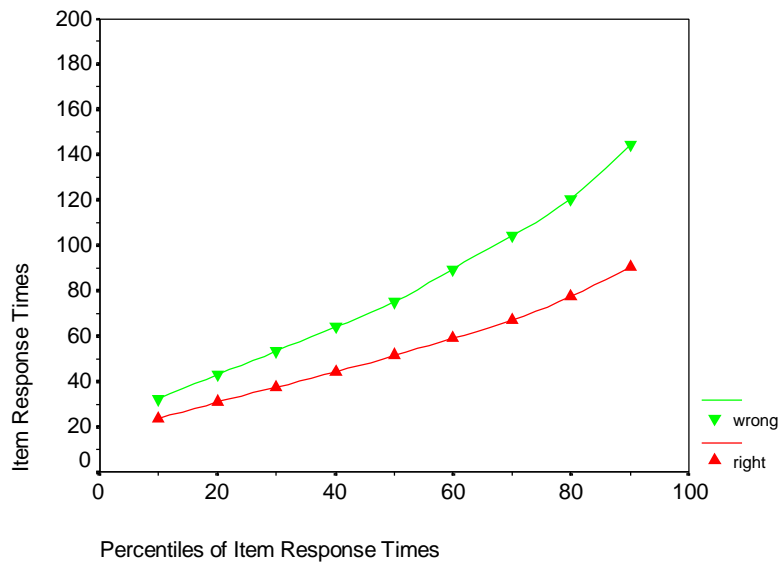


Figure 2. Mean response times differentiated for wrong and right responses.

However, total testing time and the mean response times correlate quite well, but it has to be kept in mind that they represent to a large extent part whole relations.

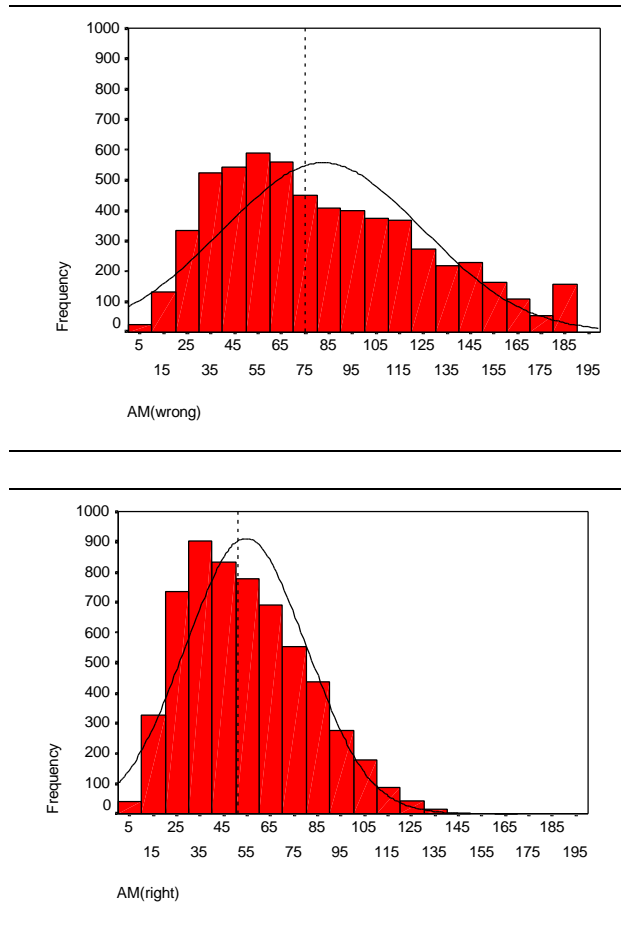


Figure 3. Distribution of mean response times for wrong and right responses. Vertical dotted lines indicate the respective medians (AM(wrong) is slightly distorted by a preset time limit of 180 seconds)

Table 2. Correlations of scores, SEM, and response times

Correlations

	THETA	SEM	N used	AM(right)	AM(wrong)	Total Testing Time	Time for Instruction
THETA	1,000	,133	,476	,603	,580	,648	-,085
SEM	,133	1,000	,412	,166	,130	,330	-,008
N used	,476	,412	1,000	,350	,275	,770	-,021
AM(right)	,603	,166	,350	1,000	,673	,755	,246
AM(wrong)	,580	,130	,275	,673	1,000	,726	,290
Total Testing Time	,648	,330	,770	,755	,726	1,000	,186
Time for Instruction	-,085	-,008	-,021	,246	,290	,186	1,000

DISCUSSION

Data from such a large sample support the advocacy of adaptive testing as an effective means of testing. With the 2PL-test model chosen, testing is quite short for most test takers ending with as few as 5 to 10 items. However, it has to be mentioned that the overall reliability level was set at .85 indicating a moderate, but in many cases, reasonable decision confidence at all score levels. The number of items is by far less than with conventional tests or alternative test models. With 16 to 24 items and the one parameter test model (Rasch-model, 1PL, see Lord & Novick, 1968) Hornke (1999a, b) was able to achieve a reliability of .80. Using the same model but a higher desired reliability the number of items would necessarily increase. Overall, these empirical results mirror earlier simulation results (see Hornke, 1999b) and clearly show that adaptive testing is an economical testing approach that does not sacrifice reliability!

However, the most striking result is the replicable difference in response times for wrong and right responses in an adaptive testing environment. Again one might expect that individual AM(wrong)- and AM(right)-scores are indicative of different processes. Whereas the latter may represent an effort based response time of a successful mental process, the former may represent a time of a lesser effort plus a loss of interest leading to failure. In some cases test takers give up mental effort on an item and guess, thus indicating that they may be employing a different cognitive process. It may be worthwhile to use these different time scores to differentiate between test takers in general. It is expected that there will be a distinct correlational pattern with other tests.

Table 3 shows the overall statistics for tests used in the screening procedure. It is a sign of the overall validity of this study that parallel test forms yield similar statistics. The variation of the number of test takers is due to the demands of the special diagnostic screening program.

Table 3. Means and standard deviations of additional tests

Test Name	Test Characteristic (all test are presented as computerized versions)	Arithmetic Mean	Standard Deviation	N
MTA	Matrices, Form A	11,92	3,38	2183
VAA	Verbal Analogies, Form A	9,67	3,31	3164
RTA	Number Problems, Form A	8,41	4,48	3156
EKA	Electronic Test, Form A	10,04	4,34	412
RSA	Spelling Test, Form A	32,58	12,50	1732
MKA	Mechanical Test, Form A	9,85	3,97	2871
MTB	Matrices, Form B	12,71	3,69	2755
VAB	Verbal Analogies, Form B	10,04	3,40	2757
RTB	Number Problems, Form B	9,20	4,60	2748
EKB	Electronic Test, Form B	9,69	4,77	366
RSB	Spelling Test, Form B	30,60	12,55	1564
MKB	Mechanical Test, Form B	9,87	4	2435
RPA	Reaction Test, Form A	44,89	17,71	2730
DOA	Doppler Auditory Test, Form A	11,90	3,59	146
SIA	Signal Detection Test, Form A	13,63	4,86	931
RPB	Reaction Test, Form B	45,61	17,24	2332
DOB	Doppler Auditory Test, Form B	12,59	3,32	143
SIB	Signal Detection Test, Form B	13,62	5	869

More revealing is Table 4 with correlations of the CAT-Matrices scores and all additional tests in the screening battery. Fortunately the CAT-Matrices and Theta, correlate well with the conventional Matrices test presented on CRT, $r=.716$ and $.749$, respectively. If one takes their reliabilities into account there is not very much more to be expected as far as their intercorrelation is concerned ($r \approx .749/\sqrt{.85 * .80} = .90$): True Matrices scores correlate substantially well! In addition, one observes quite similar correlations with all other tests, their parallel forms, and the CAT-Matrices score, which underscores the validity of the screening procedure and strongly argues in favor of CAT-testing.

Table 4. Correlations of computerized adaptive test scores with some additional tests used in a screening battery.

	THETA	N used	Total Test Time	Instruction Time	AM(right)	N (right)	N (wrong)	AM(wrong)
MTA	,716	,288	,442	-,113	,399	,570	-,270	,468
	2183	2183	2183	2183	2182	2183	2183	2182
VAA	,480	,229	,242	-,298	,165	,399	-,166	,187
	3164	3164	3164	3164	3163	3164	3164	3163
RTA	,559	,261	,288	-,278	,203	,455	-,189	,273
	3156	3156	3156	3156	3155	3156	3156	3155
EKA	,421	,150	,140	-,125	,137	,280	-,192	,169
	412	412	412	412	412	412	412	412
RSA	,475	,229	,237	-,407	,162	,391	-,140	,184
	1731	1731	1731	1731	1730	1731	1731	1730
MKA	,509	,227	,274	-,142	,204	,402	-,172	,249
	2871	2871	2871	2871	2870	2871	2871	2870
MTB	,749	,354	,507	-,115	,472	,603	-,254	,505
	2755	2755	2755	2755	2753	2755	2755	2755
VAB	,407	,161	,149	-,324	,092	,315	-,174	,120
	2757	2757	2757	2757	2755	2757	2757	2757
RTB	,586	,256	,312	-,281	,256	,452	-,209	,310
	2748	2748	2748	2748	2746	2748	2748	2748
EKB	,456	,224	,229	-,065	,222	,334	-,149	,251
	366	366	366	366	366	366	366	366
RSB	,458	,212	,223	-,427	,148	,376	-,172	,180
	1562	1562	1562	1562	1559	1562	1562	1562
MKB	,547	,240	,306	-,137	,250	,428	-,190	,298
	2434	2434	2434	2434	2431	2434	2434	2434
RPA	,440	,203	,222	-,264	,134	,373	-,167	,190
	2732	2732	2732	2732	2731	2732	2732	2731
DOA	,285	,115	,060	-,085	,088	,203	-,108	-,003
	145	145	145	145	145	145	145	145
SIA	,255	,169	,111	-,173	,011	,210	-,025	,099
	930	930	930	930	930	930	930	930
RPB	,441	,171	,227	-,227	,189	,354	-,195	,228
	2331	2331	2331	2331	2328	2331	2331	2331
DOB	,164	,112	,073	-,121	,045	,133	,004	-,025
	143	143	143	143	143	143	143	143
SIB	,336	,171	,131	-,199	,103	,242	-,096	,105
	869	869	869	869	868	869	869	869

Note. Centered entries represent Pearson correlation coefficients and off centered entries give the number of test takers available for the correlation. Test characteristics see Table 3.

There is however, *no* indication that the respective average time scores for wrong and right responses correlate differently with the other tests used in screening applicants (see figure 4).

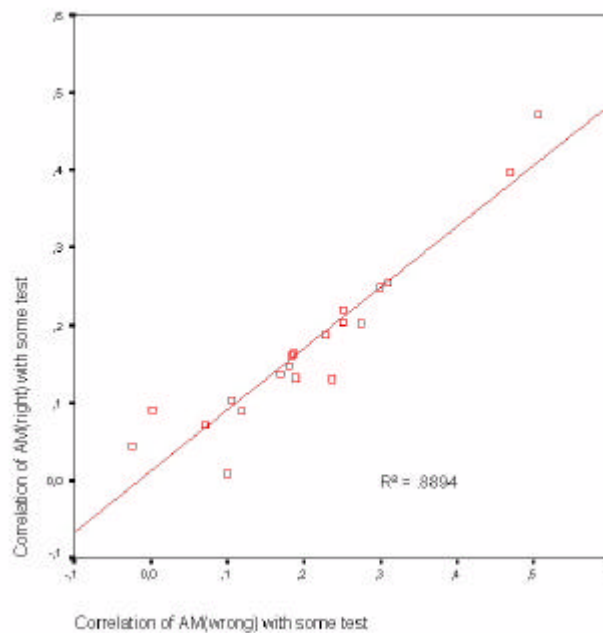


Figure 4. Correlations of all other tests with both mean response times

They do *not* indicate differential predictability even with tests that are speed prone. Currently, achievement tests with differential information from right/wrong response times in CAT-Matrices do not yield additional diagnostic insight or predictability. In addition, an often cited result of Mollenkopf (1960) could not be replicated. He found that “2. The same test material given under speeded conditions may not measure the same behavior as under power conditions” (p. 229). The power-CAT correlates very high with the speeded computer based Matrices tests; $r_{\text{righted for unreliability}} \approx .90!$

In need of investigation is the assumed correlation of speed of right or wrong responses with personality traits as alluded to by Margaret Kennedy (1930) or described as Reflection-Impulsivity by Messer (1976). Other researchers have considered cognitive styles that appear as of yet

inconclusive (Messick, 1984, Tiedemann, 1988). Still others have looked at response times with regard to test takers' strategies (Nährer, 1982, Tatsuoka & Tatsuoka, 1980). However, all this awaits further empirical study and theoretical elaboration.

In addition, investigation is necessary to determine whether differentiated speed influence is linearly related to test scores or if some other interactive relation holds. This question stems from the observation in Figure 5, where right solutions for item response times larger than $\Theta = +0.75$ do not increase to the same degree as before. It seems as though time to solve an item correctly remains more or less constant despite ability.

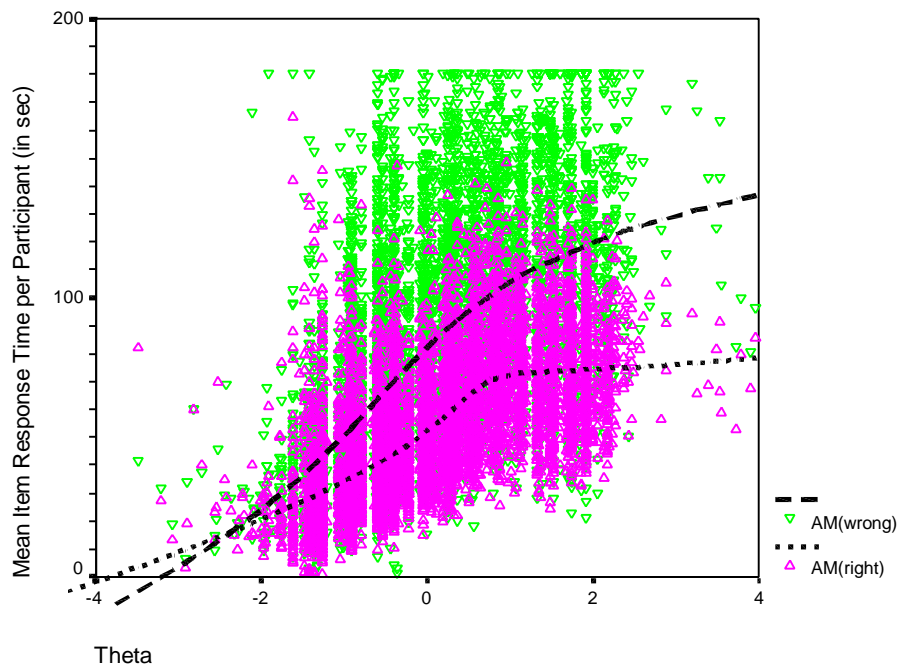


Figure 5. Mean item response times and Theta scores compared for items solved correctly or incorrectly (Both lines represent locally iteratively weighted least squares)

All the psychometric modeling and cognitive theorizing can be of great value if they unite. This would enable researchers to capture what an examinee does while taking tests, not simply evaluating him on an abstract scale and comparing his results to empirical norms. However, sophisticated observations, whether machine or experimenter-based, are needed to yield an

added value from accuracy and time data on CAT which goes beyond its already proven economical gain.

RESUMEN

Tiempos de respuesta al ítem en tests adaptativos informatizados. Los tests adaptativos informatizados (TAI) proporcionan puntuaciones y a la vez tiempos de respuesta a los ítems. La investigación sobre el significado adicional que se puede obtener de la información contenida en los tiempos de respuesta es de especial interés. Se dispuso de los datos de 5912 jóvenes en un test adaptativo informatizado. Estudios anteriores indican mayores tiempos de respuesta cuando las respuestas son incorrectas. Este resultado fue replicado en este estudio más amplio. No obstante, los tiempos promedios de respuesta al ítem para las respuestas erróneas y correctas no muestran una interpretación diferencial de la obtenida con los niveles de rasgo, y tampoco correlacionan de manera diferente con unos cuantos tests de capacidad. Se discute si los tiempos de respuesta deben ser interpretados en la misma dimensión que mide el TAI o en otras dimensiones. Desde los primeros años 30 los tiempos de respuesta han sido considerados indicadores de rasgos de personalidad que deben ser diferenciados de los rasgos que miden las puntuaciones del test. Esta idea es discutida y se ofrecen argumentos a favor y en contra. Los acercamientos más recientes basados en modelos también se muestran. Permanece abierta la pregunta de si se obtiene o no información diagnóstica adicional de un TAI que tenga una toma de datos detallada y programada.

Palabras clave: Item response times, adaptive computerized testing, response time meaning.

REFERENCES

- Beckmann, J.F. (2000). Differentielle Latenzzeiteffekte. *Diagnostica* (im Druck).
- Beckmann, J.F., Guthke, J. & Vahle, H. (1997). Analysen zum Zeitverhalten bei computergestützten adaptiven Intelligenz-Lerntests. *Diagnostica*, 43, 40-62.
- Carroll, J.B. (1993). *Human cognitive abilities - A survey of factoranalytic studies*. New York : Cambridge University Press.
- Green, W.H. (1970). Some comments on tailored testing. Chapter IX. In: W.H. Holtzman (Ed.). *Computer assisted instruction, testing, and guidance*. New York: Harper and Row.
- Hornke, L.F. & Habon, M.W. (1986). Rule based item construction and evaluation within the linear logistic framework. *Applied Psychological Measurement*, 10, 360-380.
- Hornke, L.F. (1994). Erfahrungen mit der computergestützten Diagnostik im Leistungsbereich. In: Bartussek, D. & Amelang, M. (Hrsg.). *Fortschritte der*

- Differentiellen Psychologie und Psychologischen Diagnostik*. Göttingen: Hogrefe. 321-332
- Hornke, L.F. (1999a). Item Generation Models for Higher Order Cognitive Functions. In: Sidney Irvine. *Item Generation*. Lawrence Erlbaum : Hillsdale, N.J. (in press).
- Hornke, L.F. (1999b). Benefits from computerized adaptive testing as seen in simulation studies. *European Journal of Applied Psychology*, 15, 91-98.
- Hornke, L.F. (1999c). Benötigte Itemanzahlen beim meß- und entscheidungsorientierten adaptiven Testen. *Zeitschrift für Differentielle und Diagnostische Psychologie* (in press).
- Hornke, L.F. (1997). Untersuchung von Itembearbeitungszeiten beim computergestützten adaptiven Testen. *Diagnostica*, 43, 27-39.
- Hornke, L.F. (1976). *Grundlagen und Probleme adaptiver Testverfahren*. Frankfurt: Haag+Herchen.
- Iseler, A. (1970). *Leistungsgeschwindigkeit und Leistungsgüte*. Weinheim : Beltz.
- Kennedy, Margaret (1930). Speed as a personality trait. *Journal of Social Psychology*, 1, 286-298.
- Lord, F.M. & Novick, M.R. (1968). *Theory of mental test scores*. Addison-Wesley: Reading, Mass.
- Messer, S.B. (1976). Reflection-Impulsivity: A Review. *Psychological Bulletin*, 83, 1026-1052.
- Messick, S. (1984). The nature of cognitive styles: Problems and promise in educational practice. *Educational Psychologist*, 19, 59-74.
- Mollenkopf, W.G. (1960). Time limits and the behavior of test takers. *Educational and Psychological Measurement*, 20, 223-230.
- Nährer, W. (1982). Zur Beziehung zwischen Bearbeitungsstrategie und Zeitbedarf bei Denkaufgaben. *Zeitschrift für experimentelle und angewandte Psychologie*, 24, 147-159.
- Rammsayer, T. (2000). Zum Zeitverhalten beim computergestützten adaptiven Testen: Antwortlatenzen bei richtigen und falschen Lösungen. *Diagnostica* (im Druck).
- Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 311-327.
- Roskam, E.E. (1997). Models for speed and time-limit test. In W.J. van der Linden and R.K. Hambleton (Eds.) *Handbook of modern item response theory* (pp. 187-208). New York : Springer.
- Samejima, Fumiko. (1983). A latent trait model for differential strategies in cognitive processes (*Technical Report ONR/RR 81-1*). Knoxville, TN : University of Tennessee.
- Scheiblechner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S.E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 219-244). Orlando, FL : Academic Press
- Schnipke, Deborah L. & Scrams, D.J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34, 213-232.
- Schnipke, Deborah L. & Scrams, D.J. (1997). Representing response-time information in item banks. *Law School Admission Council, Report 97-09*.
- Schnipke, Deborah L. & Scrams, D.J. (1999). Exploring Issues of test taker behaviour: Insights gained from response-time analyses. *Law School Admission Council, Report 98-09*.

- Scrams, D.J. & Schnipke, Deborah L. (1999). Response-time feedback on computer-administered tests. Paper presented at the *1999 Meeting of the National Council on Measurement in Education*, Montreal.
- Tatsuoka, K.K. & Tatsuoka, M.M. (1980). A model for incorporating response-time data in scoring achievement tests. In D.J. Weiss (Ed.) *Proceedings of the 1979 computerized adaptive testing conference* (pp. 236-256). Minneapolis, MN : University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Thissen, D. (1983). Timed Testing: An approach using item response theory. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179-203). New York : Academic Press.
- Tiedemann, J. (1988). Zur Diagnostik kognitiver Stile. *Diagnostica*, 34, 289-300.
- Verhelst, N.D., Verstralen, H.H.F.M. & Jansen, M.G.H. (1997). A logistic model for time-limit tests. In W.J. van der Linden and R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169-185). New York : Springer.