Response time in computer-aided testing:  
a “Verbal Memory” test for routes and maps  

LUTZ F. HORNKE¹

Abstract

Computerized testing, be it assisted or adaptive, provides not only the test scores but also the item response times. It is therefore particularly interesting to investigate what additional meaning can be gained from the item response times. This paper draws partly on the test- and item-based results for 1622 young men who took part in a computerized verbal memory test.

Since the early 1930’s response times have been considered as indicators of personality traits which should be differentiated from ability scores. Recent psychometric models for response times have also adopted this approach. However, the question as to whether any significant diagnostic information can be gained from CAT with detailed and programmed test-taking protocols, remains unresolved.

Earlier studies show that the response times for wrong solutions is noticeably longer than for correct solutions. This was confirmed in the present study. Average item response times for wrong and right responses do not show a marked correlation with scores (.387 and .290 respectively). However, times for wrong answers correlate higher with scores than times for correct answers.

Key words: response latencies, verbal memory, computerized testing, F > C phenomenon

¹ Correspondence concerning this article should be addressed to Prof. Dr. Lutz F. Hornke, RWTH-Aachen, Institut für Psychologie, D-52056 Aachen, phone: +49-241-8096013; email address: Lutz.Hornke@psych.rwth-aachen.de
Objective/Rationale

It is essential that people can find their way on the basis of route descriptions. People differ in their ability to follow route descriptions or to give good directions to others. Therefore, orientation is a cognitive ability and accomplishment, which can according to Thorndyke’s theory (1982, cf. Hornke & Kluge-Kläßen, 1991) be divided into knowledge about

- Perceptual icons
- Procedures and
- Overviews

Regarding perceptual icons, Thorndyke and Hayes-Roth (1982) argue that when looking for orientation in a new environment or a foreign town one memorizes prominent places/locations which serve as perceptual icons. But these icons alone are not enough for orientation, because they need to be supplemented by procedural knowledge, which gives further support in the form of when-then clauses: “When I reach the petrol station, then I have to turn right.” On top of that there has to be the overview, which is a symbolical abstraction of the perceptual icons and the procedures. This comprises not only knowledge of particular places/locations, not even knowledge of the meaning of these places/locations (“opera house”, “post office”, “petrol station”), but also spatial information like direction and distance which are not always instantaneously accessible to direct perception (Tolman, 1932), “navigatory knowledge”.

Kosslyn (1976) as well as Thorndyke (1982) assume that these parts of knowledge are memorized propositionally as well as pictorially. These considerations led to a corresponding test development (cf. Hornke & Kluge-Kläßen, 1991), the example items of which are described below (Figure 1).

Since this involves “cognitive processes” in the function of problem-solving, observations about individual differences and the conclusions that follow from them are relevant. Here, the ratio between the speed of performance and the quality of performance lies in the focus in order to study the relation between the response times and the usual test score (s. Hornke, 1997, 2000).

Because there is a difference between the “quality of performance” and the “speed of performance” (cf. Iseler, 1969), the one test can yield more differential information than the usual “quality of performance” or “test score”/ “IQ” tests alone. Young diagnosticians are especially advised to observe their test persons carefully (response behaviour, personal work organisation, emotional interjections and the like) in order to critically check and improve their interpretation. With computerized testing (Green, 1970; Hornke, 1976), some observation tasks can be transferred to the computer – the key words here being “response registration” and “response time registration”.

If two test persons obtain the same score in the same test but differ in the time spent, would it not be sensible to employ the person who works faster? Here, the processing speed and not the test score would decide about the selection. In a different case the applicants might show only small differences regarding their efficiency but the “weaker” person makes no mistakes while the other one makes 20 mistakes. Who should be employed now?

This thought was already expressed by Margaret Kennedy (1930): “There is a popular theory that some people are of a slow, stolid type and others 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). A literature survey of the time before 1930 shows that this issue had already been occasionally studied. The observed correlation between quality of performance and speed was between 0.30 and 0.80. This was later pursued by Iseler (1970), Carroll (1993), Neubauer (1995), who drew a clear distinction between psychometric and cognitive interpretation. Psychometric interpretation compares persons while cognitive focuses on processes within persons. At the moment, little is known about the diagnostic surplus value of latencies and processing times in power tests. It appears that they can be used as trait indicators if the cognitive demand is very low. Psychometric models have been developed for such cases (cf. e.g. White, 1973), Samejima (1983) and Scheiblechner (1985). Schnipke & Scrams (1999) refer to Samejima and assert, “... that response times for more complicated tasks [as in the matrices items used below] would require more complicated modelling 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). Further they cite Tatsuoka & Tatsuoka (1980) who use response times in order to classify test persons according to response strategies. Thissen (1983) is the first to present a model integrating quality and speed of performance. However, his model does not differentiate among correct and incorrect responses. Other models by Verhelst, Verstralen & Jansen (1997) as well as Roskam (1997) come closer to the issue raised here: The momentarily observed performance of a test person is a mixture of her or his mental ability and the time spent on the solution. A correct answer is given if the mental ability is present and there is sufficient time for a systematic solution (resource allocation). A wrong answer becomes more probable the more the test person turns away from the given problem and eventually stops to work on it in a systematic way, or forgets alternatives that have already been checked.

On the level of test problems, Schnipke & Scrams (1977) discussed the “answer-giving behaviour” and the “guessing behaviour”, and they recognize two aspects: “’item not reached’ as an aspect of pure speededness and ‘rapid guessing behavior’”. The latter does not yield information about a test taker's mental ability. It just represents the odd tendency to incidently gain scores by little or no mental work. Test takers “may skim items briefly for key words, 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). But for computer-aided testing it is true that: "...[I]deal on a CAT, time limits would be relaxed, and rapid-guessing behavior would not be an issue" (p. 230).

The study below tries to show what results from a combined inspection of the speed and quality of performance, as already raised by Beckmann (2000), Rammsayer (1999), Beckmann, Guthke & Vahle (1997) as well as Hornke (1994, 1997).

Material and procedure

The test material was the computerized test on “verbal memory” (Etzel & Hornke, 1999), in which a sequence of street names depicted on a map has to be memorized. Only one street name at any given moment is shown on the screen.
21 items were assembled for such a test; an example item is shown in Figure 1. On the screen, first the starting position is displayed and then, after varying intervals, the name of the next “stop” is shown for a moment and disappears. The test person has to memorize the names and their order.

Format of response:
By dragging a stop name from the list on the right to the list position on the left.

Figure 1:
VISGED example item.
At the end a selection list of potential stop names appears which have to be dragged to the right position. This results quasi in a “production procedure” for responses. Though it is based on multiple choice, there are always more potential choices than required. The distractors, too, are designed rationally (initial letters, relative control of word length and the like, cf. Hornke & Etzel, 1995).

The construction shown above is based on the following rules for item formation:

- **Complexity** ( = Number of memorized information; 3 ... 9 street names)
- **Pictorial nature** of the street names presented (e.g., "Squirrel Road" should provoke some concrete idea, while “Logic Alley” hardly can lead to concrete associations, which makes the item more difficult) and
- **Presentation time** for the street names / the material to memorize (1.5 ... 2.5 seconds presentation time).

In this way a total of 114 Rasch homogeneous test problems were generated (s. Hornke & Etzel, 1995). The rules of item construction account for 88% of the variance (adj. $R^2 = 0.877$) in the degree of difficulty of items (cf. Hornke, 2002, p. 176).

Altogether, 1622 army members and applicants were tested. The average age was 19.6 years. The distribution of the education level was as follows: Sonderschule or special schools for handicapped children (0.4%), Hauptschule or secondary school basic level (16.6%), Realschule or secondary school medium level (22.7%), Gesamtschule or comprehensive school (6.5%), Gymnasium or grammar school (53.5%).

**Results**

In agreement with previous studies (Ebel, 1953, Beckmann, 2000; Hornke, 1994, 1997; Rammsayer 1999) the average values of the mean response time show the typical relation: Correct answers need less time than wrong answers! After all, the difference between the two in this test amounts to 21 seconds. This demonstrates again that it makes sense to look separately at wrong response times and correct response times (cf. Lohmann, 1989, Lavergne et al., 1997).

There is only a moderate correlation between quality and speed of solution. The correlation of the total score and the mean time for correct solutions is $r = 0.290$, and the correlation of the total score and the mean time for wrong responses is $r = 0.387$. Therefore, the test persons can be differentiated independently according to the dimensions of quality and speed. Other studies (Tab. 1) found only low averages for the correlation between total scores and mean response times for correct answers, $AM(r) = 0.26$, and wrong answers, $AM(r) = 0.31$, but invariably, there is a higher correlation between the mean time for wrong answers and the total score than the mean time for correct answers.

Reading the performance rates separately for the various types of schools, one can get hints about the dependency on the level of education and about selection mechanisms. It is no surprise, that there is the following relation between total scores: [Hauptschule<Realschule<Gymnasium]. But since the PISA study the Gesamtschule occupies a special position, which can be seen here too.
Response time in computer-aided testing: a "Verbal Memory" test for routes and maps.

Figure 2:
Comparison of the average response times.
In the light of the PISA study, it is however worth noting that the medium score at the Gesamtschule is relatively low.

**Discussion**

The present study could demonstrate like Ebel (1953) that the mean response times can be separated by looking at wrong and correct response latencies. The degree to which these two differ is very surprising, because it suggests that

- **Correct answers** are given as the result of a speedy cognitive process: The correct answer seems to “catch the eye” of the test person, while
- **Wrong answers** are rather the result of a longer process. Perhaps item details are repeatedly considered, then discarded, and finally forgotten. The effort to find the answer drags on, and in the end it might be terminated by a random guess.

The temporal relations show up in other criteria and samples, too:
Response time in computer-aided testing: a “Verbal Memory” test for routes and maps

Table 1: Comparison of mean response times for correct and wrong answers in other studies and criteria

<table>
<thead>
<tr>
<th>Study</th>
<th>Mean time for correct answers in seconds</th>
<th>Mean time for wrong answers in seconds</th>
<th>Ratio</th>
<th>Test type</th>
</tr>
</thead>
<tbody>
<tr>
<td>this study</td>
<td>12.14</td>
<td>32.56</td>
<td>0.373</td>
<td>Verbal memory exercise</td>
</tr>
<tr>
<td>AMT-06, Prieler (norming study)</td>
<td>44.82</td>
<td>60.30</td>
<td>0.541</td>
<td>Figural matrices, very easy</td>
</tr>
<tr>
<td>AMT-04, Prieler (norming study)</td>
<td>57.05</td>
<td>100.8</td>
<td>0.565</td>
<td>Figural matrices, very easy</td>
</tr>
<tr>
<td>AMT-05, Prieler (norming study)</td>
<td>49.64</td>
<td>87.59</td>
<td>0.567</td>
<td>Figural matrices, very easy</td>
</tr>
<tr>
<td>AMT-07, Prieler (norming study)</td>
<td>53.24</td>
<td>82.89</td>
<td>0.642</td>
<td>Figural matrices, very easy</td>
</tr>
<tr>
<td>CAT, Hornke, 2000</td>
<td>51.30</td>
<td>75.00</td>
<td>0.684</td>
<td>Figural matrices, computer adaptive test</td>
</tr>
<tr>
<td>CAT, Tosch, 2002</td>
<td>13.61</td>
<td>19.12</td>
<td>0.686</td>
<td>Verbal analogies, computer adaptive test</td>
</tr>
<tr>
<td>Preckel, 2003</td>
<td>304.70</td>
<td>413.98</td>
<td>0.736</td>
<td>Figural matrices, very difficult for intellectually gifted</td>
</tr>
<tr>
<td>ADANA, Beckmann et al., 1997</td>
<td>12.34</td>
<td>16.46</td>
<td>0.750</td>
<td>Adaptive verbal analogy test</td>
</tr>
<tr>
<td>CAT, Hornke 2000</td>
<td>17.50</td>
<td>23.0</td>
<td>0.761</td>
<td>Verbal analogies, computer adaptive test</td>
</tr>
<tr>
<td>Duration discrimination, Rammsayer &amp; Brandler, 2002</td>
<td>0.921</td>
<td>1.209</td>
<td>0.762</td>
<td>Duration of discrimination exercises, milliseconds</td>
</tr>
<tr>
<td>Duration discrimination, Rammsayer &amp; Brandler, 2002</td>
<td>0.785</td>
<td>1.034</td>
<td>0.759</td>
<td>Duration of discrimination exercises, milliseconds</td>
</tr>
<tr>
<td>CAT, Athenstädt, 2002</td>
<td>41.09</td>
<td>52.68</td>
<td>0.779</td>
<td>Figural matrices, computer adaptive test</td>
</tr>
<tr>
<td>Duration discrimination, Rammsayer, 1999</td>
<td>0.941</td>
<td>1.188</td>
<td>0.792</td>
<td>Duration of discrimination exercises, seconds</td>
</tr>
<tr>
<td>Duration discrimination, Rammsayer, 1999</td>
<td>0.752</td>
<td>0.932</td>
<td>0.806</td>
<td>Duration discrimination exercises, milliseconds</td>
</tr>
<tr>
<td>CAT, Hornke, 2000</td>
<td>55.3</td>
<td>66.4</td>
<td>0.833</td>
<td>Arithmetic problems, computer adaptive test</td>
</tr>
<tr>
<td>ADAFI, Beckmann et al., 1997</td>
<td>27.66</td>
<td>33.15</td>
<td>0.834</td>
<td>Adaptive figure sequence test</td>
</tr>
<tr>
<td>AZAFO, Beckmann et al., 1997</td>
<td>62.48</td>
<td>71.7</td>
<td>0.871</td>
<td>Adaptive number sequence test</td>
</tr>
<tr>
<td>Cooperative Test of Recent Social and Scientific Developments, Ebel, 1953 - upper 27%</td>
<td>28.60</td>
<td>31.7</td>
<td>0.902</td>
<td>General knowledge, paper-pencil test</td>
</tr>
<tr>
<td>Cooperative Test of Recent Social and Scientific Developments, Ebel, 1953 - lower 27%</td>
<td>31.3</td>
<td>33.2</td>
<td>0.943</td>
<td>General knowledge, paper-pencil test</td>
</tr>
<tr>
<td>CAT, Wiggers, 2002</td>
<td>38.78</td>
<td>42.50</td>
<td>0.912</td>
<td>Arithmetic problems, computer adaptive test</td>
</tr>
</tbody>
</table>
On average, the ratio of the mean response time for correct answers to the mean response time for wrong answers is about 2:3, and this poses the question as to what the test persons are doing cognitively with the extra one third of time for wrong answers.

Certainly, the total time for working on the test can be indicative, but its correlation with the total score is only $r = 0.089$; a similar result was found by Baxter (1941), Nährer (1982) and Lachmann (1993). Rammsayer and Brandler (2003) show the lack of correlation with intelligence. Hence, the response times could be of particular value for differential diagnosis. Without an appropriate model, nothing much can be said about this at the moment. Often, “cognitive styles” in the sense of Messick (1984: broad versus narrow categorizing, cognitive complexity versus simplicity, field independence versus dependence, levelling versus sharpening, scanning versus focussing, converging versus diverging, automation versus restructuring and reflection versus impulsiveness) have been proposed as dimensions for differential diagnosis, and these can be related to the time measures mentioned above (cf. also Messer, 1976). But Tiedemann (1988) rightly complains that the operationalisation of styles is problematic and doubts whether one should speak of styles at all. Even the formal model of Schnipke and Scrams (1997) does not offer any hints for interpretation; Thissen
Response time in computer-aided testing:
a “Verbal Memory” test for routes and maps

(1983), too, has to admit in respect to formal modelling that “it needs an additional parameter to absorb the relatively consistent difference between ... correct and incorrect responses” (p. 200). Therefore, the above statements should only be regarded as the initial approaches towards an interpretation which needs further cognition and in depth research. Beckmann et al. (1997, p. 57) see in the consideration of latency times a possibility to “characterize post-hoc the motivational state of the test person or the working style ...[but]... We, too, cannot offer a satisfying answer.”

The question arises, what additional importance for differential diagnosis can be attached to distinguishing a dimension of speed independent from the dimension of quality. Baxter (1941) had already observed and formulated this: “Speed and level ... vary independently” (p. 295), which is an argument in favour of incremental validity, if both measures are used for the prediction: "Prediction through the combination of speed and level ... is greater" (p. 296). This is significant even though the correlation between response times, for wrong or correct answers respectively, is only $r = 0.346$. One has to assume that the times for wrong or for correct answers are each diagnostically relevant in a different way.

Figure 5:
Combined decision rule for recruitment.
One option is to formulate a combined decision rule for (recruitment) selection. As an example, Figure 5 depicts the rule that “only those test persons will be accepted who score at least 12 points in the test AND show a mean response time for correct answers in, at most, 14.5 seconds”. Thus one favours candidates with strong and fast solutions. The lower right quadrant contains exactly those cases which have to be looked at more closely. It is obvious that the additional time criterion leads to a restriction in the number of candidates who have to be considered further. Different cut-offs in both dimensions are of course possible and justifiable, depending on the application.

Morrison (1960, p. 231) notes that “scores obtained under time-limit conditions may differ in factor content from scores obtained on the same test under no-time-limit conditions”. While Ebel (1953) is interested in an improvement of the construction and selection of items by considering the response times, Primi (2001) studies response latencies with regard to the cognitive demand of the items. Baxter (1941) succeeds in demonstrating the incremental validity of the response times. Kyllonen (1997) discusses whether deadlines for response times might make sense: “Considerable empirical research is probably necessary to establish the feasibility of treating response time (deadlines) as simply a difficulty-altering facet. It may be that for some kinds of tasks altering deadlines alters the ability being tapped by the task” (pp. 362-363). Ebel (1953) and Beckmann et al. (1997) agree with regard to an interaction between quality and speed of performance; the latter even records that the more capable the test person is, the larger the individual difference between response times for correct and wrong answers turns out to be. Elliott’s and Osburn’s (1965) idea to control the time allowed for each item, is also interesting; “subjects taking the partial paced test attempted significantly more items than subjects taking the same test under unpaced conditions. Moreover, as a result of having attempted more items, the partial paced group made significantly higher scores as compared to the unpaced group”. Would this not mean, in respect of the above mentioned findings, that one should interrupt very long, unsuccessful attempts by the test persons, giving them a good chance to prove their capacity with a new item?

Mollenkopf (1960, p. 228) makes the general comment: "If time is an inherent aspect of a given complex performance, then the time employed should be such as to make time an appropriately significant factor in the resulting scores." Eysenck (1979) even alludes to three independent components of the IQ: "(1) mental speed, (2) persistence, and (3) error-checking" (p. 188), although especially with regard to (1) there have been crucial and well-founded comments (cf. Neubauer, 1995).

In the present study it was undoubtedly demonstrated that with computer assisted diagnostics not only individual quality measures of performance but also additional speed measures can be taken into account. By doing so, the diagnostic situation and the information about the test persons becomes richer and, hopefully, more “gainful”, and there are certainly further fields of application where the information about response times would be extremely useful.
Response time in computer-aided testing: a “Verbal Memory” test for routes and maps

References

Response time in computer-aided testing:
a “Verbal Memory” test for routes and maps