To cite this article:

Partchev, I., De Boeck, P., & Steyer, R. (2013). How much power and speed is measured in this test? *Assessment*, *20*(2), 242-252. https://doi.org/10.1177/1073191111411658

How much power and speed is measured in this test?

Ivailo Partchev

K. U. Leuven

Paul De Boeck

University of Amsterdam, K.U.Leuven

Rolf Steyer

University of Jena

Author Note

Ivailo Partchev, Department of Psychology, K.U.Leuven, presently at Cito, Arnhem (the Netherlands); Paul De Boeck, Department of Psychology, University of Amsterdam, and Department of Psychology, K.U.Leuven; Rolf Steyer, Department of Psychology, University of Jena.

We are grateful to Lutz Hornke for his help with making the data available for this study, to Peter Halpin for his comments on an earlier version, and to the Associate Editor and the reviewers for their suggestions on how to improve the manuscript. The research reported in this paper was supported by grant GOA/05/04 of the K.U.Leuven.

Correspondence concerning this article should be addressed to Paul De Boeck, Department of Psychology FMG, University of Amsterdam, Roeterstraat 15, 1018WB Amsterdam. E-mail: paul.deboeck@uva.

Abstract

An old issue in psychological assessment is whether speed and power are different cognitive performance traits, and if they are indeed, to what extent each is measured by a given test. Starting from accuracy and response time data, we propose an approach based on posterior time limits (cut-offs of recorded response time) leading to three kinds of recoded data: time data (whether or not the response precedes the cut-off), time-accuracy data (whether or not a response is correct *and* precedes the cut-off), and accuracy data (as time-accuracy data, but coded as missing when not preceding the time cut-off). Each type of data can be modeled as binary responses. Speed and power are investigated through the effect of posterior time limits on two main aspects of the data: (i) the kind of latent variable that is measured: whether it is more power-related or more speed-related; (ii) item discrimination and the test information function as an indication of how well the latent variable (of whatever kind) is measured through the item(s). As empirical data, we use responses and response times for a verbal analogies test. The main findings are that, independent of the posterior time limit, basically the same latent speed trait was measured through the time data, and basically the same latent power trait was measured through the accuracy data, while for the time-accuracy data the nature of the latent trait moved from power to speed when the posterior time limit was reduced. It was also found that a reduction of the posterior time limit had no negative effect on the reliability of the latent trait measures (of whatever kind). Finally, giving fast correct responses (accuracy data) is not more difficult than giving correct responses within a lenient time interval.

Speed and power are two basic concepts to be considered for cognitive tests. It is an old issue in the test literature (Kelley, 1927) whether speed and power are different cognitive performance traits, and if they are indeed, to what extent they each are measured in a given test. From a review of the literature, it is clear that the two concepts can be interpreted in different ways and consequently that they are measured in different ways. We will first briefly discuss the literature before deciding on an operational definition.

Early on in the history of psychological tests, it was assumed that there was no reason to distinguish between speed and quality in the domain of intelligence (Spearman, 1927). In empirical studies later on, speed was operationally defined in various ways. Examples include speeded tests of a rather high complexity (e.g., reasoning tests), tests with easy tasks (e.g., perceptual speed tasks), response time per item, and information processing time measures. From an analysis of correlations and from factor analysis results, it was concluded that speed is different from power (Carroll, 1993). For example, Davidson & Carroll (1945) showed that for tests with unlimited time, speed and accuracy measures load on different factors, while time-limit scores load on both. The time-limit scores were obtained through a procedure in which the respondents first were given a strict time limit but afterwards could work further on the test. Although respondents seem able to adapt successfully to speeding (Morrison, 1960), it was also found that a test in speeded conditions measures a somewhat different ability than in unspeeded conditions (Lord, 1956).

The issue of power versus speed seems to have lost importance more recently. Gulliksen (1950) gave the topic a prominent place in his book, and Anastasi (1976) still mentioned the issue. More recent psychometric approaches have considered speededness a confounding factor to be avoided (de Ayala, 2009). The analysis of intelligence tests in terms of speed, accuracy, and

persistence, and the corresponding Nufferno tests (Furneaux, 1956) are no longer receiving much interest, notwithstanding their fundamental aspects and relevance to the issue of power and speed. For one of the most recent studies based on this approach, see Frearson, Eysenck, and Barrett (1990).

In contrast to the trend just sketched, the topic of speed versus power has seen a revival in more recent research on the nature of intelligence. More specifically, the question is whether speed of information processing (SIP) contributes to what is called psychometric intelligence (as measured by IQ-tests). To measure speed of information processing, so-called elementary cognitive tests (ECTs) are used (e.g., Neubauer & Bucik, 1996). Although it is accepted that speed does play a role in the process of intelligence, views still differ on its nature (Danthiir, Wilhelm, Schulze, & Roberts, 2005; Stankov & Roberts, 1997). In a review of 50 years of research and 172 studies in that period, Sheppard and Vernon (2007) came to the conclusion that the correlation between speed (as measured through ECTs) and intelligence is in the range of 0.30 -- 0.40.Wilhelm and Schulze (2002) found that speeded intelligence tests correlate more highly with SIP than do non-speeded tests.

Other evidence for the difference between speed and power has been found when using the same test to determine accuracy and response time. There is clear empirical evidence for the latter two being independent. For example, van der Linden, (2009) mentioned a number of studies with very low or even negative correlations across persons between accuracy and response time.

In sum, there seems to be evidence from different perspectives to differentiate between power and speed. It makes therefore sense to ask how much power and speed is measured in a given test. In preparation of the approach to be followed, we will first discuss various ways to

measure and model speed and describe some important conceptual issues about the notions of speed and power.

Approaches for Speed

In comparison with accuracy, response times, especially for individual items, have been difficult to observe in a test situation, although this has been alleviated by advent of computerized testing. Response time data are also more difficult to model. Various interesting and promising models exist for response times, but in comparison to modeling accuracy, there is much less agreement on which models to use. For instance, there is a major difference between approaches that model response times either exclusively (Rasch, 1960; Scheiblechner, 1979; Maris, 1993), or concurrently but separately from responses and their accuracy (Gorin, 2005), or concurrently and jointly. Some approaches adapt a basic IRT model for responses to include response times (Roskam, 1997; Verhelst, Verstralen, & Jansen, 1997; Wang & Hanson, 2005), while others incorporate responses into models basically oriented towards latencies (Thissen, 1983). Recent joint models for responses and response times have utilized mixed models (van Breukelen, 2005), hierarchical structures for response time and accuracy (van der Linden, 2007) or hybrids between IRT and survival models (Ranger, 2009). There is also a great variety in the statistical distributions that have been used to model response times: log-normal (van der Linden, 2007), gamma (Maris, 1993), Weibull (Rouder, Sun, Speckman, Lu, & Zhou, 2003), and an approach based on the Box-Cox transformation that promises to approximate almost any distribution (Klein Entink, van der Linden, & Fox 2009).

There is an alternative for response times when one wants to measure speed. Speed can be measured by the sum score in a test with a limited time and very easy items, while power can be measured by the sum score in a test with unlimited time and difficult items (Gulliksen, 1950).

However, if speed is measured with easy items from one test and power with difficult items from another test, then the nature and the difficulty of items are confounding factors. Further sources of confounding may play a role. For example, because a test consists of a sequence of items and the respondent can decide when to move from one item to the next, various kinds of sequential effects may play a role in adapting to time limits. After a further discussion of the notions of speed and power, we will propose a method to avoid these sources of confounding.

The Notions of Speed and Power

Despite the methodology used to investigate speed and power these notions are ambiguous for two main reasons. First, power and speed cannot be directly observed and need to be derived from observables, and they are therefore latent variables. They are also relative notions. Power is the power to overcome something, speed is the time to cover a distance, and so a latent counterpart from the item side is required. Item response theory offers a solution for the power side of the problem, with ability as the latent person variable, and with difficulty as the item counterpart to be overcome with one's power. Speed, on the other hand, is the time taken to cover a certain distance, but there is not much clarity or agreement over the latter's nature. In his joint modeling approach for response time and accuracy, van der Linden (2009) proposed the notion of latent time intensity for the item counterpart, which can be interpreted as the distance to be overcome by one’s speed

Second, speed and power are ambiguous because they can easily vary for the same task and for the same person. They both refer to a range of possible levels from which a specific level is implemented while solving a cognitive task. Which level of speed and of power is used depends on the situation, the strategy, and the state of the respondent. For example, stress may have a positive effect on speed and a negative one on power, and motivation may have a positive

effect on both. Another important given is that a respondent may use a strategy trading-off speed against accuracy. The phenomenon has primarily attracted the attention of experimental psychology (Pachella, 1974), but it has been given some attention in the literature on tests as well (Heim & Batts, 1948; Dennis & Evans, 1996; Thurstone, 1937; van der Linden, 2009). All these factors and more can play a role in the precise level of power and of speed that is invested while taking a test. Distinguishing maximum performance, typical performance, and actual performance in the test under consideration, one can possibly come to a disambiguation, but it would require rather difficult experimental set-ups in order to make the former two notions work in an empirical study.

A Pragmatic Approach

A pragmatic way to deal with the conceptual problems from the previous section is to make two concessions. First, speed and power are understood as the level of power and speed as used in the test under consideration, which may be influenced by other factors such as those just discussed. Therefore, no absolute claims can be made unless a much larger study is set up in which these factors are manipulated in a systematic way. This is an inherent limitation in the study of speed and power. Second, because it is still far more common to rely on responses and models for these, and because response times require a choice for how to analyze them that is to a certain degree still arbitrary (e.g., with respect to their distribution), it seems attractive to use the same kind of data and the same kind of models for speed as for power, namely coded responses instead of response times as data, and a common IRT approach to model these data. The responses can be coded in different ways. Speed can be derived from whether or not a person responds within a stringent posterior time limit, and power can be derived from whether or not the response is correct given a lenient posterior time limit. The notion of a distance to be

overcome with one's speed can then be specified as the item difficulty of giving a response within a stringent time limit.

In order to deal with possible sources of confounding as discussed earlier, and to stay in line with the pragmatic approach just described, we will use *posterior time limits*. Starting with the empirically recorded response times within a sufficiently large limit for all or almost all respondents, one can consider various time cut-offs. Only responses preceding the cut-off are counted as successes (if correct), and slower responses are considered either as failures or as missing.

An approach based on posterior time limits starting from an unlimited time procedure has several advantages. One can choose whatever posterior limit that is required without doing a new experiment; the test taking attitude does not change with the size of the time limit; and the design is by definition a within-subject design, whereas a priori time limits would require multiple testing and is therefore difficult to implement with a within-subject design. The disadvantages of posterior time limits are that one cannot generalize to the situation where the corresponding time limits are imposed a priori, and that the variables derived from different posterior time limits, such as sum scores, are not independent.

Kinds of Data

Using the approach of posterior time limits and responses as data, one can recode the data in three ways for each item and for a range of posterior time limits:

Binary *time data:* Success means that the response is given within the time limit, failure means it is not;

Binary *time-accuracy data*: Success means that a response is given within the time limit and is correct, failure means the response is either not given within the time limit or it is

incorrect;

Incomplete binary *accuracy data*: Responses that are not given within the time limit are considered as missing, which makes the data incomplete, and the other responses are coded as correct or incorrect.

The time data with the most stringent time limit can function as a basis to derive an anchor measure for speed, while the accuracy data without time limit can function as a basis to derive an anchor measure for power. As an alternative speed anchor, we have also computed the average reciprocal response time over all items in the test.

The kind of test we will use is an inductive reasoning test with verbal analogies. Verbal analogies have attracted much attention in the study of intelligence (Spearman, 1927; Sternberg, 1977; Whitely, 1977a). They are the most commonly used type of analogy items in intelligence tests (Ullstadius, Carlstedt, & Gustafsson, 2008), and they are used also for the SAT and GRE. Verbal analogy tests measure g and to some extent also crystallized intelligence (Bejar, Chafin, & Embretson, 1991; Levine, 1950; Thurstone, 1938), the latter depending on how much word knowledge is required (Ullstadius, Carlstedt, & Gustafsson, 2008). In sum, verbal analogies may be considered a very popular type of intelligence test, and a measure of core aspects of intelligence.

Research Issues

The main research issue is *how much speed and power is measured by a test*, and more in particular by the verbal analogies test under consideration in this study. To answer this question, we will use a methodology based on posterior time limits. The answer will depend on how much speed and power are related. First, if they are highly related, this will show in a strong convergence between all latent variables as measured through the time data, the time-accuracy

data, and across the moving posterior time limits. In that case, the conclusion must be that both speed and power are always measured to a large extent. Second, if speed and power are not highly related, then we expect that the trait being measured depends on the kind of data and on posterior time limits. For example, the more stringent the time limit is, the more the time-accuracy data are expected to measure speed instead of power. It is an exploratory question whether the same would be true for the accuracy data.

A second research issue concerns the *reliability of the measurements obtained through the posterior time limits* for the various kinds of data. This is an important issue for two reasons. First, the reliability of a score must be taken into account for an interpretation of how highly it correlates with other scores. Second, the influence of the posterior time limits on reliability is in itself an issue.

Method

Instruments

We use data from the calibration studies for a computerized adaptive test developed by Hornke (Hornke & Rettig, 1993; Hornke, 1999; Hornke, 2001). The median proportion of correct responses over all items is 0.75 (Q1=0.62, Q3=0.83). The tests for the calibration study were administered in a computerized but non-adaptive format, with a very generous time allowance of 180 seconds per item. The complete data set contains responses (right or wrong) and response times in seconds for a total of 254 items, arranged into a complex design of 25 partially overlapping test forms of 24 items each. The average sample size per test form is about 500 persons. In this paper, we concentrate on the first test form which enjoys the largest sample size, 904 persons, but we have examined all samples and test forms to make sure that results are similar. The set of 254 items is constructed based on a typology of verbal analogy items derived

by Whitely (1977b) and found to correlate moderately high with standard intelligence tests (Whitely, 1976; Whitely, 1977a). Six of the eight Whitely types are represented among the 254 (the numbers in the subset of 24 are given within parentheses): opposites (4), functional relations (2), quantitative (3), class membership (7), conversion (5), and class naming (3). The difficulty of the types as represented among the 24 items is similar to the difficulties among the 254. The Cronbach alpha for the 24 items is .81.

Participants

All participants were at age between 17 and 27 as of the time of testing, with a median age of 20 (Q1=19, Q3=21). Almost all (98.6%) were male. With respect to schooling level and occupation, they represented a realistic sample of the male population in this age group.

Data Analysis and Statistics

Posterior time limits. To treat items in a statistically comparable way and avoid unrealistic time limits, we choose time limits corresponding to a given empirical quantile of the response times recorded for each item. In other words, the time limits (in seconds) for each item are different, but they are equivalent in the sense that the same proportion of all subjects (say, 80%) could make the limit in the original survey. We have opted for a fairly fine-grained analysis, fixing the time allowance for each item at seven percentiles: the 90%, 80%, … , 30% percentile of the RT observed.

At each such set of time limits, we produce *three binary data sets*: a time data set, a time-accuracy data set, and an accuracy data set. This means that in total three times seven binary data sets are generated: ti90 to ti30 for the time data, ta90 to ta30 for the time-accuracy data, and ac40 to ac40 for the accuracy data.

These series of data sets are complemented by the original response data produced under

conditions approaching unlimited time, i.e., the power condition. These original response data, not manipulated by the introduction of time limits, can be denoted both as ac100 or ta100, but we have chosen to represent them as power. The variable labeled as 1/RT is the average reciprocal RT over all items for a subject. Remember that this is one of the anchors that we use for speed: the other one is the time data under the most stringent posterior time limits in the design, which happens to be ti30.

Item response modeling. To address the research questions, we need estimates of the latent variables underlying the various data sets, and of the reliability of the individual items and the tests as a whole. Thanks to the decision to treat speed as a latent variable with dichotomous indicators, we can apply the same methodology to all data sets. We have used the two-parameter logistic model (2PL, see de Ayala, 2009). Under this model, the probability of a correct response by subject *j* to item *i* is modeled as:

$$P_{ij} = \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]} \qquad (1)$$

with θ_j as the latent ability score, b_i as the difficulty parameter, and a_i as the discrimination (slope) parameter for item *i*. The item parameters in the 2PL model (a_i and b_i) were estimated by maximum marginal likelihood using the program ICL (Hanson, 2002). Based on these estimates, the latent ability scores for each subject (θ_j) were estimated by maximum likelihood.

The *reliability of the individual items* can be derived from the estimates of the discrimination parameters, a_i, which can be regarded as the IRT analogue of item-test correlation.

In item response modeling, the *reliability of the total test* can be measured through the test information function. For the 2PL model the formula is:

$$\hat{T}_j = \sum_{i=1}^{I} \hat{a}_i^2 \hat{P}_{ij}(1 - \hat{P}_{ij}), \quad (2)$$

where $\hat{T}_j$ is the estimated test information function corresponding to the ability level of subject j, $\hat{a}_i$ is the estimated item discrimination for item i, and $\hat{P}_{ij}$ is the estimated probability of success of person j for item i. Because of the $\hat{P}_{ij}$ component, it is clear that the test information value is a function of the ability (see Figure 4 for an illustration). $\hat{T}_j$ reflects the inverse of the estimated error variance at the ability level in question. The larger the information at a given location on the latent continuum, the better measurement can be achieved there. de Ayala (2009) describes the test information function as the instrument's estimation capacity.

Correlations and PCA. To address the vital question of the nature of latent variable measured in each case, we examine the correlation of the ability estimates. The correlations obtained from the three kinds of data (time, time-accuracy, and accuracy), at various posterior time limits, are compared to the anchors: the score under power conditions (time limit of 180 seconds) as an anchor for power; and the score under the most stringent time limit, along with the average reciprocal response time, as two anchors for speed. In order to have an overall view on the relationships, we will conduct a principal components analysis, expecting power and speed as the major two components.

A speededness coefficient. To interpret the nature of the tests, we also make use of a coefficient τ developed by Cronbach and Warrington (1951) to measure the degree of speededness in a test. Assuming that the test has been given in a power and a speed condition, and that parallel versions *A* and *B* are available, the coefficient is computed as:

$$\tau = 1 - \frac{r_{A_p B_s} r_{A_s B_p}}{r_{A_p B_p} r_{A_s B_s}} \qquad (3)$$

where for example, $r_{A_p B_s}$ is the correlation between the sum scores when test form *A* is given in a power condition (subscript p) and test form *B* is given in a speed condition (subscript s), and similar notation applies for the other three correlations. The index lies between 0 and 1, and shows what proportion of the reliable variance in the score obtained with a given time limit reflects the same factor as the test does when given under unspeeded conditions. We have applied this measure on all pairs of data sets formed by responses (right or wrong) under practically unlimited time as the power condition, and each of the manipulated data sets in turn as the speed condition; as parallel versions we use the sum scores of odd- and even-numbered items.

Analysis of the item parameter estimates. In order to test the trends in the item discriminations as functions of the posterior time limit, we have made use of linear mixed model analysis (Gelman & Hill, 2007), based on the lme4 software package (Bates & Maechler, 2009). In a first step, only the linear component was included, and in a second step the quadratic component was added. The intercept was random in all cases, and when a component was included (linear, quadratic), the corresponding random effect was also added. We have done a similar analysis on item difficulties as well – in fact, these results will be presented first, so that results for item discriminations may be examined in the light of results for item difficulties. For an interpretation of the estimated effects, it is important that the quantiles (from Q30 to Q90) are transformed into proportions (from .30 to .90).

Results

Posterior Time Limits and the Nature of the Latent Variables

A straightforward way to find out about the nature of the latent variable measured by

each of the three kinds of data at each degree of speededness, is to explore the correlations between the subject scores they produce. The scores are the ML estimates under the 2PL model. Figure 1 shows the correlations between the three anchor variables and the various ability estimates, using small squares in shades of grey corresponding to the correlation value.

Figure 1 about here

From Figure 1, it is easy to observe that *speed* is virtually uncorrelated with *power.* Speed is highly correlated with all abilities estimated from data of the *time* type. Power is strongly correlated with all abilities estimated from *accuracy* data. The abilities estimated from *time-accuracy* data move from correlating with power to correlating with speed when the time limits become more demanding.

Figure 2 about here

Another useful representation of the correlational structure is provided by Figure 2, based on the first two components of a PCA starting from the correlations between the various ability estimates and the anchor variables. The first five PCA eigenvalues are 11.82, 7.62, 0.87, 0.67, and 0.45, supporting the two-dimensional structure. Figure 2 shows that power and speed are virtually uncorrelated; their vectors form more or less a right angle. The abilities estimated from accuracy data cluster around the power anchor. In a similar fashion, the abilities estimated from time data cluster around the speed anchors. The time-accuracy data seem to measure a composite of power and speed that varies smoothly with the posterior time limit, and can cover practically the whole range between pure power and pure speed.

Figure 3 about here

The speed component introduced into the latent variable by manipulating the posterior time limits can also be demonstrated with traditional measures of speededness such as the

coefficient due to Cronbach and Warrington (1951). Figure 3 shows results for the three kinds of data as function of the time limit. The speed contained in the accuracy data is very low and not really influenced by the time limit; the time data are always measuring speed; and the speed character of the time-accuracy data increases smoothly as the time limit becomes more and more stringent.

Posterior Time Limits and Item Difficulty

The effect of posterior time limits on the difficulty of the items can be studied through the change in the difficulty parameters of the 2PL model. The relationship between time limits and difficulty seems to depend on the kind of data. Time limitation has a strong positive influence on the difficulty for the time data, a moderately positive influence for the time-accuracy data, and no influence for the accuracy data. The results of the statistical tests are summarized in Table 1 and can be described as follows.

Table 1 about here

As can be expected, the difficulty of the items in the time data set increases dramatically as posterior time limits become more stringent ($t(899) = -73.82$, $p < .001$); the quadratic trend is also significant when added ($t(895) = -7.91$, $p < .001$), which means that, given the signs, the trend is slightly concave. For the time-accuracy data, the difficulty also increases with a shrinking posterior time limit ($t(899) = -16.40$, $p < 0.001$), and again the quadratic trend is significant when added ($t(895) = -3.43$, $p < 0.001$) and slightly concave, but less than for the time data. For the accuracy data, neither the linear trend ($t(899) = 1.24$, *n.s.*), nor the quadratic trend ($t(895) = 1.31$, *n.s.*) are significant.

Posterior Time Limits and Reliability

To understand what happens to the reliability of the test as the posterior time limits

become increasingly demanding, we concentrate in turn on item discrimination, on the test information function, and, briefly, on Cronbach's alpha coefficient. Discrimination can be approached through the slope parameters in the 2PL model. Similar to item difficulties, the relationship between time limits and discrimination depends on the kind of data. Compared to the dramatic effects on the difficulties for the time data and the time-accuracy data, the effects seem much weaker. Time limitation has a slight curvilinear influence on the discrimination for the time data, a slight positive curvilinear influence for the time-accuracy data, and a slight positive linear influence for the accuracy data. In all three cases the discrimination does not decrease from a lenient to a stringent time limit. The results of the statistical tests are presented in Table 1 and can be described as follows.

For time data, the linear trend is not significant, but adding the quadratic component does lead to a significant result ($t(895)=7.13$, $p<0.001$). The quadratic trend is convex. Discrimination tends to be higher towards the higher and lower posterior time limits. For time-accuracy data, the linear trend is significant and slightly increasing when the time limit becomes shorter ($t(899)= -2.04$, $p<0.05$), and also the quadratic component is significant when added ($t(895)=7.42$, $p<0.001$). The trend is again convex, possibly because of the convex relationship for the speed component. There is no reason to expect that the change in difficulty is at the basis of the convex trend, because the highest discrimination is found when the difficulty is very low. A possible substantive interpretation is that both fast and slow responding are rather stable response styles, while in the medium range less stable factors are at work. A possible artefactual interpretation is that the actual difference between the quantiles is smaller toward the middle quantile range, so that response time differences are somewhat less reliable in that range.

Finally, for the accuracy data, the linear trend is significant and slightly increasing when

the time limit becomes shorter ($t(899)=-3.56$, $p<0.001$), but when the quadratic trend is added, it is not significant ($t(895)=-0.285$, *n.s.*). Two possible explanations for the increase in discrimination for the time-accuracy and accuracy data when the time limit becomes sharper are that (i) removing the slowest responses may have eliminated many lucky guesses (which typically are made later on, as other answering strategies fail), thus improving the measurement precision, and (ii) the overall increase in item difficulty may have improved our ability to estimate the slopes more precisely, as the data tend to cover better the central part of the item trace lines.

Figure 4 shows the effects of time limits on the test information function. In line with the results for the discriminations, it can be concluded from the height of the curves for all three kinds of data that the reliability is not lower for stringent than for lenient time limits. The results obtained for the test information function are nicely in line with those obtained for the difficulties and the discriminations. This follows from the test information formula, which is a summation over the squared discrimination estimates multiplied with the product of the estimated probability of success and the estimated probability of failure. The latter product reaches its maximum when these two probabilities are equal, and hence where the respondent's ability matches the difficulty of the item. When the items become more difficult, the maximum of the product and therefore also the test information curve move toward higher abilities. When the item discrimination increases, the whole curve raises.

Figure 4 about here

For time data (left) the curves gradually and fairly uniformly shift to the right as the time limit becomes increasingly stringent, in line with the increasing difficulty. Their heights exhibit a marked quadratic effect, in line with the quadratic effects for discrimination. Similar effects can

be observed for the time-accuracy data (middle), but the effects are weaker, in line with the results for difficulty and discrimination. For accuracy data (right), the curves are all on top of each other because manipulating time limits does not change difficulty for this kind of data. The increase in discrimination explains why the curves raise.

We have computed also Cronbach's coefficient alpha for all recoded data sets. Its computation is not equally convenient across the three kinds of data because of the extreme sum scores found in some of the time data sets and the missing data in the accuracy data sets. Computed with these limitations in mind (including pairwise deletion in the case of accuracy data), the internal consistency of the data is quite high, also for the most stringent posterior time limit, and in comparison with the original (not recoded) data: .86 for the ti30 data, .84 for the ta30 data, and .84 for the ac30 data.

Clearly, decreasing the posterior time limit does not have a negative effect on reliability. This is an interesting result in itself, and it also implies that changes in the correlations of the estimated abilities with relevance for the power and speed issue, cannot be interpreted as following from a decreased reliability.

Discussion and Conclusion

The present study is relevant in two respects. First of all, a certain methodology is followed for the analysis of the data, in order to obtain a rather fine-grained view on the issue of power and speed. The methodology can be used for other tests as well. Second, regarding the substantive research issues outlined in the Introduction, we found that speed and power seem to be more or less uncorrelated in a verbal analogy test. Furthermore, all scores obtained from time data are strongly correlated, and are relevant measures for speed, but perhaps slightly less so when the limit is more lenient. For the time-accuracy data, the scores move gradually from

power to speed when the posterior time limit is reduced, as most easily seen on Figure 2. For the accuracy data, the scores remain primarily based on power. It is clear that it depends on the kind of data and for the time-accuracy data also on the posterior time limits, to what extent the data measure power and speed. As announced, the analysis for the selected set of 24 items was replicated for other such item sets with verbal analogies, and similar results were obtained.

Regarding item discriminations and the more fundamental issue of reliability, we have the interesting result that reliability can actually increase as the time limit becomes more stringent. For the time data, this is clearly visible from the test information function (Figure 4). For the measurement of power, the best explanation, in our opinion, is that decreasing the time limit may eliminate some lucky guesses that typically occur after a longer search for the right answer. An alternative explanation would be that having items that are not too easy might improve chances for better estimates of the IRT slope coefficients.

The fact that the time-accuracy scores diverge under decreasing posterior time limits means that, in practical applications, it is not possible to apply time limits as a sort of item generating device, automatically producing individual items of arbitrary difficulty. Added to a test of unspeeded items, such an item will have a lower discrimination than it would have in a test of speeded items. However, it may be possible to make a whole test more difficult and have better reliability and a number of further advantages, such as better coverage of the latent continuum and better estimation of discrimination parameters, which are both highly desirable in the context of adaptive testing. The price to pay is in terms of validity, and it will depend on the occasion whether the inclusion of a speed component is acceptable – “a societal judgment concerning the value of high intelligence combined with quickness of response or problem solving”, as Carroll (1993) so succinctly put it.

We believe that the conceptual framework and study of power and speed can still be further enriched with the important third concept of persistence in Furneaux's (1961) approach. It is a way to include so-called "orectic" aspects (stress, motivation) of the test situation and the respondents' personality as determinants of intelligence test scores.

Considering different aspects of the data from the same test offers the possibility of a broader qualitative range of scoring, in our case a range from speed to power, and therefore also the possibility of incremental validity. It is of course an empirical issue whether actually the validity can be increased through the inclusion of a broader range of scores. This possibility is shared with other approaches which focus simultaneously on accuracy and response times, including modeling approaches such as the one described by van der Linden (2009). It may be of interest to broaden the qualitative range of scoring even further. Persistence scoring such as in Furneaux' approach seems an interesting extension. Taking multiple aspects into account is not only beneficial with an eye on incremental validity, but also to find out to what extent various aspects play a role in test scores, so that correlations of a test with other variables can be interpreted in a more profound way.

References

Anastasi, A. (1976). *Psychological Testing.* New York: Macmillan.

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory.* New York: Guilford Press.

Bates, D., & Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes.* R package version 0.999375-32, retrieved from http://CRAN.R-project.org/package=lme4

Bejar, I., Chafin, R., & Embretson, S. (1991). *Cognitive and Psychometric Analysis of Analogical Problem Solving.* New York: Springer.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytical Studies.* New York: Cambridge University Press.

Cattell, R. (1971). *Abilities: Their Structure, Growth, and Action.* Boston: Houghton-Mifflin.

Cronbach, L. J., & Warrington, W. G. (1951). Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika, 16* , 167-188.

Danthiir, V., Wilhelm, O., Schulze, R., & Roberts, R. D. (2005). Factor structure and validity of paper-and-pencil measures of mental speed: Evidence for a higher-order model? *Intelligence*, 33, 491-514.

Davidson, W. M. (1945). Speed and level components in time-limit scores: A factor analysis. *Educational and Psychological Measurement, 4*, 411-427.

Dennis, I., & Evans, J. (1996). The speed-error trade-off problem in psychometric testing. *British Journal of Psychology, 87*, 105-129.

Frearson, W., Eysenck, H.J., & Barrett, P.T. (1990). The Furneaux model of human problem solving: Its relationship to reaction time and intelligence. *Personality and Individual Differences*, *11*, 239-257.

Furneaux, W.D. (1956). *Nufferno Manual of Speed Tests*, and *Nufferno Manual of Level Tests*. London: National Foundation of Edicational Research.

Furneaux, W.D. (1961). Intellectual abilities and problem solving behavior. In H.J. Eysenck (Ed.), *Handbook of Abnormal Psychology* (pp. 167-192). New York: Basic Books

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* New York: Cambridge University Press.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension items: The feasibility of verbal item generation. *Journal of Educational Measurement, 42*, 351–373.

Gulliksen, H. (1950). *Theory of Mental Tests.* New York: Wiley.

Hanson, B. A. (2002). *ICL: IRT Command Language.* Retrieved from www.b-a-h.com

Heim, A. W., & Batts, V. (1948). Upward and downward selection in intelligence testing. *British Journal of Psychology, 39*, 22-29.

Hornke, L. (1999). Benefits from computerized adaptive testing as seen in simulation studies. *European Journal of Psychological Assessment, 15*, 91-98.

Hornke, L. (2001). Item generation models for higher order cognitive functions. In S. Irvine, & P. Kyllonen (Eds.), *Item Generation* (pp. 159-178). Hillsdale, N.J.: Lawrence Erlbaum.

Hornke, L., & Rettig, K. (1993). Evaluation und Revision einer Itembank von Analogieaufgaben [Evaluation and revision of an item bank of verbal analogy items]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14*, 113-128.

Kelley, T. (1927). *Interpretation of Educational Measurements.* Yonkers, N.Y.: World Book Co.

Klein Entink, R. H., van der Linden, W. J., & Fox, J. P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology, 62*, 621-640.

Levine, A. (1950). Construction and use of verbal analogy items. *Journal of Applied Psychology,*

34, 105-107.

Lord, F. (1956). A study of speed factors in tests and academic grades. *Psychometrika, 21*, 31-50.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika, 58*, 445-469.

Morrison, E. J. (1960). On test variance and the dimensions of the measurement situation. *Educational and Psychological Measurement, 20*, 231-250.

Neubauer, A. C., & Bucik, V. (1996). The mental speed-IQ relationship: unitary or modular? *Intelligence, 26,* 23-48.

Pachella, R. G. (1974). The interpretation of reaction time in information processing research. In B. H. Kantowitz (Ed.), *Human Information Processing: Tutorials in Performance and Cognition* (pp. 41-82). Hillsdale, N.J.: Erlbaum.

Ranger, J. M. (2009). *Der Nutzen von Reaktionszeiten bei Psychologischen Tests im Rahmen von Item Response Modellen [Using Response Times in Psychological Tests Based onItem Response Models].* PhD dissertation, University of Giessen.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: University of Chicago Press.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden, & R. Hambleton. (Eds.), *Handbook of Modern Item Response Theory* (pp. 187-208). New York: Springer.

Rouder, J., Sun, D., Speckman, P., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika, 68*, 589-606.

Scheiblechner, H. (1979). Specifically objective stochastic latency mechanisms. *Journal of Mathematical Psychology, 19*, 18-38.

Sheppard, L. D., & Vernon, P. A. (2007). Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences, 44*, 247-259.

Spearman, C. (1927). *The Abilities of Man.* New York: Macmillan.

Stankov, L., & Roberts, R. (1997). Mental speed is not the basic process of intelligence. *Personality and Individual Differences, 22*, 69-84.

Sternberg, R. (1977). *Intelligence, Information Processing, and Analogical Reasoning: The Componential Analysis of Human Abilities.* Hillsdale, N.J.: Erlbaum.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J.Weiss (Ed.), *New Horizons in Testing* (pp. 179–203). New York: Academic Press.

Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika, 2*, 249–254.

Thurstone, L. L. (1938). *Primary Mental Abilities.* Chicago: University of Chicago Press.

Ullstadius, E., Carlstedt, B., & Gustafsson, J.-E. (2008). The multidimensionality of verbal analogy items. *International Journal of Testing, 8*, 166-179.

van Breukelen, G. J. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70*, 359-376.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*, 247-272.

Verhelst, N., Verstralen, H., & Jansen, M. (1997). Models for time-limit tests. In W. van der Linden, & R. Hambleron, *Handbook of Modern Item Response Theory* (pp. 169–185). New York: Springer.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that

incorporates response time. *Applied Psychological Measurement, 29*, 323–339.

Whitely, S. E. (1976). Solving verbal analogies: Some cognitive components of intelligence test items. *Journal of Educational Psychology, 68*, 234-242.

Whitely, S. (1977a). Information processing on intelligence test items: Some response components. *Applied Psychological Measurement, 1*, 465-476.

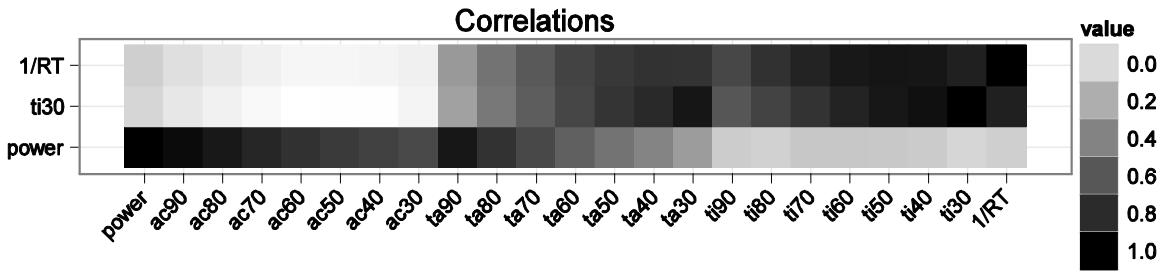
Whitely, S. (1977b). Relationships in analogy items: a semantic component of a psychometric task. *Educational and Psychological Measurement, 37*, 725-739.

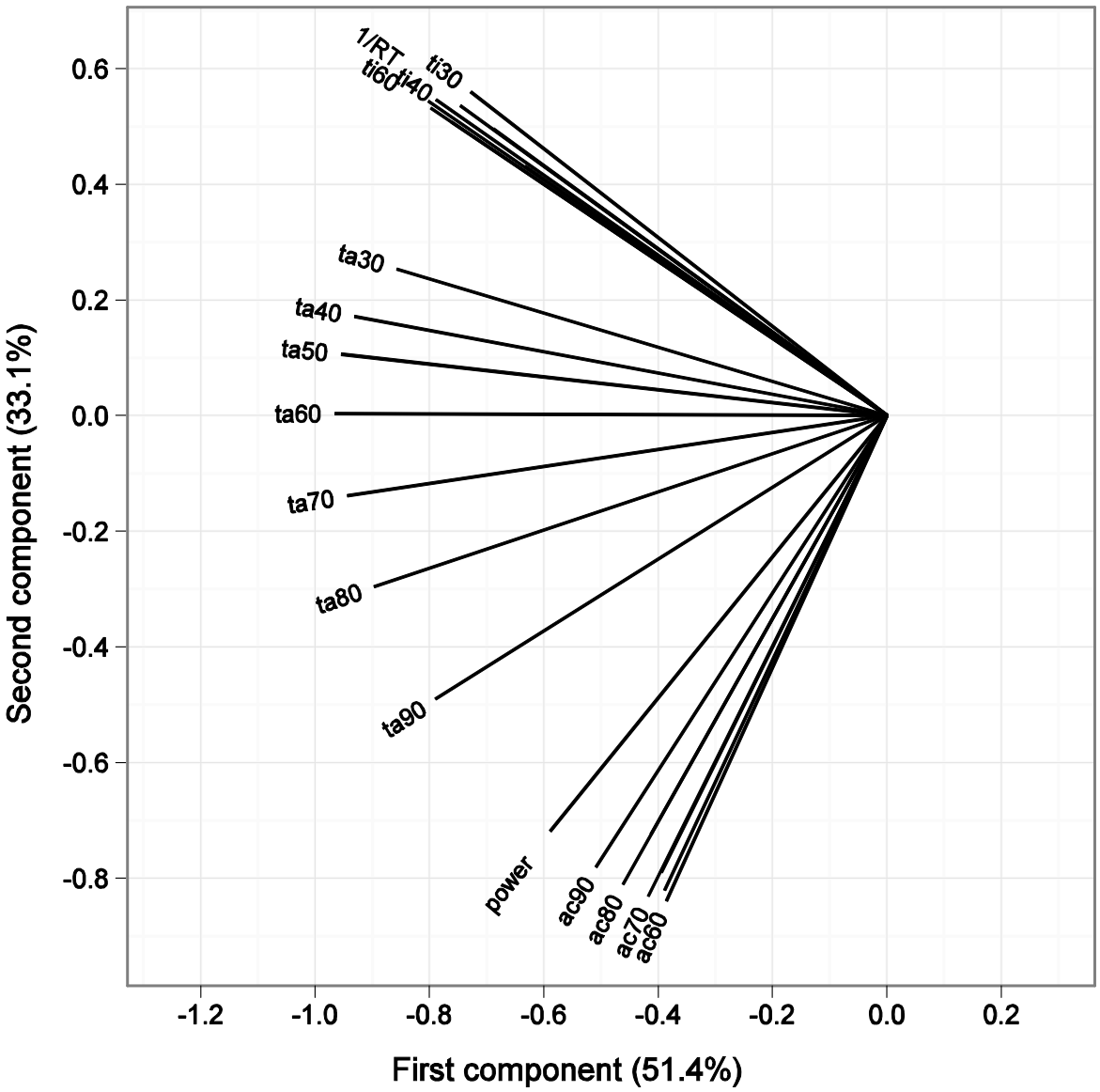
Wilhelm, O., & Schulze, R. (2002). The relation of speeded and unspeeded reasoning with mental speed. *Intelligence, 30*, 537-554.

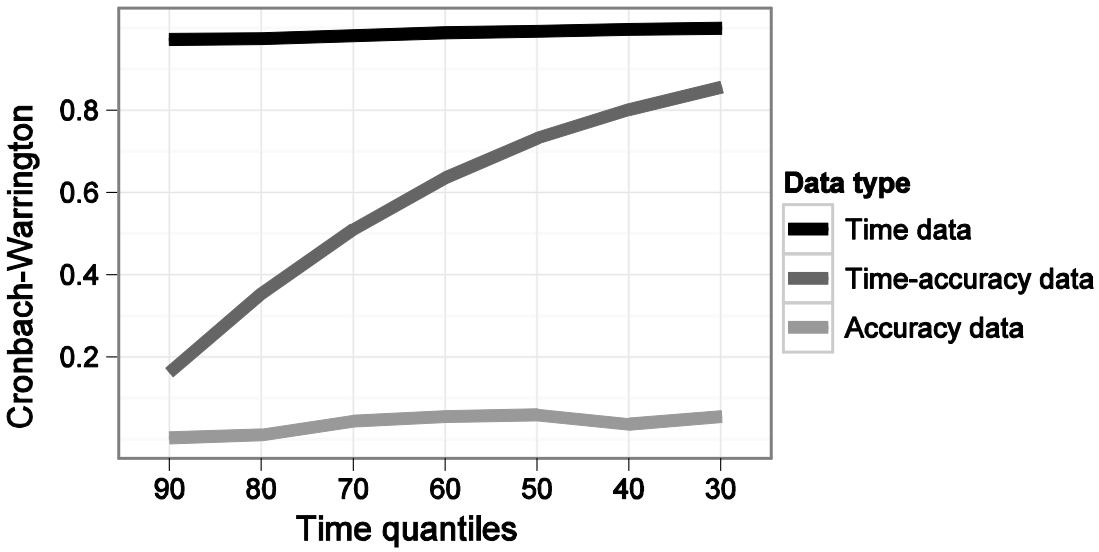
Table 1: Estimated linear and quadratic trends[a] in the item difficulty and discrimination as a function of posterior time limits for the three kinds of data[b].

	Difficulty		Discrimination	
Kind of data	Linear	Quadratic	Linear	Quadratic
Time data	-5.129***	-2.378***	0.038	1.254***
Time-accuracy data	-3.840***	-0.940***	-0.229*	1.291***
Accuracy data	0.182	0.689	-0.332***	-0.110

***p<.001
*p<.05
Note a. The linear effects are from the linear model, while the quadratic effects are from the quadratic model.
Note b. The time limits are transformed to proportions of the full time.


Correlations
1/RT
ti30
power
power
ac90
ac80
ac70
ac60
ac50
ac40
ac30
ta90
ta80
ta70
ta60
ta50
ta40
ta30
ti90
ti80
ti70
ti60
ti50
ti40
ti30
1/RT
value
0.0
0.2
0.4
0.6
0.8
1.0


1/RT
ti60
ti40
ti30
ta30
ta40
ta50
ta60
ta70
ta80
ta90
power
ac90
ac80
ac70
ac60
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.2
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
Second component (33.1%)
First component (51.4%)


Cronbach-Warrington
0.8
0.6
0.4
0.2
90
80
70
60
50
40
30
Time quantiles
Data type
Time data
Time-accuracy data
Accuracy data

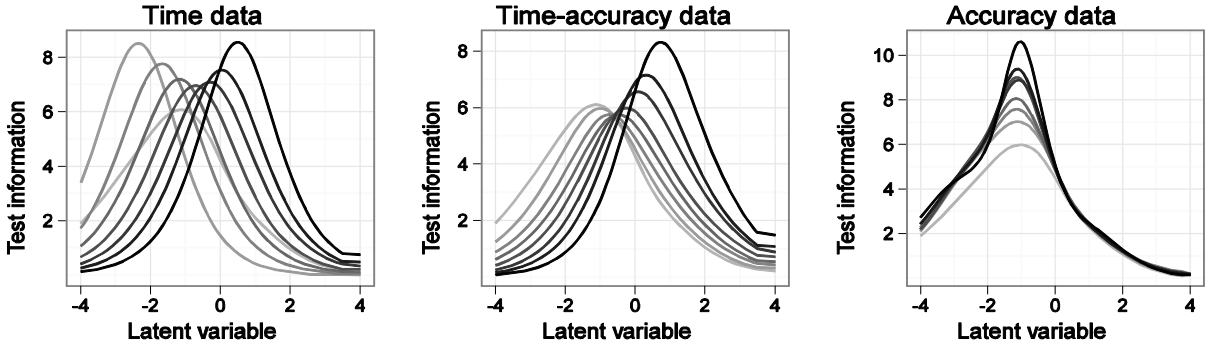

Time data
Test information
Latent variable
Time-accuracy data
Test information
Latent variable
Accuracy data
Test information
Latent variable

Figure 1. Image plot of the correlations of the ML ability estimates from time data (ti), time-accuracy data (ta), and accuracy data (ac) at various posterior time limits, with the anchor measures for speed and power. The measures range from the power anchor on the left (power) to the two speed anchors (ti30, 1/IRT) on the right. Darker grey denotes higher correlations, lighter grey denotes lower correlations. The bottom row represents the correlations with *power.* Going from left to right on that row, the grey becomes lighter, meaning that the correlations decrease from the accuracy data through the time-accuracy data to the time data, and from the more lenient time limits to the more stringent ones. The upper and middle rows represent the correlations with *speed*. Going from right to left on these two rows, the grey becomes lighter, meaning that the correlations decrease from the time data through the time-accuracy data to the accuracy data, and going from stringent time limits to the more lenient time limits.

Figure 2. PCA loadings based on correlations between ML ability estimates computed from responses under unlimited time (power); from time data (ti), time-accuracy data (ta), and accuracy data (ac) at various posterior time limits; and subject's average reciprocal response time (1/RT). Some of the labels have been removed to reduce visual clutter. The two kinds of anchors: power for power and ti30 and 1/IRT for speed, define almost orthogonal vectors. The power anchor clusters together with the accuracy vectors, and going upwards in clockwise direction from that cluster, one encounters the time-accuracy vectors in an order based on decreasing time limits, ending in a cluster with the time anchors and the time vectors.

Figure 3. Coefficient of speededness (Cronbach & Warrington, 1951) for time data, time-accuracy data, and accuracy data, as a function of the posterior time limits.

Based on the Cronbach-Warrington coefficient, the time data (upper curve) are uniformly measuring speed, and the accuracy data (bottom curve) are uniformly not measuring speed, while for the time-accuracy data (middle curve) it depends on how stringent the time limit is: the more stringent, the larger the extent to which the data measure speed.

Figure 4. Test information functions in the 2PL model for time data, time-accuracy data, and accuracy data, as a function of the posterior time limits; time limits are indicated by the shade of the curves, with black denoting the most stringent limits. The figure shows two movements of the curves when they become darker which corresponds to a sharper time limit. First, moving to the right means that the information present in the data and therefore also the reliability, focuses more on higher regions of the ability. Second, moving upwards and downwards means that the ability is estimated in a more reliable or less reliable way, respectively. For the *time data*, the reliability moves to higher ability regions and goes down and then up again. For the *time-accuracy data*, the reliability moves to the higher ability regions and increases. For the *accuracy data*, the reliability keeps focusing on the same ability region and increases.