Psychology Science, Volume 46, 2004 (4), p. 433 - 450

An algorithm for the separation of skill and working style[1]

JOACHIM HÄUSLER[2,3]

Abstract

This paper is aimed at the problem, that the causal origin of an ability test result can no longer be located, if the existence of a speed-accuracy trade-off is accepted. In the light of this problem, a method for the separation of skill and working style is discussed. Based on this, a speed & power test model for multiple choice situations – using aspects of the inspection time paradigm and signal detection theory – is described and practical suggestions for implementing it into psychological tests are given.

The postulated model is consecutively applied on a derivative of the vocabulary test LEWITE (Wagner-Menghin, 2002) and used for the prediction of test pre-knowledge in the sense of coached faking. The results show, that test pre-knowledge leads to better results, but also that it is possible to detect pre-knowledge by monitoring the working style of a person, operationalised with the model presented in this paper.

Key words: multiple choice, inspection time, signal detection theory, speed and power testing, decision quality, reflexivity, working style, test pre-knowledge

---

[1] The paper is based upon a master thesis, supervised by Prof. Klaus D. Kubinger

[2] Department of Psychology University of Vienna, Division for Assessment and Applied Psychometrics

[3] Dr. G. Schuhfried GmbH

Introduction

It appears rather easy to judge the achievement of a person according to whether they give a correct or wrong answer. But as soon as one realises that a speed-accuracy trade-off exists (Mazur & Hastie, 1978; Newell & Rosenbloom, 1981; Nährer, 1989; Kliegl, Mayr & Krampe, 1994; Roskam, 1997), the following questions might come to mind:

Was an answer correct, only because the subject took more time to decide carefully in comparison to others?

Was an answer wrong, only because the subject took less time in comparison to others and therefore decided impetuously?

Has a subject taken a lot of time for an item because he/she was overtaxed by the task or because he/she has a reflexive working style?

Has a subject taken little time for an item because he/she was unchallenged by the task or because he/she has an impulsive working style?

This indicates that causal attribution of test results is (in analogy to the competence-performance problem) not well defined. But in actual fact, the process of the assessment of a person's abilities requires this causal attribution process and therefore results in problems.

Because of a certain latent trait, the subject achieved a certain score.

This central link between a latent skill and a test score is not possible when one takes a speed-accuracy trade-off into consideration.

A generalisation of the Vickers Model on multiple choice situations

The Vickers Model (1970) describes the behaviour of a subject, trying to decide on the equality-inequality of two stimuli. In this situation, an algorithm of stepwise information search is postulated. In each step, the person inspects the stimulus material and accumulates a certain quantity of information.

The process of information accumulation can be described as a linear model if an appropriately small step size is selected as in figure 1.

$$L = (N * IT) + t_0 \tag{1}$$

Where N is the number of inspections necessary, IT is the inspection time for one single inspection and t_0 is the time for processes not relevant for the decision (eg. motor time).

This way of thinking can be easily applied to more complex item material. A condition necessary for this to take place is that a multiple choice answer format exists. Multiple choice decisions can be split up into single inspections, in which the stimulus is compared with one of the answer alternatives. The discontinuous process that Vickers formulated is in this case even more reasonable because the single inspections of the same stimulus are no longer hypothetical constructs but describe a typical solution strategy.

In a generalized Vickers model for multiple choice items (figure 2), information is accumulated in each inspection of a stimulus/answer-alternative pair. A decision is possible if the amount of information for the momentarily preferred answer exceeds the sum of information for all other answer alternatives plus a subject constant. This constant describes the individual caution a subject exhibits when forced to decide.

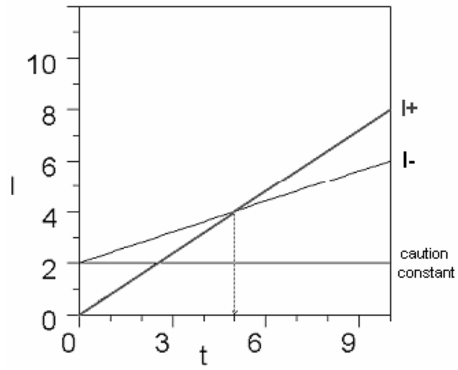


Figure 1:
Graph of the Vickers decision model. The accumulated information for (I+) and against (I-) a certain decision is drawn as a function of the total answer time for an item. A person comes to a decision, when the I+ > I- + caution constant

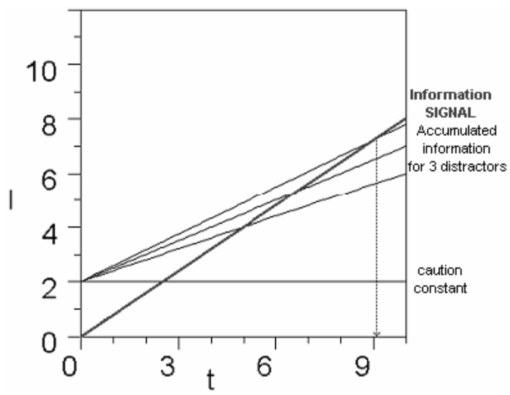


Figure 2:
Graph of the generalisation of the Vickers decision model. The accumulated information for and against a certain decision is drawn as a function of the item total answer time. A person comes to a decision, when the information (signal) > Σ (Information Distractor(k)) + caution-constant

A discontinuous decision process in multiple choice situations

By using the signal detection theory, the decision process for multiple choice items can be decrypted in more detail and described with a test-theoretical model.

If every single comparison between stimulus and an answer option is regarded as a signal in the context of the signal detection theory, there are 4 different implicit outcomes. Those and their probabilities are pointed out in the tables 1 and 2.

Table 1:
Possible outcomes of a signal-detection situation

	Equality assumed	Inequality assumed
Equality of stimuli	Correct Answer	Error Type II (β Error) => Return to decision process
Inequality of stimuli	Error Type I (α Error) => wrong answer	Return to the decision process

Table 2:
Probabilities for the outcomes in a signal-detection situation

	Equality assumed	Inequality assumed
Equality of stimuli	$1-\beta$	β
Inequality of stimuli	α	$1-\alpha$

Based on this, the decision process can be schematized like in figure 3.

Unlike in Vickers model, an item is not only a single comparison of two stimuli but a composite task consisting of several stimuli (one correct and several wrong answer alternatives) and higher mental tasks (like transformations).

Assuming that there are no differential solution strategies, the probability of inspecting one of the k answer alternatives is $\frac{1}{k}$.

In each of these inspections, the signal detection theory applies where a decision towards equality ends the item and a decision towards inequality continues the process with the inspection of another answer alternative.

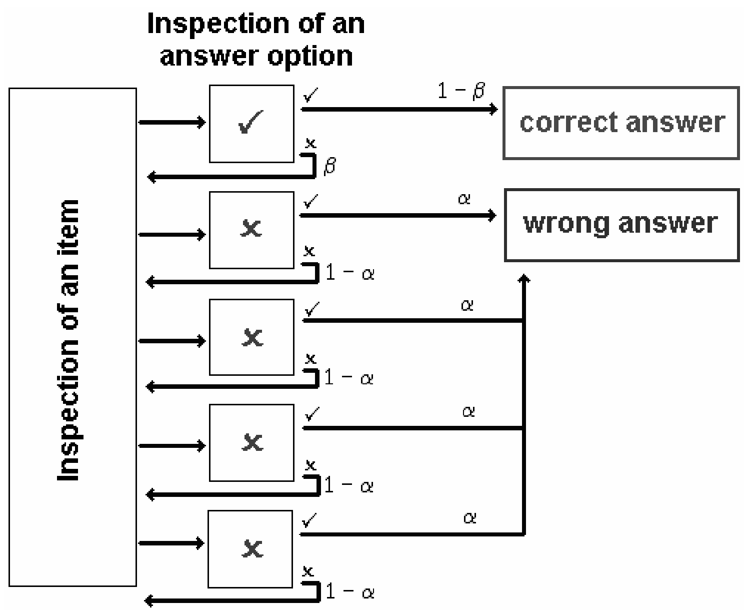


Figure 3:
Schematic plot of the decision process for a multiple choice item with 5 answer alternatives with one correct answer among them

The probabilities for the 3 possible manifest outcomes can be derived from this diagram as it is done in table 3.

Table 3:
The 3 possible outcomes of a multiple choice situation

Correct answer	P (+)	$\frac{1}{k}(1-\beta)$
Wrong answer	P (-)	$\frac{k-1}{k}\alpha$
No answer => return to inspection process	P (~)	$\frac{k-1}{k}(1-\alpha)+\frac{1}{k}\beta$

By building the odds ratio (s), which describes the relation of correct and wrong answers, the first part of the equation can be set up. Note, that the sum of P(+) and P(-) does not equal 1 since there is a third possible outcome which results in a return to the inspection process. Nevertheless, the ratio between P(+) and P(-) is still the same as the manifest odds-ratio.

$$s = \frac{P(+)}{P(-)}$$

$$s = \frac{\frac{1}{k}(1-\beta)}{\frac{k-1}{k}\alpha} = \frac{1-\beta}{\alpha * (k-1)}$$

Thus the relation between α and β can be expressed by the odds ratio.

$$\alpha = \frac{1-\beta}{s * (k-1)} \tag{2}$$

In order to completely resolve the equation system, additional information needs to be inserted. This is provided by the number of inspections that a subject needs to come to a final decision.

Based on the probability that no decision can be made – and therefore the item inspection must be continued, the following Maximum Likelihood solution for the number of necessary inspections (ct) can be established, where t is the mean duration of inspections.

$$ct = \frac{1}{P(+) + P(-)} = \frac{1}{1 - P(\sim)} = \frac{1}{\frac{1}{k}(1-\beta) + \frac{k-1}{k}\alpha}$$

If one substitutes the relation between α and β.

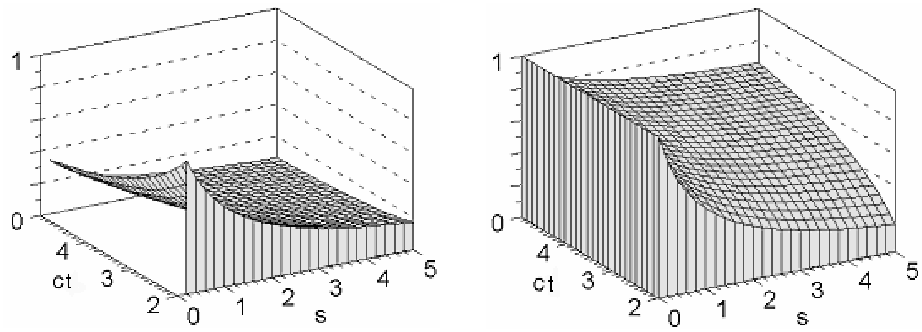
$$ct = \frac{1}{\frac{1}{k}(1-\beta) + \frac{1}{k}(\frac{1-\beta}{s})} = \frac{1}{\frac{1}{k}(1-\beta + \frac{1-\beta}{s})}$$

However, this still does not lead to a direct relation between the error probabilities and the time (t) a subject needs to solve an item. All one can conclude is the number of inspections, needed in order to make a decision – not necessarily the correct one though. The relation between the number of inspections and the time needed for them still needs to be expressed by a constant c, which for the moment we will approximate as constant for all subjects, so that the number of inspections = ct.

If one transforms the term above and resolves towards the error probabilities α and β, one gets the terms (3) and (4). Their graphs are displayed in figure 4.

If one postulates that a reflexive subject works slower, but more precisely, while an impulsive subject works faster but less precisely, the theoretical construct of reflexivity can be operationalised in the context of the signal detection theory.

The reflexivity ratio (RQ) expressed in this manner, is useful as an easy means to operationalise reflexivity, allowing for a sufficient skill-free statistic for the working style reflexivity measured in a single variable.



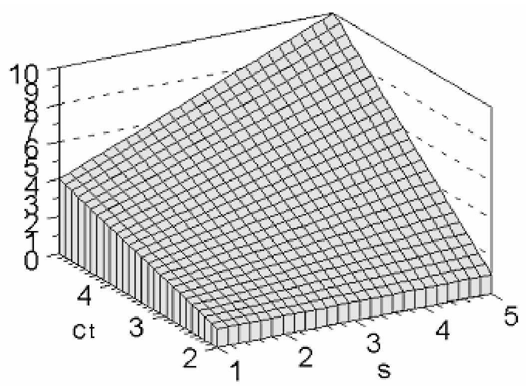
$$\alpha = \frac{1-\beta}{s*(k-1)} \tag{3}$$

$$\beta = 1-\left(\frac{k}{ct}*\frac{s}{1+s}\right) \tag{4}$$

Figure 4:
Alpha and beta error probability as a function of the odds ratio (s) and the number of inspections invested (ct), for 5 answer alternatives (k=5)

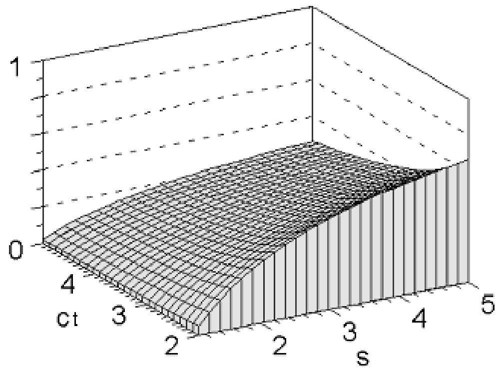
Table 4:
The cognitive construct Reflexivity in the paradigm of signal-detection theory

	α	β
Reflexive	Low	High
Impulsive	High	Low



$$RQ = \frac{\beta}{\alpha} \tag{5}$$

Figure 5:
Reflexivity ratio (RQ) as a function of the odds ratio (s) and the number of inspections invested (ct)



$$ER = 1 - \frac{\alpha}{2} - \frac{\beta}{2} \tag{6}$$

Figure 6:
Correctness of decision (ER) as a function of the odds ratio (s) and the number of inspections invested (ct)

In addition, this model creates a (working-)style-free measure for a subjects ability, which should be presented. The correctness of decisions (ER) is defined as the subjects probability of making correct decisions in the single inspection steps. This also includes (β) errors that do not manifest in wrong answers.

Therefore the correctness of the decision corresponds to a score, but a subjects behaviour is not only monitored on a distractor level but also on an item level. It is also possible to weight – according to the diagnostic question – the errors α and β differently to improve the prediction of criteria, where errors that cost time or errors that lead to wrong answers are more fatal than any other type of errors. However one should bear in mind that these weighted ER are no longer probabilities.

Model prerequisites

One should remember that this model implies several simplifications and therefore can only be adopted if special conditions apply:

1) Selection- or Decision Tasks: The correct answer shall not be produced but selected from known answer alternatives. This primarily includes pair comparisons and multiple choice alternatives.
2) Ambiguousness: For the model to be applied, complete inspections are required. This means that a person inspects every answer alternative at least once. Certain knowledge to the extent, that not all answer alternatives are inspected, contradicts the model – therefore all answer alternatives should be equally tempting. The decision should be based on preferring one alternative, not on definitely knowing the correct one.
3) Inspection Sequence: The model uses a non-systematic inspection sequence. The selection of which answer alternative to inspect next should be independent of which answer alternatives have already been inspected or what kind of information has al-

ready been accumulated. This allows all random or predefined inspection sequences but disallows strategic or adaptive selections.

5. Estimation of the constant c

As a special case of the measurement of the inspection time: Similar to the inspection time (IT) paradigm, one can expect that there are items that are easy enough so that a single inspection is sufficient to accumulate enough information to make a correct decision.

In multiple-choice this means that every answer alternative has been inspected once. The shortest possible reaction time for extremely easy items that lead to the correct answer corresponds to the inspection time.

$$c = \frac{1}{k * IT} \tag{7}$$

To avoid random effects, the mean over the 10% fastest responses could be used as a measurement.

By the definition of error probabilities: Alpha, Beta and ER are defined as probabilities. Numeric estimates of < 0 or > 1 are therefore impossible and can be derived from a wrong selection of the constant c. One way to estimate the constant c is to minimise the squared exceeding of these borders.

$$\Delta = \sum_{j=1}^{n} \begin{cases} \beta_j \geq 1 : (\beta_j - 1)^2 \\ 0 < \beta_j < 1 : 0 \\ \beta_j \leq 0 : (0 - \beta_j)^2 \end{cases} \overset{!}{=} MIN$$

As an individual measurement: The best way might be to work with individual (instead of global) IT constants, since in this way interindividual variances are eliminated. For each test, a measurement of the inspection time (for this kind of test material) has to be performed before the actual testing, which could be achieved with a set of extremely easy items. The personal IT constant could be the mean of the 10% of the fastest correct reactions of a given person.

Sequential Distractor Presentation: In variants of the sequential distractor presentation (Goethals, 1994), it is possible to measure the inspection time directly from the subjects behaviour for each answer alternative. The naming of this type of test presentation is not exactly precise as it deals with the presentation of all answer alternatives, not only with the presentation of the incorrect ones (the distractors).

While in parallel distractor presentation modes, all answer alternatives are simultaneously visible, in sequential distractor presentation modes, only one answer alternative is visible. In order to inspect another answer alternative, the respondent is required to manually activate the next answer alternative. Therefore a sequential distractor presentation is usually limited to computer-based testing.

Eye Tracking: With a certain amount of technical effort, eye movement can also be directly detected. This might be useful for research and validation purposes but of no use in practical assessment.

The error probabilities alpha & beta from an assessment-psychology point of view

From the practical point of view of assessment psychology, alpha errors can be understood as the affinity to accept wrong answer alternatives as correct and to use them as an answer for an item. This means that the subject has answered in a hasty manner or he/she was overtaxed by the task.

On the other hand, beta errors can be interpreted as a respondent's tendency to not recognize a correct answer alternative as correct or as a precautionary measure so as not to immediately determine the answer. This might be because the subject is inert or, again, overtaxed by the task.

With regard to the error probabilities alpha and beta, it is quite easy to see that no separation of skill and style variable is yet possible. Both measures are person specific or task specific mixtures of both. A fair style assessment cannot be conducted in this way. However, it is possible to model specific job specifications by using weighted alphas and betas to predict the job qualification:

For example, a task that requires extreme precision but where a fast working speed is not necessary, will primarily demand low alpha errors but might be rather invariant towards beta errors. In this case, alpha errors might receive a higher weight – the situation would be quite similar to pure power testing.

On the other hand, where performance under time pressure is required, alpha and beta errors might be equally weighted.

Such an estimation of weights for alpha and beta errors might derive from theory based on a job specification or on expert knowledge. An empirical variant could be to estimate the weights using the method of discriminant analysis whereby the predictive validity is maximized.

Reflexivity ratio from an assessment-psychology point of view

Since it is defined as the ratio of beta and alpha errors, the reflexivity ratio is a dimension free number. Therefore it is obvious from mathematical considerations that the reflexivity ratio has got to be skill-free. Therefore, one can conclude that it is the style of preference of one of the errors above the other. The risk of one of the error types is avoided more than the other.

This definition is quite similar to Kagans (not yet operational) *concern with errors* (Kagan & Kogan, 1970; Kagan, 1987).

A reflexive person tends to produce more openly visible alpha errors, which means that they respond with wrong answers. Therefore, the person will – within their own capabilities – try to work as slowly and precisely as possible.

An impulsive person tends to make more beta errors, which means that they waste time and ponder upon a task. Therefore, the person will – within their own capabilities – try to find a decision as fast as possible.

But impulsivity can also emerge from the fact that a subject has a certain amount of knowledge about the solution (the need for ambiguousness is violated in this case) and therefore has no need to completely inspect the item.

Correctness of decision from an assessment-psychology point of view

It is obvious that one chooses the probability of committing errors as a measurement for a person's ability. One can regard a person with a high correctness of decision (ER) as a person who works efficiently in the field of the ability dimension tested.

The proof that ER is a style-free measurement of a person's ability is not that trivial as seen in the case of the reflexivity ratio. For this one has to go a bit deeper into the signal detection theory.

Correctness of decision in the paradigm of the signal detection theory

If a subject is considered as a signal detection device, their efficiency can be expressed – regardless of their working style – as the area below the receiver operating characteristic (ROC) curve. The probability of making a correct decision corresponds exactly to this area and is therefore an appropriate theoretical operationalisation for the skill of that person.

But in practical terms this leads to several problems:

The problem of indeterminacy of acquired ROC curves

For a certain subject α and β error probabilities cannot be freely varied against each other. Even if a specific kind of instruction might have an effect on the working style, real subjects seem to stick to their individual style. And even if this was possible, there would be no way of creating a large pool of essentially equivalent items to measure a complete ROC curve.

On the other hand, it is impossible to estimate an unknown mathematical function with just one measurement derived from a test result. In order to facilitate this, simplifications based on theoretical considerations should be established so the real ROC can be approximated by a function that is easy to handle.

We can use the following information for the generation of this function.

ROC curves are:

- continuous
- monotonous
- include the points (0;0) and (1;1)

Using this knowledge, sensible assumptions about the shape of the ROC curve can be made in order to estimate the correctness of decision.

Assumption using the 3 points method

One possible way to assume the shape can be derived from the 3 known points on the ROC curve. Of course, on the one hand, the trivial (0;0) and (1;1) but on the other hand also the points (α; 1-β) which originate from the actual measurement.

The most straightforward way would be a linear approximation by connecting the points with straight lines like it is done in figure 7.

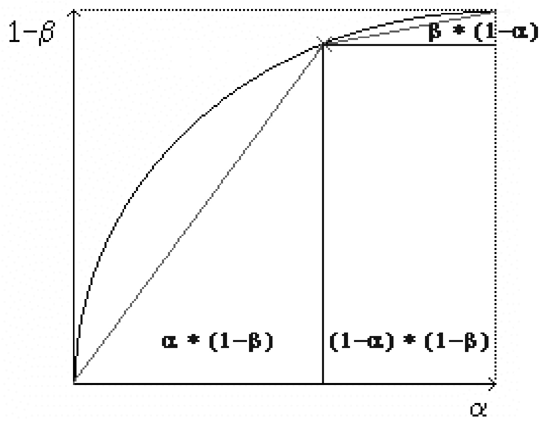


Figure 7:
ROC curve and its approximation using the 3 points method

The resulting area consists of one rectangle and two right triangles:

$$A = (1-\alpha)*(1-\beta) + \frac{\alpha*(1-\beta)}{2} + \frac{\beta*(1-\alpha)}{2}$$

which can be merged to the term for the Efficiency (6) that we postulated earlier in our theoretical considerations.

$$A = 1 - \frac{\alpha}{2} - \frac{\beta}{2}$$

Secant Assumption

Based on the knowledge that the ROC behaves in a continuous manner and increases monotonously, we can expect it to have a gradient of about 1 in the wider parts of the range where it is defined.

A parallel to the diagonal line, leading through the point (α; 1-β), as done in figure 8, might be an appropriate approximation of the unknown ROC curve.

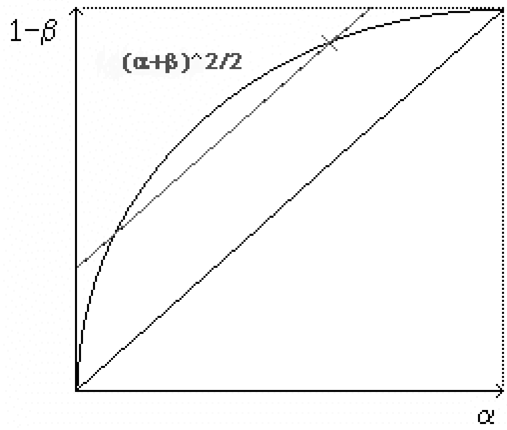


Figure 8:
ROC curve and its approximation using the secant method

Also in this case, a triangle of the area $\frac{(\alpha+\beta)^2}{2}$ can be formed, at least if $1-\beta > \alpha$, which should be the case for effective subjects. More generally formulated, this results in two possible cases:

$$A = \begin{cases} 1-\beta > \alpha : 1 - \dfrac{(\alpha+\beta)^2}{2} \\ 1-\beta \le a : \dfrac{(2-\alpha-\beta)^2}{2} \end{cases} \tag{8}$$

These two models are the easiest possible ways to make assumptions about the shape. Using nonlinear assumptions (eg. with a constant curve) might open a wide range of possible improvements for the estimation of style-free measurements of skill.

Empirical Study

In order to demonstrate this method, we will attempt to predict pre-knowledge of the test based on a single test result – in analogy to Van der Linden (2003), who used probabilistic speed & power models for this task. In his study, the specific working style of a respondent was suspected to be a predictor for test pre-knowledge – whereby an impulsive working style was considered to indicate pre-knowlede. A separate measure for the working style should therefore yield a good quality of detection.

The typical way to prepare a respondent for an ability test is to increase their specific experience with that test by providing the same or at least a similar test for training purposes. The worst case, when even the item set is identical, is simulated in this study.

This means that we are trying to detect if a person has seen a test before (or has perhaps been coached) by having a look at the working style of the subject. Is it possible to select subjects based on this method and could this contribute to the quality of a test, if regarded as a means of preventing or detecting faking?

To do so, 73 psychology students, aged between 18 and 48 years (m=22.2; sd=4.4), were tested twice with a modification (linear test form with a fixed length of 24 items, using a sequential distractor presentation) of the computer based LEWITE test (Wagner-Menghin, 2002) which measures the respondents vocabulary ability. In this case, the sequential distractor presentation was designed to react to movement of the computer mouse. In order to inspect one of the 8 answer alternatives, the respondent had to point on it, which enabled the inspection process for each respondent to be recorded. Thus, incomplete inspections, which would violate the models prerequisites, can be detected.

The secondary variable of LEWITE, the aspiration level, was not used in this experimental test form, in order not to risk a manipulation of the respondent's working style.

8.2 % of the respondents were male, 91.8% female.

A test / re-test design was chosen in order to maximize the information yielded. For the first test run, it was taken care that none of the respondents had seen the test LEWITE before. In the re-test all respondents could be considered to have test pre-knowledge to some extent since they had all done the test before. The time interval between the identical first and second testing run ranged from 2 to 4 weeks.

Results

In order to control if the model prerequisite of answer uncertainty was fulfilled, the number of inspections per item was controlled. Due to the sequential layout of the test, it was possible to record this information. No violation of this prerequisite could be detected. Both, true (m=1.19; sd=0.41) and false (m=1.08; sd=0.31) answer alternatives were inspected at least once. Therefore the prerequisite of complete item inspections can be considered to be accomplished and the model for the separation of skill and working style can be applied.

For the cases of both test runs (n=146), Cronbach's Alpha was calculated for the classical raw score and the separated measures. As expected, the disentangling of ability and working style leads to an increase in test reliability.

Table 5:
Comparison of the inner consistency (Cronbach Alpha) for the classical raw score, the separated error probabilities and the efficiency measure

	Cronbach Alpha
Score	0.826
α	0.957
β	0.952
Efficiency	0.958

In order to find out, if the selected variables were suitable to detect pre-knowledge, the data was pooled and a discriminant analysis was performed to predict which results derived from the first and which from the second (with pre-knowledge) testing runs. In order to control artificial assimilation of the model to the data, the predictions based on a jackknife algorithm were used for interpretation.

Table 6:
Prediction of test pre-knowledge based on different sets of predictors

Prediction by	No. of Predictors	Correctness of Prediction (Jackknife method)
Score	1	49.3 %
α, β	2	80.1 %
ER, RQ	2	78.8 %

It becomes visible that the use of α and β leads to a good prediction about whether a person already knows the test or not.

Since the data derives from a repeated measure, the dependent information should not be wasted. Therefore, an ANOVA for repeated measures was also used to analyze the effect of repeated testing in this setting.

Table 7:
Significance and effect size of the within-subject factor of test repetition on different dependent variables

	Sig	eta²
Score	0.004	11.0 %
α	0.011	8.7 %
β	< 0.001	57.3 %
ER	< 0.001	61.8 %
RQ	< 0.001	19.4 %

The results indicate that repeated testing has an effect on the score of a person and even without intentional preparation by using test advisers, the risk of faking with any kind of pre-knowledge is imminent. At least 11% of the variance of the score can be explained by the fact that a person has done the test before, so that they achieve better results in the second test run in comparison to the first one.

If the β of the second testing run is partialized out of this as a covariate, then this effect vanishes completely ($F=0.001$; $p=0.976$; $eta^2=0.0\%$). Furthermore, a simulation was performed to see if – after partializing the effect of β out of the result – a prediction of pre-knowledge based on the score is still possible.

The ROC in figure 9 shows, as expected, a weak prediction, based on the score. After stripping the effect of β, the score does not contain any more information on test pre-knowledge and – what is more important – pre-knowledge no longer affects the score in any way. It has become completely independent of the testing run.

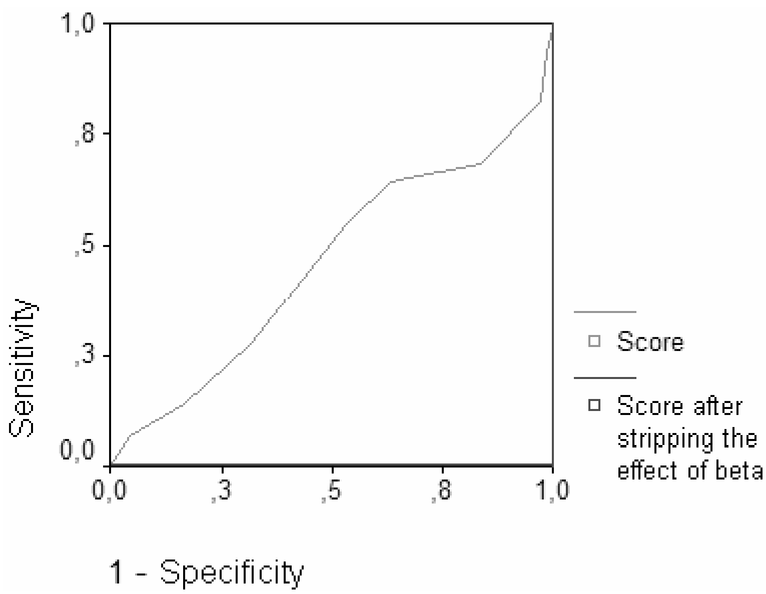


Figure 9:
ROC for the prediction of test run (test or re-test) by the score and by the score after stripping the effect of beta. The score after stripping the effect of beta remains at a Sensitivity of 0 unless a Specificity of 0 is reached; this means it carries no more information about the test run. The so cleansed score is no longer affected by test pre-knowledge.

Discussion

The results make it quite clear that at least in case of the LEWITE, intentional or unintentional faking by using specific pre-knowledge about the test material is possible and leads to better results. The alpha and beta error probabilities of a person carry information about their working style and prove to be a good predictor for pre-knowledge. However, the 80.1% correctness of prediction are not good enough to identify faking respondents, merely because a person behaves exceptionally impulsively. Mistakenly excluding such a person as a cheater would severely corrupt a tests claim of fairness.

However, a certain amount of optimisation can do the trick. Since stripping the β error from the effect of the score between the first and second test run removes all effects, we can prove that neither the score nor the variable β contain any information about the test run.

This fact allows us to find a reasonably mild threshold for the detection of pre-knowledge, so that:

1) nobody is wrongly accused
2) everybody who uses pre-knowledge and profits (in his score) from it is detected

However, of course, some subjects who have done the test before do remain undetected – but at least it can be said that none of them was able to improve their score by using any kind of pre-knowledge.

Therefore both claims – fairness and non-fakeability - can be ensured by using the appropriate working style variables in order to detect conspicuous working styles.

Outlook

The separation of skill and working style allows approaches that have been hard or even impossible to operationalise until now. The bulk of the multiple-choice tests that exist at the moment could be transformed into multifunctional tests (Wagner-Menghin, 2002), which not only measure an ability dimension but also the working style that a subject uses to accomplish this achievement. The subtest Decision Quality (Häusler, 2004a) of the test battery Intelligenz-Struktur-Batterie (Arendasy et al., 2004) and the objective personality test HKS Diagnostikum (Häusler, 2004b) are based upon this method of separation and therefore not only provide a measure of ability but also a measure of working-style.

At the same, the aspects of non-fakeability that have been expressed by latency times up until this point, can now be operationalised by skill-free working style variables as demonstrated in this study.

The separation of skill and style also opens a wide range of possible investigations into reflexivity, which has literally been impossible in the past because of the typical problems suffered by attempting to operationalise of reflexivity (Tiedemann, 1983).

References

1. Goethals, R. (1994). Die praktische Erprobung von Alternativen zur Multiple-Choice-Vorgabe bei Computertests. Unpublished Dissertation, University of Vienna
2. Häusler, J. (2004a). Entscheidungsfähigkeit. In Hornke, Arendsy, Sommer, Häusler, Wagner-Menghin, Gittler, Bognar & Wenzl. Manual: Intelligenz-Struktur-Batterie. Mödling: Schuhfried
3. Häusler, J. (2004b). Manual: HKS Diagnostikum. Mödling: Schuhfried
4. Hornke, L.F.; Arendasy, M.; Sommer, M.; Häusler, J.; Wagner-Menghin, M.; Gittler, G.; Bognar, B.; Wenzl, M. (2004). Manual: Intelligenz Struktur Batterie. Mödling: Schuhfried
5. Kagan, J. (1987). Misgivings about the Matching Familiar Figures Test: A brief reply to Block, Gjerde and Block (1986). Developmental Psychology, 23, 738-739
6. Kagan, J.; Kogan, N. (1970). Individual variation in cognitive process. In: Mussen, P.H. (Ed.): Carmichael's manual of child psychology. New York: Wiley, 1273-1365
7. Kliegl, R.; Mayr, U.; Krampe, R.T. (1994). Time-Accuracy Functions for Determining Process and Person Differences: An Application to Cognitive Aging. Cognitive Psychology, 26, 134-164
8. Mazur, J.E.; Hastie, R. (1978). Learning as an accumulation: A reexamination of the learning curve. Psychological Bulletin, 85, 1256-1274
9. Nährer, W. (1989). „Schnelligkeitsangepasstes Testen“: Testökonomie unter Berücksichtigung der Testzeit. In Kubinger, K.D. (Ed.): Moderne Testtheorie – Ein Abriss samt neuesten Beiträgen. Weinheim: Beltz, 219-236
10. Newell, A.; Rosenbloom, P.S. (1981). Mechanisms of skill acquisition and the law of practice. In: Anderson, J.R. (Ed.): Cognitive skills and their acquisition. Hillsdale: Lawrence Erlbaum Associates
11. Roskam, E.E. (1997). Models for Speed and Time-Limit Tests. In Van der Linden, W.J. (Ed.): Handbook of modern item response theory. New York: Springer, 187-208
12. Tiedemann, J. (1983). Der kognitive Stil Impulsivität – Reflexivität: Eine kritische Bestandsaufnahme. Zeitschrift f. Entwicklungspsychologie u. Pädagogische Psychologie, 15, 66-74
13. Van der Linden, W.J.; van Krimpen-Stoop, E.M.L.A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. Psychometrika, 68, 251-265
14. Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. Ergonomics, 13, 37-58
15. Wagner-Menghin, M.M. (2002). Der Lexikon-Wissen-Test (LEWITE) als Prototyp moderner Testkonzepte. In K.D. Kubinger (Ed.): Psychologische Diagnostik. Psychologie in Österreich, 22, 68-72