Psychology Science, Volume 47, 2005 (2), p. 262 - 278

Effects of feedback on performance and response latencies in untimed reasoning tests

JENS F. BECKMANN[1] & NADIN BECKMANN

Abstract

In an experimental study, a set of 12 number series problems with open-answer format had to be solved by a sample of 120 eighth and ninth graders randomly assigned to one of two test conditions (standard condition: no feedback; feedback condition: correct/incorrect item-by-item feedback). Task-related self-confidence and worry was measured before and after the performance test. Overall, results suggest that simple correct/incorrect feedback in performance tests does not provide the examinee with helpful information. Rather, it increases the level of worry, which tends to result in poorer performance. Moreover, the provision of feedback had no systematic effect on examinees' time behavior. The findings give no support for the assumption that time behavior in untimed performance tests is at least partially determined by non-intellectual variables such as self-confidence and worry.

Key words: feedback, intelligence tests, response latencies, I > C phenomenon, confidence, worry

[1] PD Dr. Jens F. Beckmann, and Dr. Nadin Beckmann, Yale University Center for the Psychology of Abilities, Competencies, and Expertise (PACE Center), Department of Psychology, Yale University; email: jens.beckmann@yale.edu
Correspondence concerning this article should be addressed to Jens F. Beckmann, The Yale University Center for the Psychology of Abilities, Competencies, and Expertise (PACE Center), 340 Edwards Street, New Haven, CT, 06520, USA

Feedback in performance tests

Feedback in the framework of psychological assessment is seen as an important intervention strategy to facilitate performance on tests. Within the concept of dynamic testing (Guthke & Wiedl, 1996; Guthke, Beckmann & Wiedl, 2003, see also Sternberg & Grigorenko, 2002), feedback is used to induce intraindividual differences in performance scores. In learning tests, for instance (Beckmann & Guthke, 1999), interindividual differences in the profit from feedback and elaborated thinking prompts has been proven to be a more valid indicator for an examinee's intellectual capacity than have test scores in traditional, feedback-free assessment procedures (Beckmann, 2001).

The implementation of feedback in performance tests in general is driven by the assumption that feedback will provide the examinee with useful information about his or her success in attempts to tackle the problems. It is expected that this information might be helpful to improve the examinee's performance on succeeding items within the test. However, feedback can be interpreted not only as a source of task-related information to the problem solver, but also as a source of information about his or her level of skills and abilities. Under this perspective, feedback might influence task motivation. The foundation for this research was laid by Thorndike's law of effect (Thorndike, 1911, 1932; Skinner, 1969). Based on the reinforcement approach, feedback – positive feedback in particular – results in positive effects in task motivation and henceforth in performance. This behavioristic approach has been overtaken by the cognitive perspective on feedback. This perspective allows one to conceptualize not only the effects of processing *task-related* information and its effect on performance, but also the effects of processing *self-related* information and its motivational effects.

There is also a considerably large body of research using non-contingent feedback. What we can learn from these studies is that the individual fit between the feedback received and the academic self-concept plays a crucial role whether feedback has a positive, negative or zero effect on performance. But the all over picture is anything but clear. Based on his study, Bossong (1982) has suggested that a perceived discrepancy between feedback and academic self-concept will result in an effort adjustment. That is to say, in the case of a positive discrepancy (feedback is more positive than the self-concept) the level of effort will be reduced, which might lead to a decrease in performance. The reaction to a negative discrepancy (feedback is more negative than expected, based on the self-concept) will be an increase in effort and henceforth in performance. However, other researchers such as Shrauger and Rosenberg (1970) have stated that feedback does have an impact on performance only if there is no discrepancy between feedback information and the level of self-esteem. Poor performers receiving rather negative feedback tend to show a decrease in performance, whereas good performers benefit from contingent positive feedback. In other words, feedback is good for the good ones and bad for the not so good ones (see also Meyer & Starke, 1981/1982; Schwarzer & Jerusalem, 1982).

A third position emphasizes the importance of congruence between self-concept and feedback independently from the type of the feedback (e.g., Stake, 1982). Congruent feedback positively influences performance, however incongruent feedback causes confusion and interferes with cognitive processes necessary to deal with the tasks. That is, examinees with high levels of self-esteem tend to profit from positive feedback, whereas examinees with low self-esteem might find negative feedback to be beneficial (see also within the framework of

self regulation theory: Idson & Higgins, 2000; Förster, Grant, Idson & Higgins, 2001; Van-Dijk & Kluger, 2004).

Coming back to studies employing performance-contingent feedback, we are confronted with empirical evidence suggesting that feedback intervention effects on performance are quite variable (Kluger & DeNisi, 1996, p. 254). In some conditions, feedback improves performance, in others, no effect on performance can be found, and in still other conditions, feedback can debilitate performance. Kluger and DeNisi's (1996) meta-analysis included 131 studies dealing with performance-related feedback effects in different settings. Surprisingly, more than a third of the 607 effects reported were negative. The overall effect size reported by Kluger and DeNisi was $d = 0.41$. Similar results were also revealed in a meta-analysis conducted by Bangert-Drowns, Kulik, Kulik and Morgan (1991), where 18 out of 58 feedback effects were negative, resulting in an overall effect size of $d = 0.26$.

What are the reasons for the high variability in the pattern of results of these feedback studies? Explanations can be expected from two perspectives: the situation-oriented perspective and the person-oriented perspective. An important situational characteristic that potentially moderates the effects of feedback is the level of elaboration of the feedback intervention itself. The level of elaboration ranges from simple correct/incorrect feedback as it is sometimes more or less deliberately – but nonetheless in contradiction to all rules of standardized testing – provided in individual intelligence assessment procedures, to the other pole of this continuum, marked by the highly individualized but standardized error-related thinking prompts as they are provided in learning tests (see e.g., Guthke & Beckmann, 2000a). The lower the level of elaboration (e.g., simple correct/incorrect), the less likely the chances are to benefit from this kind of feedback (e.g., Kulhavy, 1977) and to improve performance.

From a person-related perspective it is of particular relevance how the recipient processes the information provided as feedback. The question arises whether feedback is interpreted as a potential threat to self-esteem or whether the focus of attention is put on problem-relevant characteristics within the situation. Whereas the latter activates potentially beneficial cognitions about how to bridge the perceived feedback-standard discrepancy, the former results in less productive worry cognitions. Again, the less specific the feedback is in respect to information about successful problem-solving strategies, the more likely the chances are that the feedback – particularly the feedback "wrong" after an unsuccessful attempt to solve a problem – will be perceived merely as a negative evaluation of the examinee's own ability.

According to Meijer (2001) and Meijer and Elshout (2001), a lack of self-confidence is the central component of test anxiety. Anxious persons tend to interpret external and potentially evaluative stimuli in performance situations as threatening to their self-esteem (Sarason & Spielberger, 1975). Test anxiety or a lack of self-confidence can be seen as important performance-limiting personality factors that not only prevent potential profit from feedback, but could even enlarge the discrepancy between the manifest performance and the level of "true" ability (competency) in performance situations.

Moreover, it might be necessary to overcome the still dominant result-related perspective in feedback research that is primarily focused on test scores (number of correct responses). There is the need to consider process indicators to gain more insight about the mechanism of feedback (non-)effects. In other words, we should widen the perspective from analyzing problem-solving results to an investigation of the problem-solving process. One way is to

use response latencies in our attempts to better understand the effects of feedback on the problem-solving process.

Latencies in untimed reasoning tests

Within the framework of intelligence assessment in power tests, examinees are presented with a set of increasingly complex items without a time limit. The more problems that are solved in these tests, the higher the level of successfully mastered item complexity. The number of correct responses represents the estimator for the examinee's level of ability.

With the increasing employment of computerized test administration, not only can the responses be registered but also the time needed to answer each single item. This development and the old call for a more process-oriented evaluation of performance behavior in intelligence tests in general led to the reanimation of the interest in latencies in untimed reasoning tests (see, for example, Baxter, 1941; Ebel, 1953; Iseler, 1970; Nährer, 1982; Necka, 1992; Phillips & Rabbitt, 1995).

Besides the development in test presentation, recording, and scoring techniques, the theoretical understanding of the meaning of latencies is still quite limited. It is not clear yet to what extent latencies are meaningful indicators of task characteristics such as complexity (cf. Ebel, 1953). From a rather person-related perspective, the even more interesting, differentially-oriented question is whether response latencies in untimed intelligence tests can be seen either as additional indicators of the participants' intellectual capacity (e.g., speed of information processing, see, for example, Danthiir, Wilhelm & Schacht in this issue), or if they should rather be interpreted as indicators for non-intellectual personality factors (personal tempo, impulsivity, test anxiety, self-confidence, see Preckel & Freund and Troche & Rammsayer in this issue). In the former case, latencies would be more or less redundant to the test score (number correct) or at best they might serve as a kind of "backup indicator" for the examinee's ability (c.f. Hornke, 1997; see also Dörfler & Beckmann, 2003). In the latter case we would gain additional insight into the interplay of intellectual capacities and performance-related, non-intellectual personality factors during the process of dealing with the task complexity. However, so far no clear evidence could be found for that claim (no relationship between latencies and cognitive style such as impulsivity: Beckmann, 1999; Beckmann, 2000b; no relationship between latencies and neuroticism, extraversion, psychoticism, anxiety or need for achievement: Rammsayer, 1999). In any case, progress in understanding the meaning of response latencies in this setting will give us the chance to increase the quality of assessment tools employed in this field.

In recent studies dealing with response latencies in untimed performance tests, it has consistently been replicated that latencies for incorrect answers were longer than those for correct ones (Hornke, 1997; Beckmann, Guthke & Vahle, 1997; Beckmann, 2000a; Hornke, 2000; Rammsayer, 1999; Rammsayer & Brandler, 2003). This so called Incorrect > Correct phenomenon (I > C phenomenon[2]) is characterized by a high consistency across different domains, task complexities, and even test approaches (sequential vs. adaptive testing). Another interesting pattern of results is under discussion: The magnitude of the I > C phenomenon differs in accordance with the performance level of the examinee. More capable exami-

[2] Other authors use the term "False > Correct" (F>C phenomenon).

nees showed a larger difference between their response latencies for incorrect and correct answers (Beckmann et al., 1997; Beckmann, 2000a). Whereas the findings give very strong evidence for the generality of the I > C phenomenon, the pattern of results supporting the differential universality of this phenomenon is more heterogeneous (e.g., Hornke, 1997; Rammsayer, 1999, who reported no such differential effects of the I > C phenomenon). However, the seemingly differential universality of the I > C phenomenon raises the question of what poor performers do differently in untimed reasoning tests besides solving fewer problems. The analyses conducted by Beckmann (2000) show that there is no performance level-related difference regarding the latencies for correct responses, but there is regarding incorrect responses. Low performers tend to give their incorrect responses faster than do more successful performers. Does this mean that poor performers tend to give up too early in their attempts to solve the given reasoning problem? If so, latencies for incorrect responses might represent indicators for *mental effort* spent on hard items. Or do poor performers have to be faster than their more capable counterparts because of their "lower time horizon"? In other words, their limited capacity to process information might lead them to a seemingly too hasty pace while working on complex items. The latter speculative explanation puts the focus back on *mental efficiency* as the construct potentially covered by latencies.

Aim of the study

The goal of the study is twofold. First, we want to learn more about the mechanisms of the feedback – performance relation by analyzing the effects of feedback not only on performance scores but also on the time behavior in an untimed performance test. This introduces the second perspective in this study, the question about the meaning of response latencies in untimed performance tests. The study of relationships between individual differences in time behavior and non-intellectual personality factors within different test settings is supposed to shed light on the still open question of the validity of response latencies in untimed power tests.

To pursue these goals, hypotheses from three complexes need to be tested. The first group of hypotheses deals primarily with the replication of previous findings regarding the I > C phenomenon. In accordance with the assumption of the generality of the I > C phenomenon, latencies for incorrect responses should be longer than latencies for correct responses both under non-feedback conditions and feedback conditions (*generality hypothesis*). To test for evidence for the differential universality of the I > C phenomenon, the individual differences between latencies for correct and incorrect responses are expected to be larger for more capable subjects (*universality hypothesis*), independent of whether feedback is provided or not. The second set of hypotheses deals with direct effects of feedback. According to the discussion in the literature, simple correct/incorrect feedback should at best result in rather moderate positive effects on performance scores (*performance hypothesis*). One potential explanation for the limited performance-related feedback effects may be a higher level of worry cognition under feedback conditions (*worry hypothesis*). Besides the processing of feedback information as a potential self-threat, the provision of item-by-item feedback on the other hand should allow for a more realistic estimation of one's performance level (*confidence-performance hypothesis*).

The third set of hypotheses strives for a synopsis of ability- and personality-related factors and their relationships to individual differences in time behavior. Here, hypotheses need to be tested concerning whether non-intellectual personality factors such as worry or confidence can contribute to the prediction of time behavior in untimed reasoning tests independently of the person's level of ability (*validity hypothesis – worry, validity hypothesis – confidence*).

Method

Participants

A sample of 120 eighth and ninth graders from a northern England middle school voluntarily participated in our study. Participants range in age from 12 to 15 years old (M = 13.9, SD = 0.6), and 51 percent of them were female.

Instruments

The performance score on a set of 12 reasoning problems in the numerical domain (number series completion paradigm) served as the main dependent variable. The computerized test administration followed the principles of a power test procedure (increasing complexity of items, no time limit employed). In each item a series of 7 numbers had to be completed by providing the eighth link for the given series. The three easiest items (complexity level I) can be described as first-order arithmetic series (e.g., 44, 38, 32, 26, 20, 14, 8, ?) where the rule can be described by the subtraction or addition of a constant (e.g., " – 6" in the given example). The next three items (complexity level II) are second-order arithmetic series (e.g., 3, 4, 6, 9, 13, 18, 24, ?). Here the rule is more complex because the rule itself represents a first-order arithmetic series with a constant operation (e.g., "+1, +2, +3, …"). Items for which the determination rule uses multiplication or division belong to the category of geometric series and they therefore represent items in complexity level III. The rule to find the correct completion of a series like 64, 16, 12, 48, 52, 13, 9, ? can be described as series of a constant operand in combination with a series of three different operations (that is: "/4, +4, *4, /4, …"). The most difficult items in the test (complexity level IV) represent geometric series for which the rule combines the systematic change of the operation and the operand. For a number sequence belonging to the category of second order geometric series like 5, 6, 12, 14, 42, 45, 180, ? the rule would be "+1, *2, +2, *3, +3, *4, …".

In addition to the number of correct completions of the series, the latencies for each response were also recorded. The median response latencies for correct and incorrect answers were calculated for each subject. The difference of the median latency for correct and incorrect responses divided by the median latency for correct responses represents the individual I > C ratio.

After being presented with two example items on the computer screen, a low complexity item (as described for complexity level I) and a more complex one (representative for complexity level IV), the participants were asked to give their opinions to the following statements:

- “I am afraid I may not do as well on this test as I could” vs. “I am pretty optimistic that I will do as well on this test as I can.”
- “I feel pretty confident that I shall be able to solve most of the problems” vs. “I do not feel confident that I shall be able to solve most of the problems.”

The first set of statements is supposed to tap the participants’ level of situation-specific a priori worries as a component of task-related test anxiety. The second set of statements aims at identifying their task-related self-confidence.

The sentences in each statement represent the two extremes of an analogous scale 7 centimeters in length. The individual levels of worry and self-confidence, respectively, were operationalized by the graphic line segment on which participants had to place a mark indicating their response.

After finishing the reasoning test the same questions were administered again. Participants were asked to give their opinions to the following statements in respect to the level of a posteriori worry: “I am afraid I may not have done as well on this test as I could” vs. “I think I have done as well on this test as I could,” and in respect to the a posteriori level of task related self-confidence: “I think I have done well” vs. “I think I have done poorly.” These scores were recoded so that high scores stand for a high level of worry or high level of confidence, respectively.

Experimental design

Participants were randomly assigned to two experimental conditions. In one condition participants received item-by-item feedback regarding the correctness of the answer given. Participants in the control condition received no feedback. According to test scores collected independently of this study, both groups (no feedback group, $N_{FB-} = 60$, and feedback group, $N_{FB+} = 60$) did not differ in terms of their psychometric intelligence[3] (mean $IQ_{FB-} = 105.0$ [$SD = 11.5$]; mean $IQ_{FB+} = 106.7$ [$SD = 9.9$]; $F_{[1,118]} = 0.76$ $p > .05$).

Procedure

The test sessions took place in the school’s computer lab, and approximately 15 students participated per group session. Two test administrators were present in each session, one of the authors and a teacher. Each participant worked on his or her own. The test session started by providing a general instruction about the kind of problems to be solved. Then each participant worked on the computer at a self-paced tempo.

[3] IQ-estimations derive either from the Middle Years Information System (MidYIS, Durham University) for eighth graders, or the Cognitive Abilities Test (CAT, Thorndike & Hagen, 1993) for ninth graders, respectively. The MidYIS contains tasks to assess vocabulary, mathematical skills, information processing speed, spatial abilities, and reasoning abilities (see also www.midyisproject.org). The CAT assesses reasoning ability based on 10 subtests in three domains. For both instruments UK-specific national norms exist.

Results

In terms of the *generality hypothesis* it is expected that the I > C phenomenon emerges independently of test conditions (no-feedback vs. feedback). A multivariate analysis of variance with repeated measures (latencies for correct and latencies for incorrect) and the between subject factor "feedback" (no-feedback vs. feedback condition) reveals a significant main effect for latencies ($F_{[1,118]} = 93.238$, $p < .001$). At the same time, the nonsignificant main effect for test conditions ($F_{[1,118]} = 0.714$, $p > .05$) indicates that feedback does not have any substantial effect on time behavior. There is also no significant difference in the I > C phenomenon between the two test conditions ($F_{[1,118]} = 2.049$, $p > .05$).

Table 1 represents the mean latencies for incorrect responses and correct responses under the different test conditions.

Table 1:
The I > C Phenomenon in different test conditions.

	Test Condition	
Latencies	No Feedback	Feedback
Incorrect responses	57.58 (30.87)	52.70 (24.98)
Correct responses	27.92 (10.37)	30.70 (12.59)

Note: Time in seconds; values in parentheses are the standard deviations.

As stated in the *universality hypothesis*, the lower the performance level of the given examinee, the smaller the difference is expected to be between latencies for correct and incorrect responses. Based on the results of univariate regression analyses computed for each condition, the performance level and the I > C ratio share about 25% of variance under no-feedback conditions, and 23% variance under feedback conditions, respectively. The I > C phenomenon shows a differential universality consistently across test conditions.

The *performance hypothesis* deals with the question of whether feedback causes differences in test performance. Whereas under no-feedback conditions an average of 7.98 (SD = 2.22) items were solved correctly, the feedback sub-sample was successful on 7.23 items (SD = 1.95). This unexpected performance difference in disfavor of the feedback condition fails to pass the two-tailed threshold of statistical significance ($t_{[118]} = 1.96$, $p = .052$). Under the given circumstances (sample sizes, selected alpha level of 5%, two tailed), it can merely be concluded that feedback does not cause medium effects ($d \approx 0.50$) on performance. In contrast to our initial expectations, the results of this analysis refer potentially to a moderate *decline* in performance under test conditions in which simple correct/incorrect feedback is provided.

In accordance with the *worry hypothesis*, which was supported post hoc by the previously reported findings, it is expected that processing simple correct/incorrect feedback during performance tests does not necessarily provide the examinee with helpful information about potentially successful problem-solving strategies. On the contrary, such feedback information might rather increase the amount of worry cognitions while tackling the items. The significant interaction of test condition and level of worry before and after the test (see

Figure 1) confirms this assumption ($F_{[1,118]} = 4.366$. $p = .039$). Whereas the worry level before the test does not differ between the two sub-samples (no-feedback: 3.61 [*SD* = 2.13], feedback: 3.53 [*SD* = 2.19]), the level of worry cognitions caused by experiences during the test is apparently intensified by the item-by-item feedback (no-feedback: 3.50 [*SD* = 2.35], feedback: 4.37 [*SD* = 2.33]).

With the *confidence-performance hypothesis* the question was raised whether the provision of simple correct/incorrect feedback leads to a more adequate self-judgment of one's level of ability. In this case the relationship between the level of confidence reported immediately after the test and the actual level of performance is expected to be closer under feedback conditions. This research question calls for a moderator analysis (Aguinis, 2004; Baron & Kenny, 1986; Bartussek, 1970; Jäger, 1978; Saunders, 1956, 1966)[4]. The Multivariate

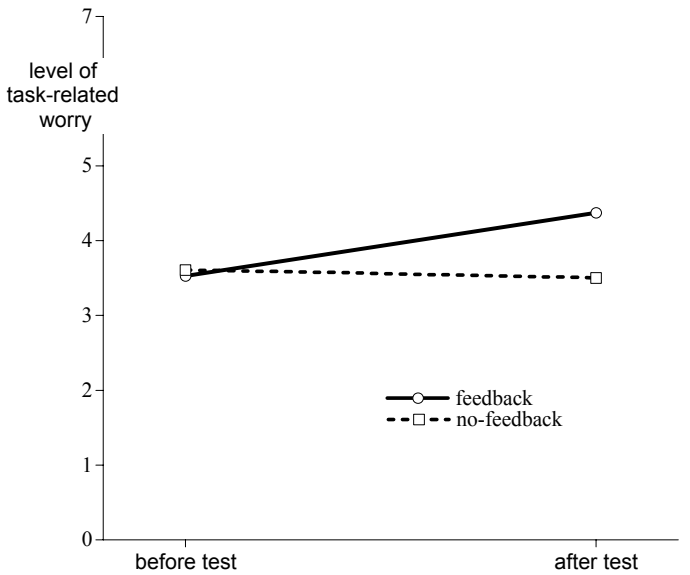

Figure 1:
Levels of task-related worry before and after the test depending on test condition.

[4] Often, for testing this kind of hypotheses the correlational method is employed. This bears at least two serious problems. First, if the variance in the independent variable (in our case: confidence) differs between the two subsamples (no-feedback vs. feedback) the correlation between confidence and performance will be (artificially) reduced in the subsample in which the variance is smaller. The effect virtually caused by the restriction of range will then inappropriately be interpreted as a "true" effect. Second, if the measurement error in the dependent variable differs as a function of the moderator variable spurious differences in correlations will be the result. This means in our case, if the performance measure under feedback condition contains more error variance than under no-feedback condition then the correlation between confidence and performance under feedback condition will be (artificially) smaller. Since regression coefficients are not influenced by differences in variances it is strongly recommended to use the method of moderated multiple regression (MMR).

Moderated Regression (MMR) analysis shows that the relationship between the level of confidence and performance is not moderated by feedback (see nonsignificant increase in R^2, or the nonsignificant standardized η weight of the moderator term, respectively, in Table 2). Feedback does not lead to a more realistic estimation of one's own level of ability.

Table 2:
Test for a moderator effect of feedback on the relationship between performance and confidence.

Step	Variables Entered	Standardized Coefficients	R^2 change	F
1	confidence feedback	.526 (6.83) * -.152 (-1.97) *	.308	26.01 *
2	confidence feedback confidence x feedback	.613 (5.46) * -.011 (-0.07) -.181 (-1.06)	.007	1.12

*Note: Values in parentheses represent t-values. The degrees of freedom for the F-test of R^2 change in the step 1 model are (2,117) and for the step 2 model they are (1,116). * significant on $\zeta \leq .05$.*

Post hoc, this result can be seen as in harmony with the worry-related findings. The increase of (rather unrealistic) worries under feedback conditions does not necessarily put the examinee in a better position for a more appropriate self-judgment regarding his or her level of ability. On the other hand, the bivariate correlation between the level of confidence and performance under no-feedback conditions of $r = .57$ clearly indicates that examinees do have a fairly realistic picture about their performance level even without receiving any information about the accuracy of their responses.

The final set of hypotheses deals with whether time behavior can be explained by non-intellectual personality factors. Therefore the regression of time behavior (I > C-ratio) on either the level of worry experienced during the test or the level of confidence acquired during the course of the test is calculated. Since we are interested in the unique portion of variance in the time behavior that could potentially be explained by these predictors, we needed to control for the examinee's level of intellectual ability (IQ). As we have learned from the analysis in relation to the *worry hypothesis,* examinees tend to produce a higher level of worry under test conditions in which simple correct/incorrect feedback is provided. Based on this result in extension of the *validity hypothesis–worry* we might expect a higher relationship between time behavior and worry under feedback conditions. If so, the test condition (no-feedback vs. feedback) serves as a moderator of the relationship between worry and the I > C ratio. To test this hypothesis a MMR[5] analysis was conducted (Table 3).

[5] For all MMRs reported in this paper the assumption of homogeneity of the error variances is tested by utilizing the routines provided at http://carbon.cudenver.edu/~haguinis/mmr/.

The results of this MMR analysis do not qualify feedback as a moderator of the relationship between time behavior and level of worry during the test. Although the worry level is increased in the feedback condition, its relevance as a potential determinant of time behavior in the test does not change. The result for the Step 1 model in Table 3 reveals also that worry does not predict time behavior during the test when controlled for IQ.

An analogous analysis for testing the *validity hypothesis-confidence* was conducted. Here the research question was whether the time behavior in the untimed reasoning test depends – at least partially – on the examinee's level of confidence.

The results in Table 4 give no support for the *validity hypothesis–confidence*. Similar to the results for worry as a potential predictor of time behavior, confidence is not related to the I > C ratio when controlled for IQ. Time behavior seems to be independent from the level of confidence, consistent across different test conditions.

Table 3:
Test for a moderator effect of feedback on the relationship between Worry and Time Behavior (I > C Ratio) when controlling for Intellectual Ability (IQ).

Step	Variables Entered	Standardized Coefficients	R^2 change	F
1	IQ	.262 (2.88) *	.097	4.16 *
	worry	-.088 (-0.96)		
	feedback	.205 (-0.11)		
2	IQ	.262 (2.88) *	.001	0.07
	worry	-.065 (-0.50)		
	feedback	-.068 (-0.39)		
	worry x feedback	-.056 (-0.27)		

*Note: Values in parentheses represent t-values. The degrees of freedom for the F-test of R^2 change in the step 1 model are (3,116), for the step 2 model they are (1,115). * significant on $\zeta \leq .05$.*

Table 4:
Test for a moderator effect of feedback on the relationship between Confidence and Time Behavior (I > C Ratio) when controlling for Intellectual Ability (IQ).

Step	Variables Entered	Standardized Coefficients	R^2 change	F
1	IQ	.245 (2.53) *	.097	4.17 *
	confidence	.093 (0.97)		
	feedback	-.118 (-1.33)		
2	IQ	.240 (2.44) *	.001	0.09
	confidence	.124 (0.89)		
	feedback	-.070 (-0.39)		
	confidence x feedback	-.061 (-0.30)		

*Note: Values in parentheses represent t-values. The degrees of freedom for the F-test of R^2 change in the step 1 model are (3,116), for the step 2 model they are (1,115). * significant on $\zeta \leq .05$.*

Discussion

The study presented focused primarily on the effects of feedback on performance in an untimed reasoning test. However, the operational perspective on test performance shall widen with the consideration of response latencies in addition to the number of correct answers. This serves the secondary goal of this study, to learn more about the meaning of latencies in untimed power tests and their potential value as a source of additional diagnostic information.

To address these research questions, an experimental design was chosen: A set of number series problems with open-answer format had to be solved either under standard conditions (no feedback) or under feedback conditions (item-by-item correct/incorrect feedback). Feedback is seen as an important intervention strategy in the framework of psychological assessment. On one hand, there are reasons to be optimistic that the provision of feedback during the test is not only helpful for the examinee to improve his or her performance but also helpful for the examiner to gain valuable information above and beyond what is gained if tests are administered in the traditional, non-dynamic way. This optimism is nurtured by empirical findings from studies evaluating the incremental validity of learning tests in which the provision of feedback is one important feature (Beckmann, 2001; Guthke & Wiedl, 1996). On the other hand, numerous findings from feedback research reduce the optimism regarding the beneficial effects of feedback on test performance (e.g., Kluger & DeNisi, 1996).

In the study presented, the performance scores (number of correct responses) of participants on the number series problems under feedback conditions did not differ positively from those who worked without feedback. Rather, a slightly negative effect of feedback on performance occurred. This result is in line with findings from other feedback-oriented studies (Delgado & Prieto, 2003; Rousseau & McKelvie, 2000; Stankov & Crawford, 1997), in which no or even negative feedback effects on test performance were reported.

That test condition (non-feedback vs. feedback) does not serve as a moderator of the relationship between confidence and performance in the study presented suggests that examinees do not gain new insight into their performance level when feedback is provided. Rather, it is more likely that simple correct/incorrect feedback is interpreted as mainly evaluative information. In this respect, the feedback "wrong" in particular might be processed as a potential threat to the examinees' self esteem (see also MacLeod, Williams, & Bekerian, 1991). This interpretation is supported by another finding in our study: Examinees working under feedback conditions reported a significantly higher level of worried thoughts after the test.

Worry may affect an allocation of attentional resources, which results in an absence of feedback benefits or even in performance deficits (see also Morris, Davis, & Hutchings, 1981; Thompson, Webber, & Montgomery, 2002). As a consequence of worry in a study by Metzger, Miller, Cohen, Sofka, and Borkovec (1990), impaired performance and even slowed response latencies in solving categorization tasks with feedback were reported. Interestingly, in our study the increase in worry under feedback is not reflected in examinees' time investment. According to Davis and Montgomery (1997), worried cognitions are associated with reduced problem-solving confidence, delays in decision-making, and poor performance (see also Dugas, Letarte, Rhéaume, Freeston, & Ladouceur, 1995). Our results indicate neither any effect of feedback on response latencies nor on confidence ratings but

they do give evidence for an increased level of worry and a tendency toward performance decline.

Although a process-oriented approach – not only in the assessment of intellectual capacities – is often claimed to be more appropriate than the predominant product-oriented approach (merely reflecting on the number of total correct answers), little is known about the validity of potential process variables. By analyzing the meaning of response latencies the question is addressed whether time behavior in power tests can serve as such a process-oriented variable.

As mentioned before, in previous studies focusing on time behavior in untimed performance tests, it was consistently found that latencies for incorrect answers are longer than those for correct answers, independently from the item paradigms employed, the complexity of the items, and the test presentation modes. The I > C phenomenon is replicated in the study presented under both the non-feedback condition and the feedback condition. This result gives further support for the generalizability of the I > C phenomenon. In respect to the also hypothesized differential universality of the I > C phenomenon, the I > C ratio was found to be larger the higher the performance level. However, the performance-related I > C effect was not affected by test condition.

The overall perspective on the findings in the study presented suggests that the I > C phenomenon is "merely" related to the examinee's level of capacity. We gained no support for the assumption that time behavior in untimed reasoning tests is at least partially determined by non-intellectual personality factors such as worry or confidence experienced while solving the items.

We still do not know the exact meaning of response latencies in untimed performance tests, but at least we might know better now what their meaning is not. Further research attempts should concentrate on the evaluation of the relevance of latencies as a process-oriented and ability-related variable (see Danthiir et al. or Hornke, both in this issue).

With respect to feedback in performance tests, based on the results reported here, we can conclude that the provision of simple correct/incorrect feedback in performance tests is not helpful, since it (a) does not contain any "new" or helpful information when the examinee is familiar with the test demands, and (b) causes worry, which interferes potentially with task-related information processing. Our suggestion to test administrators therefore must be: Do not provide feedback! Our recommendation to test takers is: Try to ignore it if it is provided!

References

1. Aguinis, H. (2004). Regression analysis for categorical moderators. New York, NY: Guilford Press.
2. Bangert-Drowns, R. L., Kulik, C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. Review of Educational Research, 61, 213–238.
3. Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychology research: Conceptual, strategic and statistical considerations. Journal of Personality and Social Psychology, 51, 1173–1182.
4. Bartussek, D. (1970). Eine Methode zur Bestimmung von Moderatoreffekten [A method for the determination of moderator effects]. Diagnostica, 16, 57–76.

5. Baxter, B. (1941). An experimental analysis of the contributions of speed and level in an intelligence test. Journal of Educational Psychology, 32, 285–296.
6. Beckmann, J. F. (1999). Latenzzeiten in Reasoning-Tests – nur Ausdruck eines kognitiven Stils? [Latencies in reasoning tests - Just an manifestation of cognitive style?]. Paper presented at the 5. Arbeitstagung der Fachgruppe Differentielle Psychologie, Persönlichkeits-psychologie & Psychologische Diagnostik, Bergische Universität Gesamthochschule Wuppertal.
7. Beckmann, J. F. (2000a). Differentielle Latenzzeiteffekte bei der Bearbeitung von Reasoning-Items [Differential effects of response latencies in reasoning tests]. Diagnostica, 46, 124–129.
8. Beckmann, J. F. (2000b). Zur diagnostischen Relevanz des Latenzzeitverhaltens in Reasoningtests. [On the diagnostic value of timing behavior in reasoning tests]. Paper presented at the 42. Kongreß der Deutschen Gesellschaft für Psychologie, Jena.
9. Beckmann, J. F. (2001). Zur Validierung des Konstrukts des intellektuellen Veränderungspotentials [The validation of the construct "intellectual change potential"]. Berlin: Logos.
10. Beckmann, J. F., & Guthke, J. (1999). Psychodiagnostik des schlussfolgernden Denkens [Psychological assessment of reasoning ability]. Göttingen: Hogrefe.
11. Beckmann, J. F., Guthke, J., & Vahle, H. (1997). Analysen zum Zeitverhalten bei computergestützten adaptiven Intelligenz-Lerntests [Analyses of time behavior in computerized adaptive learning tests]. Diagnostica, 43, 40–62.
12. Bossong, B. (1982). Kognitive Dissonanz, Attribution und Leistung. [Cognitive dissonance, attribution and performance]. Zeitschrift für Sozialpsychologie, 13, 194–208.
13. Delgado, A. R., & Prieto, G. (2003). The effect of item feedback on multiple-choice test responses. British Journal of Psychology, 94, 73–85.
14. Dörfler, T., & Beckmann, J. F. (2003). Mental effort or mental efficiency. The meaning of response latencies in reasoning tasks. Paper presented at the 11th Biennial Meeting of the International Society for the Study of Individual Differences, Graz, Austria.
15. Dugas, M. J., Letarte, H., Rhéaume, J., Freeston, M. H., & Ladouceur, R. (1995). Worry and problem-solving: Evidence of a specific relationship. Cognitive Therapy and Research, 19, 109–120.
16. Ebel, R. L. (1953). The use of item response time measurements in the construction of educational achievement tests. Educational & Psychological Measurement, 13, 391–401.
17. Förster, J., Grant, H., Idson, L.C., & Higgins, E.T. (2001). Success/failure feedback, expectancies, and approach/avoidance motivation: How regulatory focus moderates classic relations. Journal of Experimental Social Psychology, 37, 253–260.
18. Guthke, J., & Beckmann, J. F. (2000). The learning test concept and its application in practice. In C. S. Lidz & J. G. Elliott (Eds.), Dynamic Assessment: Prevailing models and applications (pp. 17–69). Oxford, UK: Elsevier.
19. Guthke, J., Beckmann, J. F., & Wiedl, K. H. (2003). Dynamik im dynamischen Testen [Dynamics in dynamic testing]. Psychologische Rundschau, 54, 225–232.
20. Guthke, J., & Wiedl, K. H. (1996). Dynamisches Testen. Zur Psychodiagnostik der intraindividuellen Variabilität [Dynamic testing: Assessment of intraindividual variability]. Göttingen: Hogrefe.
21. Hornke, L. F. (1997). Untersuchung von Itembearbeitungszeiten beim computergestützten adaptiven Testen [Investigation of response latencies in computerbased adaptive testing]. Diagnostica, 43, 27–39.
22. Hornke, L. F. (2000). Item response times in computerized adaptive testing. Psicologica, 21, 175–189.

23. Idson, L. C., & Higgins, E. T. (2000). How current feedback and chronic effectiveness influence motivation: Everything to gain versus everything to lose. European Journal of Social Psychology, 30, 583–592.
24. Iseler, A. (1970). Leistungsgeschwindigkeit und Leistungsgüte [Performance speed and accuracy]. Weinheim: Beltz.
25. Jäger, R. S. (1978). Differentielle Diagnostizierbarkeit in der psychologischen Diagnostik [Differential diagnosticability in psychological assessment]. Göttingen: Hogrefe.
26. Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin, 119, 254–284.
27. Kulhavy, R. W. (1977). Feedback in written instruction. Review of Educational Research, 47, 211–232.
28. MacLeod, A. K., Williams, J. M. G., & Bekerian, D. A. (1991). Worry is reasonable: The force of explanations in pessimism about future personal events. Journal of Abnormal Psychology, 100, 478–486.
29. Meijer, J. (2001). Learning potential and anxious tendency: Test anxiety as a bias factor in educational testing. Anxiety, Stress, & Coping, 14(3), 337–362.
30. Meijer, J., & Elshout, J. J. (2001). The predictive and discriminant validity of the zone of proximal development. British Journal of Educational Psychology, 71, 93–113.
31. Metzger, R. L., Miller, M. L., Cohen, M., Sofka, M., & Borkovec, T. D. (1990). Worry changes decision-making: the effect of negative thoughts on cognitive processing. Journal of Clinical Psychology, 46, 78–88.
32. Meyer, W.-U., & Starke, E. (1981/1982). Das Einholen begabungsrelevanter Informationen in Abhängigkeit vom Konzept eigener Begabung: Eine Feldstudie. [The seek for ability-related information and its dependence on self-esteem: A field study] Archiv für Psychologie, 134, 105–115.
33. Morris, L. W., Davis, M. A., & Hutchings, C. A. (1981). Cognitive and emotional components of anxiety: Literature review and revised worry-emotionality scale. Journal of Educational Psychology, 73, 541–555.
34. Nährer, W. (1982). Zur Beziehung zwischen Bearbeitungsstrategie und Zeitbedarf bei Denkaufgaben [The relationship between strategy and time consumption in thinking problems]. Zeitschrift für experimentelle und angewandte Psychologie, 29, 147–159.
35. Necka, E. (1992). Cognitive analysis of intelligence: The significance of working memory processes. Personality and Individual Differences, 13, 1031–1046.
36. Phillips, L. H., & Rabbitt, P. M. A. (1995). Impulsivity and speed–accuracy strategies in intelligence test performance. Intelligence, 21, 13–29.
37. Rammsayer, T. (1999). Zum Zeitverhalten beim computergestützten adaptiven Testen: Antwortlatenzen bei richtigen und falschen Lösungen [On time behavior in computerized adaptive testing: Response latencies for correct and false responses]. Diagnostica, 45(4), 178–183.
38. Rammsayer, T., & Brandler, S. (2003). Zum Zeitverhalten beim computergestützten adaptiven Testen: Antwortzeiten bei richtigen und falschen Lösungen sind intelligenzunabhängig [Time behavior in computerized adaptive testing: Response times for correct and incorrect answers are independent from intelligence]. Zeitschrift für Differentielle und Diagnostische Psycholgy, 24, 57–63.
39. Rousseau, F. L., & McKelvie, S. J. (2000). Effects of bogus feedback on intelligence test performance. Journal of Psychology, 134, 5–14.

40. Sarason, I. G., & Spielberger, C. D. (1975). Stress and anxiety (Vol. 2). Washington, DC: Hemisphere Publishing.
41. Saunders, D. R. (1956). Moderator variables in predicition. Educational and Psychological Measurement, 16, 209–222.
42. Saunders, D. R. (1966). The moderator variable as a useful tool in predicition (paper 1954). In A. Anastasi (Ed.), Testing problems in perspective: 25th anniversary volume of topical readings from the Invitational Conference on Testing Problems (pp. 301–306). Washington, DC: American Council on Education.
43. Schwarzer, R., & Jerusalem, M. (1982). Selbstwertdienliche Attributionen nach Leistungsrückmeldungen. [Self-serving attributions after performance feedback] Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 14, 47–57.
44. Shrauger, J. S., & Rosenberg, S. E. (1970). Self-esteem and the effects of success and failure feedback on performance. Journal of Personality, 38, 404–417.
45. Skinner, B. F. (1969). Contingencies of reinforcement: A theoretical analysis. New York, NY: Appleton-Century-Crofts.
46. Stake, J. E. (1982). Reactions to positive and negative feedback: Enhancement and consistency effects. Social Behavior & Personality, 10, 151–156.
47. Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. Intelligence, 25, 93–109.
48. Sternberg, R., & Grigorenko, E. (2002). Dynamic testing. Cambridge, UK: Cambridge University Press.
49. Thompson, T., Webber, K., & Montgomery, I. (2002). Performance and persistence of worriers and non-worriers following success and failure feedback. Personality & Individual Differences, 33, 837–848.
50. Thorndike, E. L. (1911). Animal intelligence. New York, NY: Macmillan.
51. Thorndike, E. L. (1932). The fundamentals of learning. New York, NY: Teachers College of Columbia University.
52. Thorndike, R. L., & Hagen, E. P. (1993). Cognitive abilities test. Itasca, IL: Riverside Publishing.
53. Van-Dijk, D., & Kluger, A. N. (2004). Feedback sign effect on motivation: Is it moderated by regulatory focus? Applied Psychology: An International Review, 53, 113–135.

For appendix see the next page.

Appendix

Descriptives and correlations of the measures for the subsample Working under No-Feedback Condition ($N = 60$).

Measure	*M*	*SD*	conf 1	conf 2	wrry 1	wrry 2	per-form	lat corr	lat incorr	I > C-ratio
confidence pre	4.77	1.38								
confidence post	3.42	1.89	.41							
worry pre	3.61	2.13	-.18	-.35						
worry post	3.50	2.35	-.27	-.55	.53					
performance	7.98	2.22	.47	.57	-.17	-.37				
latency correct	27.92	10.37	.08	-.10	.19	.03	-.09			
latency incorrect	57.58	30.87	.24	.16	.11	-.11	.42	.27		
I > C-ratio	1.23	1.08	.27	.27	-.09	-.13	.49	-.42	.70	
IQ	105.06	11.45	.30	.51	-.04	-.22	.65	-.29	.13	.34

Descriptives and correlations of the measures for the subsample Working under Feedback Condition ($N = 60$).

Measure	*M*	*SD*	conf 1	conf 2	wrry 1	wrry 2	per-form	lat corr	lat incorr	I > C-ratio
confidence pre	4.13	1.90								
confidence post	3.22	1.01	.27							
worry pre	3.53	2.19	-.34	-.10						
worry post	4.37	2.33	-.27	-.54	.25					
performance	7.23	1.95	.13	.50	.08	-.53				
latency correct	30.70	12.59	.06	-.15	-.02	.10	-.18			
latency incorrect	52.70	24.98	.07	.16	.03	-.19	.58	-.07		
I > C-ratio	0.99	1.20	.02	.13	.01	-.16	.49	-.57	.77	
IQ	106.72	9.92	-.03	.29	-.07	-.25	.46	-.12	.23	.23