Coaching for the PISA test [1, 2]

Martin Brunner, Cordula Artelt, Stefan Krauss & Jürgen Baumert

Abstract Coaching is known to improve student performance on tests with high personal relevance ("high-stakes tests"). We investigate whether the same holds for a test that has no personal relevance for the students taking it ("low-stakes test"). More specifically, we explore whether student performance on the reading and mathematics assessments of the OECD's Programme for International Student Assessment (PISA) can be fostered by coaching (and administering a pretest). Coaching and pretest effects were studied for each content domain separately in a pre-/posttest quasi-experimental design. To examine differential effects of academic tracks, samples were drawn from German Hauptschule and Gymnasium schools. Results show that only the combined effects of pretesting and coaching have substantial positive effects on student performance. Implications for the interpretation of large-scale assessment programs are discussed.

Keywords *Test coaching, low-stakes test, reading achievement, mathematics achievement*

1. Coaching for low-stakes tests

1.1 Background to the study

In this article we address an aspect of learning and instruction that targets student outcomes on achievement tests, namely, test coaching. The issue of coaching is of great relevance to students whose future academic and occupational careers hinge on their performance in high-stakes tests such as the SAT. Meta-analyses indicate that, if an individual is highly motivated, coaching can boost performance on high-stakes test to a moderate degree (e.g., Becker, 1990; Kulik, Bangert-Drowns, & Kulik, 1984) and consequently benefit individual careers.

In contrast, large-scale assessment programs, such as the Programme for International Student Assessment (PISA) set up by the Organisation for Economic Co-operation and Development (OECD) or the National Assessment of Educational Progress (NAEP) mandated by the US Congress and administered by the National Center for Education Statistics (NCES) at the US Department of Education, do not influence individual student careers and thus can be considered low-stakes tests. It cannot be assumed that the results of meta-analyses concerning coaching for high-stakes tests generalize to low-stakes situations such as these.

Why might it be important to consider the effects of coaching for low-stakes tests? Although low-stakes tests do not usually have any direct implications for the students taking them, their results are highly relevant to politicians responsible for the outcomes of the school system as well as to the principals of individual schools. For instance, the No Child Left Behind Act makes schools responsible for the achievement of their students. Not reaching a certain benchmark may have negative consequences for a school's funding and reputation. School principals thus have a vested interest in fostering students' test performance, and may use test coaching as a means to improve their results on assessments. However, the fundamental idea of school assessment programs is to capture the overall effect of schooling on students' achievements rather than the effects of coaching in a certain school. Therefore, "low-stakes" tests that influence political decisions or the allocation of funding to schools should be as immune to coaching as possible.

Interestingly, to our knowledge, there is no research on whether coaching prescribed by outside agents (low-stakes situations) produces similar effects to coaching programs that students elect to join in order to boost their test scores and hence enhance their future prospects (high-stakes situations). Thus, in this article, we address the question of whether low-stakes tests are susceptible to coaching.

1.2 Components of coaching

Activities designed to prepare students for a specific test with the aim of optimizing their performance outcomes are subsumed under the term "test coaching" or simply "coaching." Although there is much variety in the activities involved, most coaching programs comprise at least one of the three following components (Allalouf & Ben-Shakhar, 1998):

(1) They familiarize participants with key elements of the test. If test material (e.g., previous versions of the test) is available, participants are exposed to typical test instructions, items, time limits, and

1 This research was supported by a grant from the German Research Foundation (DF 025).
2 Accepted Author's Manuscript. This article was published in:
Learning and Instruction 17 (2007), 111-122doi: 10.1016/j.learninstruc.2007.01.002

question-and-answer formats by training under authentic conditions ("familiarity approach"). From this perspective, implementing a pretest as a control condition in an experimental design also qualifies as coaching for the posttest.

(2) Participants are prepared for the content of the test. Students studying for a mathematics test, for example, receive targeted coaching in the topics likely to come up ("content approach").

(3) Participants are taught "test-wiseness" strategies. In one of the seminal works on this topic, Millman, Bishop, and Ebel (1965) characterize test-wiseness as the individual's ability to utilize the characteristics and formats of the test or test-taking situation to do well ("test-wiseness approach"). For instance, students learn general test-taking strategies (e.g., not to waste too much time on difficult items) as well as specific strategies for certain item types (e.g., how to use distractors in multiple-choice questions to infer the right answer).

Components 1 to 3 can be applied to both low- and high-stakes tests. Note that component 2 bears the strongest resemblance to the instruction provided in ordinary school lessons. Regular school instruction and cognitive training programs usually differ from coaching programs in that the former aim to enhance more general, transferable abilities, rather than to improve student outcomes on a specific test (Hasselhorn & Hager, 2001).

1.3 Effects of coaching and pretesting

To date, research on coaching has concentrated on how, and to what extent, it is possible to improve students' scores on high-stakes tests. In the following, we summarize the relevant findings from this field of research. Numerous studies have explored the effects of test coaching and pretesting on performance in ability tests (e.g., Allalouf & Ben-Shakhar, 1998; LeGagnoux, Michael, Hocevar, & Maxwell, 1990; Powers, 1985, 1987; Powers & Rock, 1999). The major findings of these studies are summarized in the reviews by Bond (1989) and Powers (1993), and in various meta-analyses (Becker, 1990; DerSimonian & Laird, 1983; Kulik, Bangert-Drowns, et al., 1984; Kulik, Kulik, & Bangert-Drowns, 1984; Messick & Jungeblut, 1981; Samson, 1985; Willson & Putnam, 1982). Table 1 documents the results of the most frequently cited meta-analyses by Becker (1990) and Kulik, Bangert-Drowns, and Kulik (1984).

Table 1 reports both the incremental effects of coaching (upper half) and the effects of pretesting (lower half). Although pretesting can be considered a coaching activity in its own right, it is considered separately. This is because the incremental effects of a coaching program are usually measured by comparing the achievement gains of a group that worked on pre- and posttests *and* took part in coaching with the gains of a group that only took the pre- and posttests. Hence, the incremental coaching effects reported in Table 1 reflect the effects of all coaching activities (as summarized in section 1.2) *except for* taking the pretest.

Table 1: Effects of coaching and of pretesting on test performance

Meta-analysis	Sample	Number of studies	Test	Mean effect size
Incremental effects of coaching				
Becker (1990)	All coaching studies relating to the SAT	70	SAT-M	0.38
		70	SAT-V	0.26
	– of which published	25	SAT-M	0.18
		25	SAT-V	0.13
	– of which unpublished	45	SAT-M	0.46
		45	SAT-V	0.31
Kulik, Bangert-Drowns et. al. (1984)	Coaching studies with a control group design	38	SAT (total score) and other assessments	0.33
	– of which SAT	14	SAT (total score)	0.15
	– of which other assessments	24	Various assessments	0.43
Effects of pretesting				
Becker (1990)	All coaching studies (incl. pretest) relating to SAT	16	SAT-M	0.16
		28	SAT-V	0.23
Kulik, Bangert-Drowns et. al. (1984)	Coaching studies (incl. pretest) with a control group design	20	SAT (total score) and other assessments	0.24
	– of which SAT	14	SAT (total score)	0.21
	– of which other assessments	14	Various assessments	0.25

Note: Effect sizes correspond to mean achievement gains from pre- to posttest. Becker (1990) calculated the change in the standardized mean scores from pre- to posttest for coached und uncoached groups separately, and took the difference between these figures as the mean effect size. This yields the incremental effect of coaching over and above simply taking the same test (or a parallel version of it) twice. The mean effect sizes calculated by Kulik, Bangert-Drowns, and Kulik (1984) using a slightly different method can be interpreted in the same way. SAT-M = SAT mathematics score; SAT-V = SAT verbal score.

The effects of pretesting constitute the impact that the very act of working on a pretest has on performance in the posttest. The effects observed for participants who took a pre- and posttest without participating in a coaching program are presented in the lower half of Table 1. First, we discuss coaching effects.

1.3.1 Effects of coaching

Coaching effects tend to be smaller for the Scholastic Aptitude Test (SAT, i.e., the test that regulates college admission in the United States) than for other assessments (Kulik, Bangert-Drowns, et al., 1984). Moreover, coaching has more pronounced positive effects on performance in the mathematics part of the SAT (SAT-M) than in the verbal part (SAT-V) (Becker, 1990; Kulik, Bangert-Drowns, et al., 1984).

Which components of coaching programs are particularly effective? In the most comprehensive meta-analysis to date, Becker (1990) found instruction in test-wiseness strategies (cf. Samson, 1985) and exposure to typical test items to be the most effective elements of SAT coaching programs (for an overview see Powers, 1988). Moreover, because it is difficult to disentangle the effects of test-wiseness instruction from those of the other components of coaching programs, its effectiveness may in fact be underestimated (Becker, 1990; Kulik, Bangert-Drowns, et al., 1984).

Flippo, Becker, and Wark (2000) determined that the most effective coaching programs last between six and nine hours. According to Bunting and Mooney (2001), the minimum time needed to produce an effect over and above the effect of simply taking a pretest is three hours of coaching. However, the relationship between the length of the program and achievement gains is very weak. On average, ten hours of coaching increases test performance by 0.07 standard deviations (Becker 1990).

It is interesting to note that the published and unpublished coaching studies in Table 1 differ in terms of their mean effect sizes. Because published studies tend to be better controlled than unpublished studies, the true effects of test coaching are likely to be better represented by the more conservative effect sizes reported in the published studies. Powers and Camara (1999) also point out that many of the studies conducted by commercial coaching companies fail to use a control group design and have numerous other methodological flaws, making it reasonable to assume that they tend to overestimate the total effects of test coaching.

1.3.2 Effects of pretesting

In many studies designed to measure the effects of coaching, a pretest is administered prior to the coaching program to assess participants' baseline capabilities. The achievement gains of students who sit the pre- and a posttest without taking part in any coaching activities provide an estimate of the effects of simply taking a pretest. In these cases, the pretest is usually a parallel form of the posttest. As shown in the lower half of Table 1, pretesting has a small incremental effect that is, rather surprisingly, similar in magnitude to the incremental effect of test coaching. The effect of pretesting tends to be larger for the verbal part of the SAT (Becker, 1990) and seems to be independent of the test employed (Kulik, Bangert-Drowns, & Kulik,1984).

A few studies have investigated the effects of pretesting without addressing coaching. Burke (1997) and LeGagnoux and colleagues (1990), for instance, found that the effects of pretesting seem to differ across cognitive abilities subtests. Moreover, the magnitude of pretest effects is dependent on how similar the pre- and posttest material is, and on the number of pretests administered (Kulik, Kulik, et al., 1984). The strongest effects of pretesting can be expected when less than two weeks expire between the pre- and the posttest (Willson & Putnam, 1982), but pretest effects can last several years (Burke, 1997; Kulik, Kulik, et al., 1984).

1.4 Research questions

High-stakes tests (e.g., the SAT) differ from low-stakes test in one major respect – the results of high-stakes tests are crucial for the future educational careers of the students taking them (e.g., as the decisive criterion for college entry), whereas the results of low-stakes assessment have no direct implications for individual students. Given that the results of "low-stakes" tests (e.g., the large-scale assessments implemented in NAEP or PISA) can serve as a basis for political decisions (e.g., funding), however, it is surprising research has not yet considered coaching effects in the context of low-stakes tests.

Since 2000, the OECD's Programme for International Student Assessment (PISA) has been assessing 15-year-old students' reading, mathematics, and science literacy in a three-year testing cycle. More than 150,000 students from over 30 different countries participated in each of the first two assessments: PISA 2000 and PISA 2003 (OECD, 2001, 2004). The performance of German students was unexpectedly low in PISA 2000 (and only marginally better in PISA 2003), prompting a controversial political discussion. It seems reasonable to assume that those with high-stakes interests in the results may have been tempted to try to improve students' performance in PISA 2003 by prescribing coaching in their area of responsibility.

In this article, we empirically investigate whether the PISA assessments are susceptible to coaching. We cannot simply generalize findings on coaching for high-stakes tests to the low-stakes PISA assessment because previous studies have shown students' motivation to participate in coaching programs to be a crucial factor for their success (Allalouf & Ben-Shakhar, 1998, Powers, 1987). As students taking low-stakes tests have no vested interests in their personal results, they might not be motivated to learn from the coaching program.

In particular, we address the following research questions:

(1) *Effects of "authentic" coaching.* What is the effect of coaching conducted by a class teacher? Here, we are interested in the effects of coaching activities that might actually have been implemented by the teachers of students participating in PISA 2003.

(2) *Effects of pretesting*. Previous research has shown the effects of pretesting to be similar in size to the effects of coaching in high-stakes tests. Therefore, we empirically investigate the effect of pretesting on performance on the PISA test.

(3) *Domain specificity*. In high-stakes tests, the susceptibility to coaching has been shown to differ across mathematical and verbal subtests. Assuming that coaching effects (or pretest effects) are found, we will investigate whether these effects differ across content domains (mathematics and reading).

(4) *The role of prior knowledge*. Students with more prior knowledge might gain more benefit from coaching (Kulik, Kulik, et al., 1984) or from taking a pretest (Kulik, Bangert-Drowns, et al., 1984). We investigate whether the effects of coaching (or pretesting) differ between students attending the college-track Gymnasium (indicating high prior knowledge) and students attending the vocational-track Hauptschule (indicating low prior knowledge).

2. Method

2.1 Study design

Effects of coaching and pretesting on performance on the PISA test were investigated for the domains of reading and mathematics separately in a pre/post-test quasi-experimental design (*reading study* and *mathematics study*). For both domains, the effects were studied separately for students attending Hauptschule and Gymnasium schools. Our study used the original PISA materials, but was independent from and not embedded in the PISA 2003 assessment organized by the OECD. The design of the study is illustrated in Table 2.

2.2 Sample

Based on the information available on schools that had participated in PISA 2000, we selected schools that were comparable in terms of (a) the percentage of students with immigration backgrounds and (b) students' socioeconomic background characteristics. We asked the principals of these schools whether their 9th grade German or mathematics teachers would agree to participate in our study. These teachers then decided, in consultation with their principals, which condition was to be implemented at their school. Entire 9th grade classes were sampled from 11 Hauptschulen (33 classes) and 11 Gymnasium schools (33 classes), with all students in each school being assigned to the same quasi-experimental condition. Data were obtained from a total of 1,323 students. A detailed sample description is given in Table 2.

Table 2: Study design and student characteristics by group

Academic track	Group (quasi-xperimental condition)	1st week Pretest	2nd week	3rd week Posttest	N: number of students (number of classes)	Mean age (SD)	Percentage female	Percentage native speakers
Mathematics study								
Hauptschule	Pretest	M[b]	Regular lessons	M, R[b]	97 (7)	16.2 (1.0)	50.0	68.1
	Pretest & coaching[a]	M	Coaching	M, R	164 (10)	16.1 (0.9)	45.0	85.4
Gymnasium	Pretest	M	Regular lessons	M, R	154 (7)	15.6 (0.7)	54.7	86.6
	Pretest & coaching	M	Coaching	M, R	252 (10)	15.6 (0.8)	71.7	82.5
Reading study								
Hauptschule	Pretest	R	Regular lessons	R, M	139 (7)	16.0 (1.0)	43.0	87.6
	Pretest &coaching	R	Coaching	R, M	176 (9)	16.1 (0.7)	34.9	70.0
Gymnasium	Pretest	R	Regular lessons	R, M	196 (9)	15.6 (0.7)	60.5	93.3
	Pretest &coaching	R	Coaching	R, M	145 (7)	15.7 (0.7)	61.5	95.1

[a] In the text, this group is termed "coaching group."

[b] M: PISA mathematics test. R: PISA reading test. The administration of R and M was counterbalanced in the posttest

Student participation was contingent on parental consent. In order to provide informed consent, parents and students were informed about the study design (e.g., students in the pretest condition knew that students from other schools would practice for the PISA test and vice versa). Only data from students who participated in both the pretest and posttest were included in our analyses. For detailed sample descriptions, see Table 2. Note that students who participated in PISA 2000 were not included in our sample as they had either finished school or progressed to higher grades.

2.3 Coaching activities

What would be the effects of teachers deciding (or being instructed to) coach their students for an upcoming PISA assessment? In other words, what are the effects of *authentic coaching*? This question cannot be addressed by prescribing coaching activities to teachers (e.g., materials prepared by professional coaching companies); rather, teachers must be allowed to design their own coaching programs.

In a pilot study, we asked 72 teachers (not participating in the present study) to imagine that they were planning to prepare their students for the PISA test. On average, teachers stated that they would dedicate 3 (mathematics) or 4.5 (German) hours to PISA coaching. Moreover, teachers indicated that they would focus on the content of the upcoming test and use original test items from previous tests. Interestingly, none of the teachers mentioned instruction in test-wiseness strategies. In general, there was no great variability in the approaches that German and mathematics teachers identified for PISA preparation.

Based on this information, we asked the teachers in the present study to dedicate four lessons (each lasting 45 minutes) to coaching activities in the second experimental week, giving an approximate total of 3 hours of coaching. We provided the teachers in the coaching condition with released PISA items, as well as with the framework document outlining the theory behind the construction of the PISA tests. We did not provide any information on coaching activities (e.g., teaching test-wiseness skills).

2.4 Measurement instruments

All tests were conducted by trained administrators. Pre- and posttests were administered to all student participants. In order to control for selection effects, we also obtained data on students' sociodemographic and motivational characteristics as well as on their reasoning ability and school grades.

The items measuring the dependent variables (reading and mathematics literacy) were selected from the sizeable PISA 2000 item pool (for a description of the literacy framework and sample items see OECD, 1999). The time allocated for each booklet was 1 hour, allowing almost all students to work through all items without time pressure.

The items of both the reading scale and the mathematics scale can be approximated by a unidimensional Rasch model (Adams & Wu, 2002; Klieme, Neubrand, & Lüdtke, 2001). This allowed us to construct parallel test forms (with no item overlap) for both measurement points and to compare students' performance at pre- and posttest. To increase the statistical power of the study, we compiled separate test booklets that ensured maximum reliability for students attending Gymasium and Hauptschule, respectively. The student achievement data were scaled with Conquest (Wu, Adams, & Wilson, 1998) while anchoring the item parameters to the values derived in the PISA 2000 study. This produced a weighted likelihood estimate (WLE; Warm, 1989) for the reading or mathematics achievement of each student at the pre- and posttest. All WLE parameters were linearly transformed to the metric of the PISA 2000 study (PISA metric: M = 500, SD = 100). Differential booklet difficulty was controlled using the method outlined in Adams and Wu (Adams & Wu, 2002), including cognitive, motivational, and socio-demographic student characteristics as background variables.

2.5 Implementation check

We checked whether the teachers' implementation of authentic coaching met our requirements by analyzing whether the teachers spent as much time as we had instructed on coaching their students, whether they thought that the treatment was authentic, and what kind of approaches and materials/items they used in their coaching lessons.

The main implementation check concerns the time teachers spent coaching students for the test, as well as the time they spent preparing for lessons (see Table 3).

In two of the nine Hauptschule reading classes, the treatment was not implemented as intended, with only 30 minutes being allocated to coaching. Table 3 thus presents data for both the whole group of nine Hauptschule reading classes and, separately, for the group of seven classes in which the treatment was implemented as intended. Overall, our requirement of 3 hours being allocated to coaching was met in most classes. As indicated by the standard deviation, there are a few classes where fewer than 3 lessons were dedicated to coaching. Because the time actually spent on coaching exceeded one hour, however, these classes were not excluded from the analysis.

Teachers' self-reports provided additional support for the authenticity of the treatment and its implementation. Most of the teachers said that they would have coached their students in the way we suggested if they had been selected for the real PISA test. Only two teachers indicated that they would have spent more time on coaching.

Prior to coaching, teachers were also asked to predict the average achievement gain of their class, given that the maximum test score on the PISA test was 100 points. In both subjects, the mean anticipated achievement gain was 10.7 points, with a smallest anticipated gain of 5 points. Teachers thus seem to have been rather positive about the beneficial effects of their intervention, another sign that the treatment had been implemented successfully.

Furthermore, teachers were asked to write protocols about their preparation for the coaching lessons and to list the materials and items they used. In general, teachers used the materials and items we provided. Moreover, 29% of teachers also used their own materials and items (31% in reading and 27% in mathematics classes).

In terms of teaching approaches, the most notable difference between the regular classes and the coaching classes is the more frequent use of repetition

and rehearsal techniques during coaching. Although test-wiseness strategies were neither included in the materials sent to teachers nor suggested as the content of lessons, half of the coaching teachers made their students aware of at least one test-taking strategy.

2.6 Statistical analyses

As indicated by the teachers' responses in our pilot study as well as by the implementation check (see Table 3), teachers did not use pretests as a means of coaching their students. In order to calculate the effect of *authentic* coaching, we therefore have to estimate the incremental effect of coaching without the effect of the pretest.

Incremental coaching effects were estimated for each academic track and content domain separately by running regression analyses with the post-test achievement score as the dependent variable. A significant unstandardized regression weight of a dummy variable indicating the treatment condition represents the incremental effects of test coaching in the PISA metric: Students who participated in a coaching program were coded as 1 and students who worked on the pre- and posttest only were coded as 0.

To control for differences between quasi-experimental conditions, we included student characteristics such as gender, age, parents' occupational status (Ganzeboom, de Graaf, Treiman, & de Leeuw, 1992), and immigration status (Kunter et al., 2002) in the regression model. Furthermore, we included an "effort thermometer" tapping test motivation (Kunter et al., 2002), cognitive ability measures (in particular, the figural analogies subtest from the Cognitive Abilities Test (KFT; Heller & Perleth, 2000), the estimation, number sequences, fact-opinion, word knowledge, and verbal analogies subtests from the Berlin Intelligence Structure Test (Jäger, Süß, & Beauducel, 1997)) and, of course, achievement in the pretest.

As we are dealing with clustered data (students nested within classes), the standard errors of our statistics had to be corrected (e.g., Snijders & Bosker, 1999). To obtain standard errors that take the clustered nature of the data into account, we used the Mplus 3.01 program (Muthén & Muthén, 1998-2004) with the complex option for all analyses.

Data on the predictor variables were missing for some students (the highest percentage of missing data being 11% for parents' occupational status). We therefore imputed five data sets (cf. Schafer & Olsen, 1998) using Norm (Schafer, 2000). All regression analyses were run five times with Mplus, and the results combined according to the formula proposed by Rubin (1987).

The significance level was set to $p < .05$ for all inferential statistical analyses.

Table 3: Time spent on coaching-related activities by subject and school type

Academic track		Minutes allocated to coaching in class	Number of lessons allocated to coaching	Minutes teachers spent preparing for coaching lessons
	SD			
Mathematics study				
Hauptschule (N=10)	M	135.0	3.7	252
	SD	27.1	0.7	138.0
Gymnasium (N=10)	M	117.3	3.3	92.35
	SD	27.5	0.5	61.59
Reading study				
Hauptschule (N=7/9)[a]	M	128.6 / *109.4*	4.1 / *3.4*	154. 2 / *123.1*
	SD	43.1 / *53.2*	0.9 / *1.6*	93.7 / *97.8*
Gymnasium (N=6)[b]	M	116.3	3.2	50.8
	SD	23.2	0.8	13.9

a: Data based on either 7 or 9 classes, see text for details.

b: One teacher did not provide data concerning the time spent on coaching activities. However, we were able to infer from lesson protocols that this teacher did prepare her students for the PISA reading test.

3. Results

In order to address research question 1 (effects of authentic coaching), we need to disentangle the combined effects of coaching and pretesting. Therefore, we first report our findings for research question 2 (effects of pretesting, see section 3.1) before going on to document the combined effects of pretesting and coaching (see section 3.2). Finally, we estimate the incremental effects of authentic coaching by comparing the performance gains of the pretest and coaching groups while controlling for differences between the quasi-experimental conditions (see section 3.3). This procedure allows us to address research questions 3 (domain specificity) and 4 (role of prior knowledge) simultaneously.

3.1 Effects of pretesting

3.1.1 Mathematics study

Students in both academic tracks profited slightly from taking a pretest (see Table 4). On average, Hauptschule students gained 12 points (d = .20), while Gymnasium students gained 6 points (d = .11) on their pretest performance. However, neither of these effects was significantly different from zero.

3.1.2 Reading study

Surprisingly, mean performances decreased from pre- to posttest in both academic tracks (see Table 4). The mean decrease was 14 points (d = -.17) for Hauptschule students and 9 points for Gymnasium students (d = -.13). Because none of the mean differences were statistically significant different from zero, we do not interpret this tendency.

3.2 Combined effects of pretesting and coaching

3.2.1 Mathematics study

The combined effect of taking a pretest and being coached by the class teacher was slightly positive for both academic tracks (see Table 4). On average, Hauptschule students gained 9 points (d = .16), while

Gymnasium students gained 24 points (d = .36) on their pretest performance. Both effects were significantly different from zero.

3.2.2 Reading study

As shown in Table 4, the mean reading achievement of Hauptschule students slightly decreased by 4 points (d = -.04) at posttest. This effect was not statistically significant. When the two classes with just 30 minutes of coaching were excluded from the analysis (the number of students dropping to 134), performance decreased by 9 points (d = - .11). The effect was not statistically significant. Gymnasium students who took the pretest and participated in the coaching program showed slight performance gains. Their scores improved by 12 points (d = .18). This effect was statistically significant.

3.3 Incremental effects of coaching ("authentic coaching")

3.3.1 Mathematics study

When controlling for the influence of the predictor variables on the mathematics posttest scores, the incremental effect of coaching for Hauptschule students is almost zero (the unstandardized regression weight B was 1.52, corresponding to an effect size d of -.03 standard deviations).

The incremental effect for Gymnasium students was 10.39 points on the PISA metric (d = .16). Neither regression weight was statistically different from zero. Table 5 provides an overview of the regression analyses.

Table 4: Descriptive and inferential statistics for each quasi-experimental condition

Academic track	Group (quasi-experimental condition)	Pretest		Posttest		*d*	*r*
		M	SD	M	SD		
Mathematics study[a]							
Hauptschule	Pretest	445	54	457	69	0.20	0.59
	Pretest & coaching	429	59	438	69	0.16 *	0.56
	Coaching	No directly corresponding quasi-experimental group				– 0.03	0.58
Gymnasium	Pretest	562	64	568	62	0.11	0.64
	Pretest & coaching	539	63	563	68	0.36 *	0.60
	Coaching	No directly corresponding quasi-experimental group				0.16	0.61
Reading study[b]							
Hauptschule	Pretest	431	75	417	84	–0.17	0.59
	Pretest & coaching	426	81	422	93	–0.04	0.58
	Coaching	No directly corresponding quasi-experimental group				0.07	0.58
Gymnasium	Pretest	566	67	557	91	–0.13	0.58
	Pretest & coaching	565	59	577	89	0.18 *	0.54
	Coaching	No directly corresponding quasi-experimental group				0.43 *	0.56

Note: d: For the effects of pretesting and of pretesting and coaching, d reflects average gain scores divided by the pooled pretest standard deviation. For the effects of coaching, d reflects the unstandardized regression weight of the dummy variable that indicates quasi-experimental conditions divided by the pooled pretest standard deviation. r: r indicates the stability of the rank ordering of students across weeks. In the "Pretest" and "Pretest & coaching" groups, r reports the correlation within each track-specific group. In the "coaching" group, r reports the correlation within each track across the "Pretest" and "Pretest & coaching" groups.

*$p < 0.05$

[a]: statistics refer to mathematics achievement

[b]: statistics refer to reading achievement

3.3.2 Reading study

For Hauptschule students, coaching had a small incremental effect (B = 5.11, d = .07) that was not statistically different from zero. When the two classes with just 30 minutes of coaching were excluded from the analysis, the incremental effect decreased slightly and was still not statistically significant (B = -1.8, d = -.02). Authentic coaching only had a substantial (and statistically significant) incremental effect of 27.24 points on the PISA metric for Gymnasium students (d = .43).

Table 5: Unstandardized regression weights (B) for the estimation of incremental coaching effects

Predictor	Mathematics Study				Reading Study			
	Hauptschule		Gymnasium		Hauptschule		Gymnasium	
	B	z	B	z	B	z	B	z
Achievement at pretest	0.47	4.43	0.39	7.01	0.55	7.54	0.60	9.59
Mathematics grade	-1.98	-0.43	-9.51	-2.78	-0.44	-0.11	-8.20	-1.76
German grade	-1.48	-0.33	-5.56	-1.89	-11.33	-2.89	-9.29	-1.71
Fact opinion	-3.31	-0.63	7.61	2.58	0.77	0.13	4.56	0.68
Number sequences	14.12	2.95	4.41	1.47	-0.47	-0.12	10.80	2.38
Verbal analogies	10.27	2.11	5.13	1.59	8.23	1.48	10.78	2.58
Estimation	2.76	0.88	-4.51	-1.88	-9.21	-1.47	-2.02	-0.49
Word knowledge	12.81	2.51	5.57	1.70	-0.41	-0.07	8.88	1.48
Figural analogies	5.46	1.97	9.98	5.73	6.37	1.98	-2.15	-0.63
Effort thermometer	-0.67	-0.40	3.31	3.26	4.77	2.25	0.09	0.04
Gender (female = 0, male = 1)	14.39	1.86	1.29	0.18	-3.01	-0.51	-17.19	-1.89
Number of books at home	2.88	0.92	0.62	0.25	-3.47	-1.14	-6.99	-2.04
Age	-0.21	-0.05	1.83	0.63	-0.56	-0.09	-2.17	-0.42
Immigration status I: first generation (=1) vs. others (=0)	-17.79	-0.79	-4.79	-0.41	-52.80	-2.51	7.22	0.30
Immigration status II: native speakers (=1) vs. others (=0)	-25.84	-3.22	14.11	2.06	-16.95	-1.84	-9.29	-0.63
Parents' highest occupational status	0.29	0.87	0.05	0.26	0.03	0.13	0.08	0.29
Coaching (=1) vs. pretest (=0)	**-1.52**	**-0.14**	**10.39**	**1.09**	**5.11**	**0.32**	**27.24**	**3.46**
Regression intercepts	269.7	2.85	258.2	4.38	185.1	1.76	277.9	3.26
R^2	0.47		0.52		0.40		0.40	

Note: The dependent variable in all regression analyses was achievement at posttest. All predictors were measured at pretest. B: averaged unstandardized regression weight across all five imputed data sets. z: normally distributed statistical test computed by dividing the unstandardized regression weight by the corresponding standard error (Muthén & Muthén, 1998-2004, p. 481).

4. Summary of results and discussion

What are the incremental effects of authentic coaching? According to Cohen (e.g., 1992), authentic coaching has small to medium effects across content domains for students in higher academic tracks (i.e., Gymnasium schools). In the present study, however, the mean effect of d = .43 observed for Gymnasium students in the domain of reading is largely due to the students in the pretest group exhibiting an unexpected drop in performance at posttest. Authentic coaching had no incremental effects on either mathematics or reading in the lower academic track (Hauptschule).

What is the effect of pretesting? The mathematics pretest had small positive effects across academic tracks, while performance on the reading test in fact declined in both tracks. Note, however, that none of these effects were statistically different from zero. Thus, the question of whether pretesting alone has positive effects on performance on low-stakes tests warrants further research.

To summarize our results concerning research questions 1 and 2, if it were necessary to choose between pretesting and authentic coaching, our data suggest that students – at least those in higher academic tracks – might benefit modestly from coaching

Turning to the combined effect of pretesting and authentic coaching, our results again indicate that Gymnasium students benefit more than their peers at Hauptschule: Both effects (d = .36 for mathematics and d = .18 for reading) were statistically significant different from zero. Students in the lower academic

track also profited from this combined treatment in the domain of mathematics (d = .16, also significant).

How does our pattern of results for a "low-stakes" test correspond with the findings reported for high-stakes tests in published studies? The incremental effects of coaching and pretesting in high-stakes tests presented in Table 1 are slightly higher, yet comparable to the effects found in the present study (aggregated across academic tracks). Thus, the personal relevance of the test results seems to play a minor role (see also Baumert & Demmrich, 2001). Coaching in a classroom setting (e.g., for PISA) can be almost as effective as professional coaching programs (e.g., for the SAT).

What are the limitations of our study? First, the sample size was too small to isolate effective components of coaching or to identify possible moderator effects of other teacher or school characteristics. Further research with a large enough sample on the level of classes/schools is needed here. Moreover, whether or not our results hold for low-stakes tests in general remains an open question. A study by te Nijenhuis, Voskuijl, and Schijve (2001) suggests that the higher a test's loading on general intelligence, the less susceptible it is to coaching effects (see also Jensen, 1998).

What are the implications of our study for the interpretation of the PISA 2003 results? In the German PISA 2003 study, an item tapping coaching activities was included in the student questionnaire. It emerged that 26% of students had practiced for the mathematics test in some manner. However, Prenzel, Drechsel, Carstensen, and Ramm (2004) found no significant performance differences between coached and uncoached students (in either the mathematics or the reading test). This finding is in line with our results indicating that authentic test coaching alone has almost no effects on performance (with the exception of the effect for reading at Gymnasium, which is due to the unexpected decline in the performance of the control group at posttest).

Our study also shows what form *test-specific* learning and instruction may take in schools: It is only when students are administered a pretest *and* given coaching that notable positive effects on their PISA outcome are observed. One treatment alone does not guarantee success. Because the items actually employed in the PISA assessments are not publicly available (and only a few of them are released after each assessment cycle), teachers cannot easily conduct PISA pretesting. As such, we do not consider coaching to present a great threat to the validity of the PISA study.

Acknowledgements

We would like to thank (in alphabetical order): Michael Becker, Irmela Buddeberg, Andrea Derichs, Carolin Guzmán, Olaf Menzel, Henriette Paschen, Charlotte Rosenbach, Steffen Sameiske, and Annika Seehausen for their valuable help during all phases of conducting this study, and Susannah Goss for editing the English text.

References

Adams, R., & Wu, M. (2002). *PISA 2000 technical report.* Paris: OECD.

Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement, 35*(1), 31-47.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology and Education, 16*(3), 441-462.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60*(3), 373-417.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 429-444). New York: American Council on Education/ Macmillan.

Bunting, B. P., & Mooney, E. (2001). The effects of practice and coaching on test results for educational selection at eleven years of age. *Educational Psychology, 21*(3), 243-253.

Burke, E. F. (1997). A short note on the persistence of retest effects on aptitude scores. *Journal of Occupational and Organizational Psychology, 70,* 295.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review, 53*(1), 1-15.

Flippo, R. F., Becker, M. J., & Wark, D. M. (2000). Preparing for and taking tests. In R. F. Flippo & D. C. Caverly (Eds.), *Handbook of college reading and study strategy research* (pp. 221-260). Mahwah, NJ: Lawrence Erlbaum Associates.

Ganzeboom, H. B. G., de Graaf, P. M., Treiman, D. J., & de Leeuw, J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21,* 1-56.

Hasselhorn, M., & Hager, W. (2001). Kognitives Training [Cognitive training]. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (2nd ed., pp. 343-350). Weinheim: Psychologie Verlagsunion.

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision. Manual* [Cognitive Abilities Test for Grades 4-12, Revision, Manual]. Göttingen: Hogrefe.

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur: Test, Form 4.* Göttingen: Hogrefe.

Jensen, A. R. (1998). The g factor. *The science of mental ability.* Westport: Praeger.

Klieme, E., Neubrand, M., & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse [Mathematical literacy: Test conception and results]. In J. Baumert, E. Klieme,

M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 139-190). Opladen: Leske + Budrich.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C.-L. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*(2), 179-188.

Kulik, J. A., Kulik, C.-L., & Bangert-Drowns, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*(2), 435-447.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente [PISA 2000: Documentation of the study instruments]*. Berlin: Max-Planck-Institut für Bildungsforschung.

LeGagnoux, G., Michael, W. B., Hocevar, D., & Maxwell, V. (1990). Retest effects on standardized structure-of-intellect ability measures for a sample of elementary school children. *Educational and Psychological Measurement, 50,* 475-492.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 8*9, 191-216.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25,* 707-726.

Muthén, L. K., & Muthén, B. O. (1998-2004). *Mplus: The comprehensive modeling program for applied researchers: User's guide.* Los Angeles, CA: Statmodel.

te Nijenhuis, J., Voskuijl, O. F., & Shijve, N. B. (2001). Practice and coaching on IQ tests: Quite a lot of g. *International Journal of Selection and Assessment, 9*(4), 302-308.

Organisation for Economic Co-operation and Development (OECD). (1999). *Measuring student knowledge and skills: A new framework for assessment.* Paris: Author.

Organisation for Economic Co-operation and Development (OECD). (2001). *Knowledge and skills for life: First results from the OECD Programme for International Student Assessment (PISA) 2000.* Paris: Author.

Organisation for Economic Co-operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003.* Paris: Author.

Powers, D. E. (1985). Effects of coaching on GRE aptitude test scores. *Journal of Educational Measurement, 22*(2), 121-136.

Powers, D. E. (1987). Who benefits most from preparing for a "coachable" admissions test? *Journal of Educational Measurement, 24*(3), 247-262.

Powers, D. E. (1988). *Preparing for the SAT: A survey of programs and resources* (College Board Report No. 88-7). New York: College Entrance Examination Board.

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and one update. *Educational Measurement: Issues and Practice, 12,* 24-39.

Powers, D. E., & Camara, W. J. (1999). *Coaching and the SAT I (RN-06).* New York, NY: College Board.

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36*(2), 93-118.

Prenzel, M., Drechsel, B., Carstensen, C. H., & Ramm, G. (2004). PISA 2003 - Eine Einführung [PISA 2003 - An introduction]. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost & U. Schiefele (Eds.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (pp. 314-354). Münster: Waxmann.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: J. Wiley.

Samson, G. E. (1985). Effects of training in test-taking skills on achievement test performance: A quantitative synthesis. *Journal of Educational Research, 78*(5), 261-266.

Schafer, J. L. (2000). NORM for Windows 95/98/NT (Version 2.03) [Computer program]. Retrieved November 2, 2004, Available from http://www.stat.psu.edu/~jls/misoftwa.html

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: a data-analysist´s perspective. *Multivariate Behavioral Research, 33*(4), 545-571.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal, 19*(2), 249-258.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER Conquest.* Melbourne: The Australian Council for Educational Research.

.