Journal for Geometry and Graphics
Volume 2 (1998), No. 1, 71–84

Differential Transfer of Learning: Effects of Instruction in Descriptive Geometry on Spatial Test Performance

Georg Gittler, Judith Glück

Department of Psychology, University of Vienna,
Liebiggasse 5, A-1010 Wien, Austria
email: Georg.Gittler@univie.ac.at

Abstract. An important question in educational research is whether the imparting of knowledge and skills also improves pupils' intelligence. This aspect of transfer of learning is difficult to study within the framework of educational research. The longitudinal study presented here shows, based on sophisticated test materials and methods of analysis, that courses in Descriptive Geometry improve pupils' spatial ability, a primary dimension of intelligence. Also, sex differences that were clearly present at the first testing disappeared during "training" in Descriptive Geometry.

1 Introduction

Sex differences in spatial ability in favor of males have been a typical finding and an important topic of research and discussion in psychology for many years (see, e.g., [30], [6], [7], [23]). Regarding these sex differences, meta-analyses by Voyer, Voyer, and Bryden [42], and Linn and Petersen [27] have shown interesting differences between three subdomains of spatial ability named by Lohman [28]: Whereas sex differences have declined over the years for *visualization* (the ability to mentally handle and transform complex spatial configurations) and *spatial orientation* (the ability to judge how a given array would look from another perspective), they are still regularly found for *spatial relations* (the ability to mentally rotate a given object quickly). The sex differences found for this ability are usually quite large (the average difference exceeds sex differences, for example, in body height; see [38]). Note that time limitations in testing procedures may influence results because they can lead to a confounding of speed and power.

Naturally, one question of interest is whether these differences are biologically or socially determined. The fact that sex differences have largely disappeared for visualization and spatial orientation ([12]) speaks for a socio-educational basis of these differences, as well as many studies showing that spatial ability can be predicted quite well by the "environmental

ISSN 1433-8157/\$ 2.50

input" a person receives ([4], [39], [33]) and by motivational aspects ([24]). However, other studies show that biological influences are not completely negligible. Hormonal level ([38]), hemisphericity ([31]), timing of puberty ([32]), and handedness ([3], [8], [9]) have been shown to be related to spatial ability. One important aspect in this discussion is whether it is possible to make sex differences in spatial ability disappear by training female subjects properly: if this is possible, and especially if the training effect transfers to other fields of performance within spatial ability, socio-educational theories would be strongly supported.

Many studies on "spatial learning" have focused on changes in strategy or speed of problem solving that can be induced by training subjects specifically for a spatial task. For example, Embretson [11] trained subjects with the physical analogy of a paper folding task in a pretest-posttest design and found specific effects on test performance as well as (smaller) transfer effects on text editing performance. Kyllonen, Lohman and Snow [25] taught their subjects various strategies and examined various subgroups' gain in performance. Also, some studies have shown that sex differences on the Piagetian water-level task can be eliminated by training ([41]). However, as Eliot [10] states, "there are very few training or practice studies in the literature that are undertaken for more than a brief interval of time, that employ a variety of related tasks, that report transfer and retention data, and do not 'train' or provide specific practice or training for a particular criterion task" (p. 165). As mentioned above, an especially important question is whether it is possible to improve spatial ability *in general* through training.

The long-term field study presented here evaluates the transfer effects of a non-specific training — namely, Descriptive Geometry instruction at school — on performance on the *Three-Dimensional Cube Test* (3DC [20]), a spatial ability test that was developed by means of item response theory. Item response theory (see, for example, [18], [43], [26]), uses a probability function to model the relationship between manifest response behavior and a latent dimension of ability. Using certain item response models, the important assumption that a psychological test measures the same latent dimension of ability in all subjects can be statistically tested. Item response models can also be applied to the measurement of change and have some very favorable properties for this type of applications, as will be shown in the following.

In Austrian High School, pupils have to choose between several branches of upper school at an age of about 14. Some of these upper school types include Descriptive Geometry, obligatory or as a subject of choice. The advantages and disadvantages of Descriptive Geometry, as well as its usefulness in general, have been discussed quite emotionally. One of the arguments often brought up in such discussions is the question of "transfer of learning", referring to the possibilities for students to apply school knowledge or aptitudes in other contexts. Teachers, educationalists, and psychologists stress that instruction in Descriptive Geometry is beneficial to the development of spatial ability in general. However, profound empirical confirmation of the transfer of learning hypothesis has not yet been available.

2 Method and test material

A crucial point in measurement of spatial ability (as well as other factors of intelligence) is the question of *dimensionality* of the test material used. For example, sex differences in spatial ability tasks may be explained by differences in solution strategy employed by male and female subjects (the classical hypothesis is that males tend to use holistic visualization strategies, whereas females more often apply verbal-analytic strategies; see [22]). Such strategy differences may lead to differences in item difficulty: an item that is very easy for subjects applying strategy A may be rather difficult for those who use strategy B. In such cases, a different dimension of ability is measured in these two subgroups of subjects. A fair comparison of the test scores of individuals who have been using different strategies is impossible.

Item response models can handle this problem. Particularly, the *One Parameter Logistic Test Model* [36], commonly known as the "RASCH model", allows to investigate whether test items measure the same dimension of ability in different groups of subjects; items that do not fulfill this requirement can be removed from the test. In the RASCH model, item difficulty and person ability parameters have interval scale level (which ensures that parametric statistics can be used in data analysis). If the same dimension of ability is measured in, for example, men and women, the items may still be more difficult to women on the average than they are to men, but the rank order of and relations between the item difficulties must be equal for both subgroups. The development and evaluation of tests with this model takes a lot of effort; large samples of subjects have to be tested to analyze and select items. However, only the RASCH model can test the assumption that a test is actually measuring the same ability in different subpopulations. This is usually done by means of ANDERSEN's [2] likelihood ratio test. This asymptotic chi-square test examines whether the model fits the data significantly better when the item parameters are estimated separately for subgroups of subjects (H1), than when they are estimated based on the whole sample (H0). As a comparison of subjects (as well as, of course, the comparison of one subject to a population) is only legitimate ("fair") if the same ability has been tested, RASCH calibration is an important requirement for the validity of a test (cf. [43]).

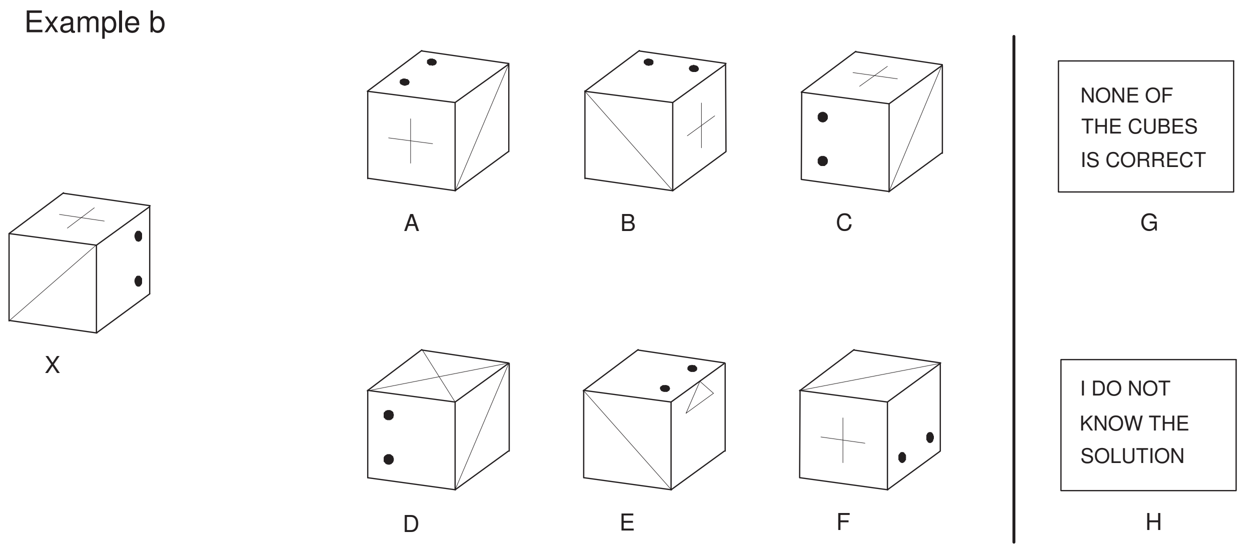

Figure 1: Practice item from the 3DC (solution: D)

The *Three-Dimensional Cube Test* (3DC [20]) is a cube comparison test. It consists of two practice items (solution given) and 18 test items; Fig. 1 displays an example. Each item shows one target cube (cube X), six alternative cubes (A to F) and two additional response categories (G, H). Every cube has six different patterns on its faces. The testees have to decide which one of the six cubes A to F (or none of them — category G) can be identical with the target cube X seen from a different (rotated) perspective. If a subject cannot find the solution, he/she is supposed to mark category H, "I do not know the solution". The test is administered without time pressure; the duration usually varies between 15 and 40 minutes. As the first item is treated as a warming-up item and not included into scoring, the test score (number of correct solutions) is computed from the remaining 17 items; therefore, scores range from 0 to 17. The test is recommended for application from 13 years onward. In many validation studies, the 3DC has been proven to measure the same latent dimension of ability in various subgroups of subjects (cf. [20], [21]).

The 3DC items can be considered to be a subsample of cube comparison tasks drawn from an explicitly defined universe of items. This universe is generated by several facets including the number of mental turns needed (one-turn vs. three-turn items; cf. Fig. 2), or the number and the type of symmetries in the visible patterns on the cube faces. In the development of the 3DC, these facets were identified both by theoretical analyses following a cognitive components approach (cf. [35]), and by empirical testing of the hypothesized item complexities using the Linear Logistic Test Model (see below).

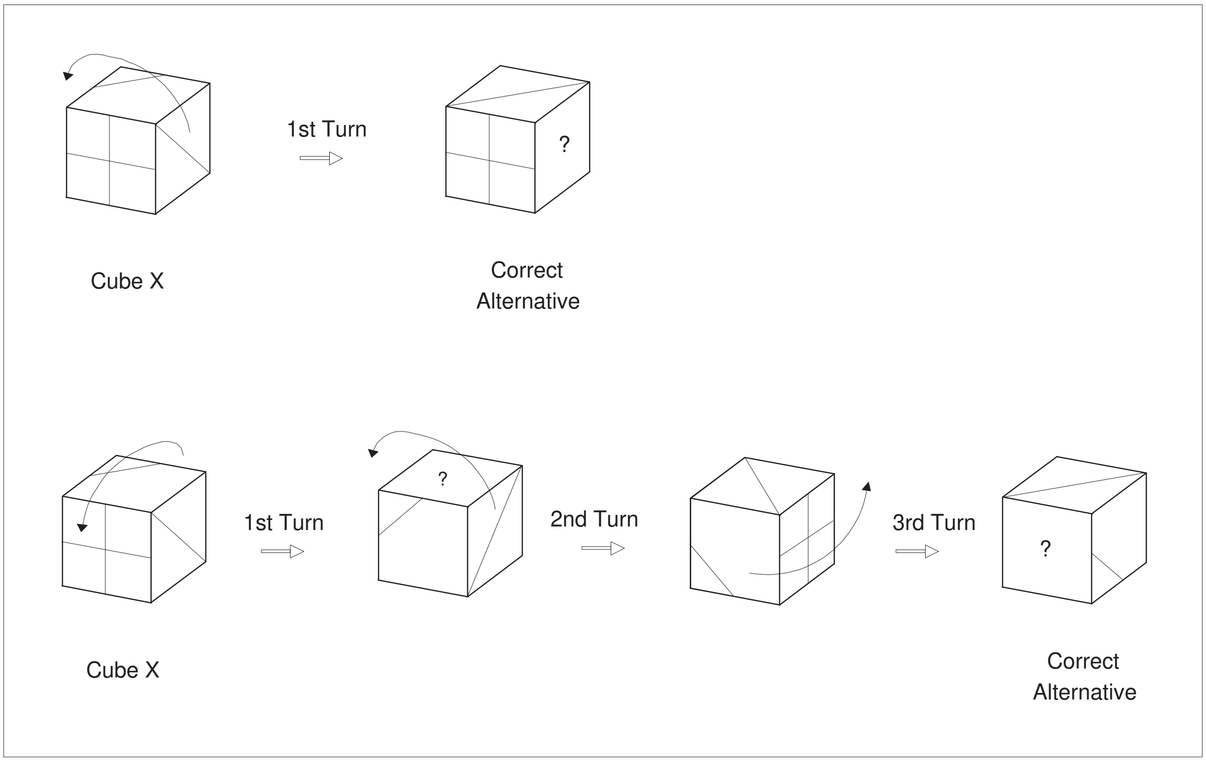

Figure 2: One-turn solution and three-turn solution items (included in the 3DC)

During the process of test development, it became evident that only the comparison of cubes which do not show the same three patterns requires a spatial solution strategy (i.e., either one or three mental turns of one cube; see Fig. 2). Items whose both cubes display the same three patterns can be solved either using a spatial (two-turn) strategy or by using a much easier non-spatial pattern rotation strategy (see Fig. 3). Thus, a test which includes two-turn-solution items does not measure unidimensionally and will not maintain RASCH-homogeneity. For this reason, only one-turn and three-turn solutions were used in the 3DC.

The response alternative G ("none") ensured that the items cannot be solved by a falsification strategy, i.e., successive exclusion of false alternatives which "leaves over" the correct alternative: as G can also be the solution, the testee must verify whether the one remaining

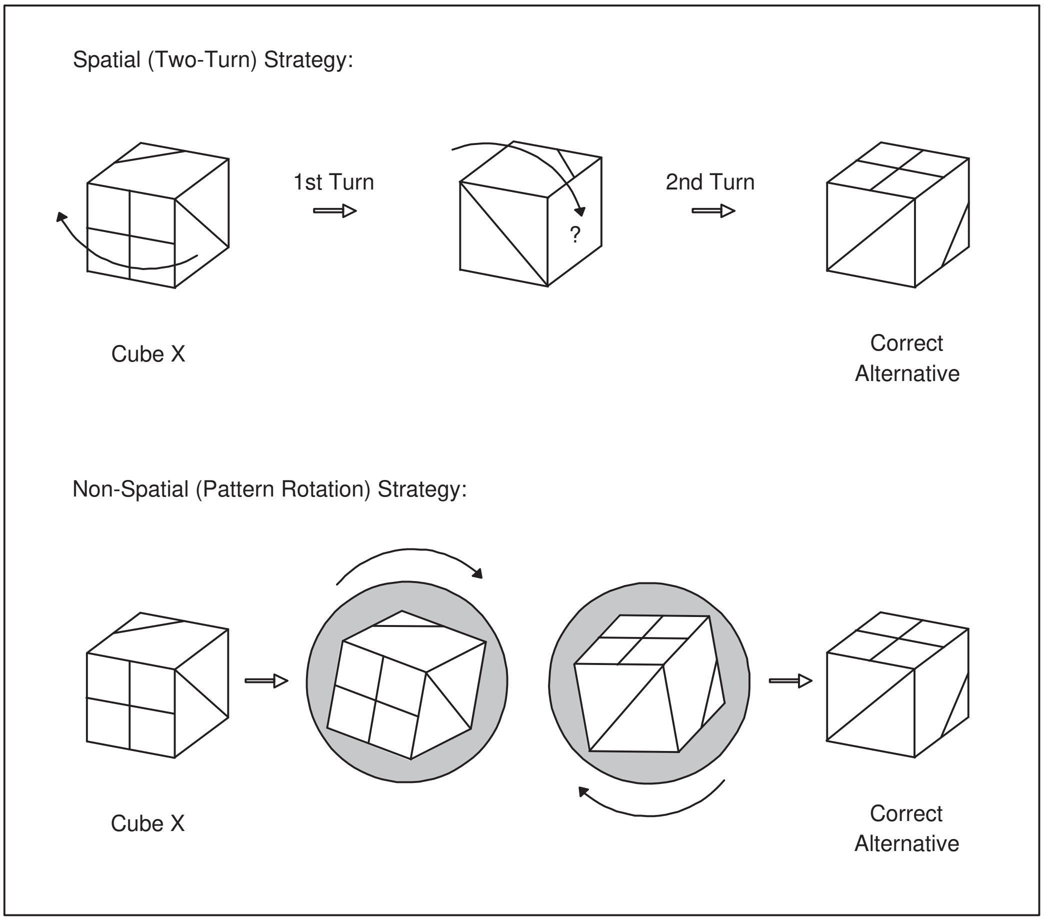

Figure 3: Two-turn solution item (excluded from the 3DC)

cube can really be the target cube seen from a different perspective. Moreover, the "none" alternative reduces guessing probability.

As the study presented here investigates the effect of a certain "treatment" (instruction in Descriptive Geometry) on subgroups of subjects which may be different in their performance *a-priori*, it is crucial that the test material must measure the same dimension of ability in all subjects. Since the 3DC has been proven to be Rasch-homogeneous, it meets this requirement.

Quantification of intraindividual change bears a couple of methodological problems, especially when raw score differences are used as a measure of change. One of these is the question of *bottom and ceiling effects* resulting from the nonlinear relationship between the raw score (number of correctly solved items; varying from 0 to the number of items given) and the latent dimension of ability. A-priori differences between subgroups (a typical problem related to quasi-experimental designs), in combination with this nonlinear relationship, can lead to differences in change which are highly artificial, though often interpreted as treatment effects. Another problem is the adequate quantification of treatment effects. The Linear Logistic Test Model for measuring change avoids these problems in part, as will be shown in the following.

3 Subjects and procedure

In the months June, July, and September of 1984 (t1; first testing), and again in February, March, and April of 1986 (t2), pupils from four Austrian provinces (Lower Austria, Upper Austria, Styria, and Vienna) were tested with the 3DC. Between the two testings, part of the students attended courses of Descriptive Geometry (DG). This "treatment" has not been designed for this particular purpose but has been part of the Austrian school system for many years; neither teachers nor their methods of teaching have been manipulated in this study. Pupils' decision to choose this subject could, of course, not be influenced, so no randomization and hence no "experimental design" was feasible. Still, to simplify matters, the group of students who attended DG courses will be referred to as the "experimental group" in the following, the other group will be named the "control group".

The interval between the two testings ranged from 19 to 22 months. Age means were 15.91 at t1 (SD = .53) and 17.62 at t2 (SD = .58).

Altogether, 690 pupils were tested, but 300 of them participated only in the first testing (t1). 76 subjects were only present at time point 2 (t2), so there remain 314 students who were actually tested twice. The following report will only refer to these paired data. Sample size had to be reduced to n=275 (see Table 1) for incomplete or mistakeable testing protocols and because testees who either solved all items or none at all in both testings are not informative for quantification of changes.

	male	female	total
experimental group	105	58	163
control group	37	75	112
total	142	133	275

Table 1: Sample sizes for all subgroups

Fig. 4 gives a picture of the mean raw scores for girls and boys at t1 and t2 which shows in advance that:

- there are no strong a-priori differences (t1) in performance level between control group and experimental group,
- male subjects start from a higher level of performance at t1, but at t2, the sex difference has disappeared for the experimental group, and
- the two experimental groups show greater improvement in performance than the control groups do.

These results are rather typical for training studies in the literature: Females often show low levels of performance at the beginning, but improve more than men and often reach the level of males after training (e.g., [29], [34], [41]). However, in most of these studies the training had been designed specifically for the type of task used. In the study presented here, training was not related to cube comparison tasks.

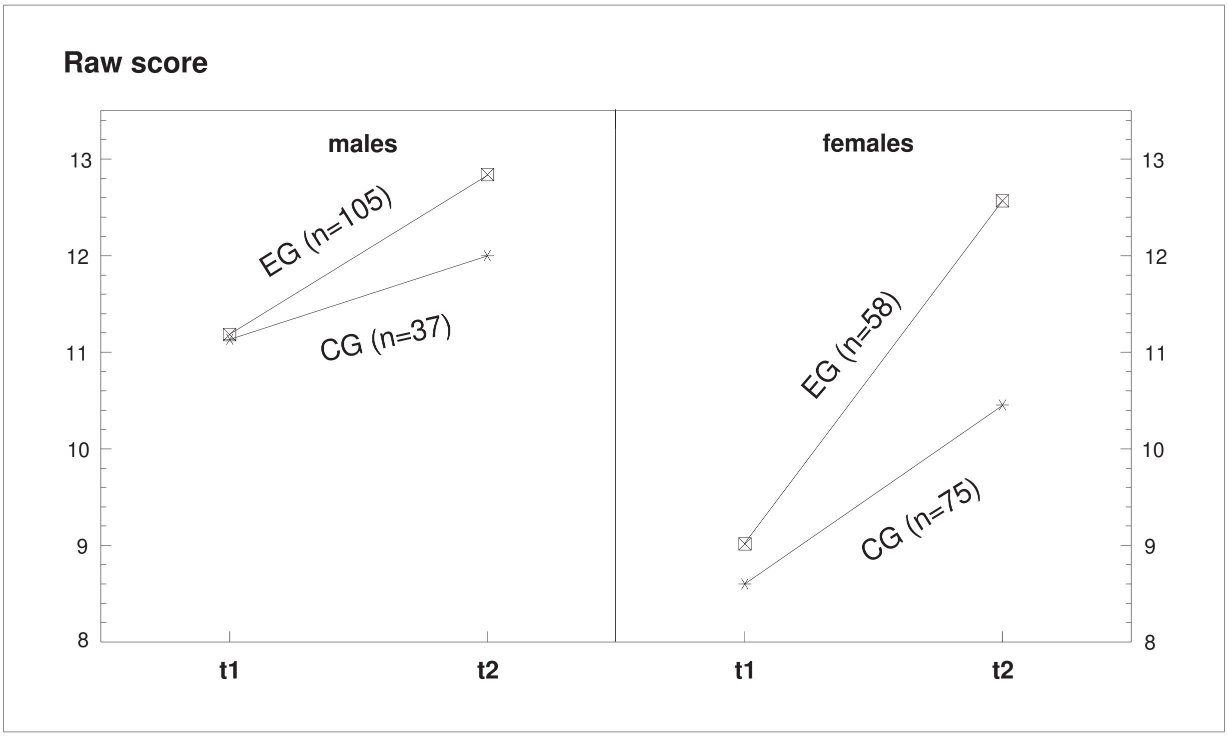

Figure 4: Raw score means for males and females at t1 and t2
(EG = experimental group, CG = control group)

4 Application of the Linear Logistic Test Model (LLTM) in measurement of change

The LLTM (Fischer [13], [14], [15], [16]) was primarily developed for analyzing the cognitive components or difficulty facets of test items. If the Rasch model holds for a set of items, one can try to "explain" their difficulties by means of cognitive operations which are considered to be relevant for solving the items. The model equation of the Rasch model is

$$P(\text{“}+\text{”} \,|\, \theta_v, \sigma_i) = \exp(\theta_v - \sigma_i) / \left[1 + \exp(\theta_v - \sigma_i)\right].$$

$P(\text{“}+\text{”} \,|\, \theta_v, \sigma_i)$ is the probability for subject v to solve item i, given the subject's ability θ_v and the item difficulty σ_i. In the LLTM, a linear restriction is imposed on the item parameters: each item parameter is formalized as the weighted sum of a number of "basic" parameters representing the cognitive operations:

$$\sigma_i = \textstyle\sum_j q_{ij}\eta_j + c\,,$$

where σ_i is the difficulty parameter of item I_i in the Rasch model,
η_j are the basic parameters of the LLTM,
q_{ij} are given weights of the basic parameters η_j (q_{ij} are usually 1, if the jth cognitive operation is assumed to be relevant for solving item I_i, and 0 if it is not), and
c is the normalization constant.

The LLTM estimates the difficulties η_j of the cognitive operations and tests the fit of a hypothetical model with empirical data by means of a likelihood ratio test. A necessary precondition for applying the LLTM is that the Rasch model holds for the data. The LLTM can also be used as a structure model for the analysis of experimental designs or treatments or for measurement of change. Items presented to the same subjects under different experimental conditions or at different time points are regarded as "structurally different", because their difficulties are expected to change with the conditions although they actually are the same items. The difference in item structure is represented in a hypothetical structure model, and the effect of each condition is estimated. In measurement of change, item difficulty parameters are estimated separately for every time point as if the items had not been the same k items at the t time points, but $k * t$ different ones. For example, the difference in difficulty between the identical items taken by the same subjects before and after a treatment is used as an estimator for the effect of the treatment. Using a control group, changes not caused by the treatment ("trend" effect) can be estimated and separated from the treatment effect.

The LLTM for measurement of change can be applied to our data as follows (see also [17], [19]). First, a quasi-saturated model (Model 0) is formalized: For the first time point, no differences in item difficulty between the four groups (experimental group /male, experimental group /female, control group /male, control group /female) are expected. Hence, for t1, the 17 item parameters of the 3DC will be estimated for all subgroups together. Regarding t2, item parameters will be estimated for each group separately — yielding 68 additional parameters. Consequently, the quasi-saturated model contains 85 parameters ("virtual items") altogether. This model allows for all kinds of group-specific changes in item difficulty between the two time points.

Next, a linear structure (Model 1) is imposed upon the item parameters from Model 0 which parameterizes a substantial hypothesis about treatment and trend effects. Differences between the four groups at t2 are explained by four effect parameters, as will be shown below. Similar to Andersen's test [2] for the Rasch model (see above), a likelihood ratio test can be used to test whether the restricted Model 1 explains the data "significantly worse" than the quasi-saturated Model 0:

$$\chi^2 = -2\ln[L(\text{data} \mid \text{Model } 1 : \sigma_1, \ldots, \sigma_{17}; \eta_1, \ldots, \eta_4)/L(\text{data} \mid \text{Model } 0 : \sigma_1, \ldots, \sigma_{85})]$$

with df = the difference in the number of parameters between Model 1 and Model 0.

Unless the LRT is significant, another more restrictive model is formalized and tested, and so on — the aim of this hierarchical procedure is to find the model that describes effects best, being theoretically adequate but as parsimonious as possible regarding the number of parameters.

As the assumption of unidimensionality of the test material must be fulfilled to apply the LLTM, the first step of analysis was to test if the Rasch model held for our data. This was done by means of Andersen's likelihood ratio test [2] for various criteria ([21]). All model tests were insignificant, implying that the Rasch model holds for the data.

In the following, the four hierarchical (nested) LLTM models that were tested against Model 0 will be described.

Model 1

Model 1 consists of only 21 parameters: the 17 item difficulty parameters from the 3DC and four effect parameters. Table 2 gives the design matrices for Model 0 and Model 1. Note that in the "item parameter" columns for Model 1 all numbers outside the diagonal line are 0, only the diagonal elements are 1 (indicating that in all four groups, the item parameters at t2 are assumed to be identical with those at t1, except for differences modeled in the effect

		Model 0	Model 1: LLTM matrix of q weights									
	3DC item no.	virtual item no.	item parameter 1^*	2	3	…	17	TREND	SEX	DG	INT
t1: all subjects	1	1	0	0	0	…	0	0	0	0	0
	2	2	0	1	0	…	0	0	0	0	0
	⋮	⋮				⋱		⋮	⋮	⋮	⋮
	17	17	0	0	0	…	1	0	0	0	0
t2: control group, male	1	18	0	0	0	…	0	1	0	0	0
	2	19	0	1	0	…	0	1	0	0	0
	⋮	⋮				⋱		⋮	⋮	⋮	⋮
	17	34	0	0	0	…	1	1	0	0	0
t2: control group, female	1	35	0	0	0	…	0	1	1	0	0
	2	36	0	1	0	…	0	1	1	0	0
	⋮	⋮				⋱		⋮	⋮	⋮	⋮
	17	51	0	0	0	…	1	1	1	0	0
t2: experim. group, male	1	52	0	0	0	…	0	1	0	1	0
	2	53	0	1	0	…	0	1	0	1	0
	⋮	⋮				⋱		⋮	⋮	⋮	⋮
	17	68	0	0	0	…	1	1	0	1	0
t2: experim. group, female	1	69	0	0	0	…	0	1	1	1	1
	2	70	0	1	0	…	0	1	1	1	1
	⋮	⋮				⋱		⋮	⋮	⋮	⋮
	17	85	0	0	0	…	1	1	1	1	1

Table 2: Design matrix **Q** for Models 0 and 1,
*… the q weights for the first item parameter are set to zero for normalization

parameters). Also, parameter estimates were normalized by setting the first item parameter zero, thus defining the unit of measurement for the parameters.
The following four effect parameters were used in Model 1:

1. TREND — the "trend" parameter models general changes between t1 and t2, that is, changes that occur in all four subgroups. This category accounts for effects of training, biological (developmental) processes, social factors, and so on. The TREND parameter is weighted zero for t1 and 1 for all items and all groups at t2.
2. SEX — this parameter refers to a sex-linked trend, representing, in addition to the general TREND parameter, changes independent of treatment which are not equal for the male and the female subgroup. The female groups have parameters of 1, men zero, which represents the hypothesis that women improve more than men (if men show more improvement, a negative effect parameter would result).
3. DG — this parameter models the effect of Descriptive Geometry instruction independent of sex. Weighting the parameter one for the experimental groups and zero for the control groups implies the assumption that subjects who have attended DG courses show greater

improvement.

4. INT (Interaction) — the last parameter models an interaction between sex and DG instruction, that is, a sex-specific treatment effect. Again, the weights imply that women profit more from DG courses than men do.

5 Results

Table 3 gives the likelihood ratio test results for all models. As shown in Table 3, the LRT of Model 1 against Model 0 is insignificant. Consequently, another, even more restrictive model (Model 2) can be tested in the same manner.

					parameter estimates η_j			
Model	k	χ^2	*df*	p-value	TREND	SEX	DG	INT
0	85	—	—	—				
1	21	63.26	64	.5027	.33	.27	.32	.29
2	20	64.96	65	.4781	.21	.45*	.49*	
3	19	68.58	66	.3901		.59*	.64*	
4a	18	142.34	67	$< .0001$		.86*		
4b	18	126.12	67	$< .0001$			.87*	

Table 3: Likelihood ratio tests (χ^2) : $k\ldots$ number of parameters in the model, $^*\ldots$ parameter significantly different from 0 ($\alpha = .05$)

Model 2

As comparatively homogeneous and small effect parameters were obtained for Model 1, none of which differ significantly from zero (according to their confidence intervals for alpha = 1%), it seemed likely that at least one more effect parameter could be eliminated by removing the corresponding column in the design matrix. For the sake of simplicity, inefficiency of the INT parameter was hypothesized in Model 2. This would imply that changes can be described using only the three other effect parameters TREND, SEX, and DG in addition to the 17 item parameters. If there is no interaction, the effects of sex and Descriptive Geometry can be considered to be additive. The insignificant likelihood ratio test (Table 3) shows that this hypothesis can be maintained.

Model 3

As, in Model 2, the global TREND parameter was not significant, it was obvious to drop this parameter next. So the null-hypothesis of Model 3 is that a sex-specific trend (SEX) and instruction in Descriptive Geometry (DG) are sufficient to explain the changes that occurred. This hypothesis was confirmed again by an insignificant likelihood ratio test (see Table 3).

Both models that result from eliminating another parameter (Models 4a and 4b) showed inadequate for our data according to significant likelihood ratio tests, leaving Model 3 as the best-fitting model with the smallest number of parameters.

As the parameterization of Model 3 showed to be an empirically valid quantification of changes, our data can be described in the most parsimonious way by 17 item parameters and the two effect parameters DG and SEX.

The actual parameter estimates for these two parameters are SEX = .59 and DG = .64, values measured on the same scale as the item parameters (which range from -.97 to +1.27). Confidence intervals ($.39 \leq \text{SEX} \leq .80\,; .45 \leq \text{DG} \leq .84\,; \alpha = 1\%$) show that both parameters are significantly different from zero. General changes (TREND) and the interaction of sex and Descriptive Geometry instruction (INT) did not prove to be relevant for describing the changes that occurred — i.e., TREND = 0 and INT=0.

To test the stability of these results, we increased the number of subjects by including data from the 3DC norm sample, which gave us a total sample size of 5351 subjects for the first time point. Hence, the 3DC item parameters were estimated from a much larger sample, which warranted a more consistent and reliable parameter estimation for the 17 item parameters at t1. The results (significances and parameter estimates) did not change at all: the parameter estimates were .59 for SEX and .63 for DG (cf. the values given in Table 3). This shows that our results cannot be explained by any a-priori properties of the subjects in our sample.

6 Discussion

As there was no sex-specific effect of instruction in Descriptive Geometry in our sample, the changes induced by this "treatment" were equal for male and female pupils. Also, there is no global trend, which seems surprising in view of THURSTONE's developmental curves [40] as well as in view of our test-retest design. The structurally relevant sex-specific trend shows that in our sample only women, in contrast to men, show a significant improvement of performance in spatial ability between the 16th and the 18th year of age, independent of the Descriptive Geometry "treatment". Naturally, this study cannot show conclusively which components are responsible for this sex-specific development.

The main point in our results is that there is a significant effect of Descriptive Geometry instruction on performance in spatial ability tasks. This clearly indicates that this school subject stimulates the development of spatial ability, inducing a non-trivial transfer of learning beyond the training of a single skill only relevant for passing school examinations.

As there is no significant interaction of sex and DG instruction, two more points must be noted here: First, men and women profit from instruction in Descriptive Geometry to the same extent (DG = .64). Second, the effects of SEX and DG are additive in the female experimental group, so this subsample shows a cumulative effect of SEX + DG = 1.23 and reaches the ability level of the male group.

This result is partly in accordance with the literature, as many studies have shown a decline of sex differences in spatial ability with higher academic education (cf. ANASTASI [1]), regardless if education is specifically technical or not. RICHARDSON [37] also found gender differences in "spatial thinking" in undergraduates, but none in postgraduates of psychology. BURNETT and LANE [5] found "training effects", that is, an improvement in spatial ability within two years, in college students even if they were majoring in psychology or other non-technical subjects. In our study, sex differences only disappear for the experimental groups. This may result from the female students in the experimental group showing the "double effect" of DG instruction *and* the sex-linked trend (which, for itself, does not suffice for the female control group to reach the level of the male controls). Perhaps it has not been possible to separate these effects clearly in earlier studies due to methodological problems.

References

[1] A. Anastasi: *Psychological testing.* 6th ed., Macmillan, New York 1988.

[2] E.B. Andersen: *A goodness of fit test for the Rasch model.* Psychometrika **38**, 123–140 (1973).

[3] M. Annett: *Spatial ability in subgroups of left- and right-handers.* British Journal of Psychology **83**, 493–515 (1992).

[4] M. Baenninger, N. Newcombe: *Environmental input to the development of sex-related differences in spatial and mathematical ability.* Learning and Individual Differences **7**, 363–379 (1995).

[5] S.A. Burnett, D.M. Lane: *Effects of academic instruction on spatial visualization.* Intelligence **4**, 233–242 (1980).

[6] P.J. Caplan, G.M. MacPherson, P. Tobin: *Do sex-related differences in spatial ability exist? — A multilevel critique with new data.* American Psychologist **40**, 786–799 (1985).

[7] P.J. Caplan, G.M. MacPherson, P. Tobin: *The magnified molehill and the misplaced focus: Sex-related differences in spatial ability revisited.* American Psychologist **41**, 1016–1018 (1986).

[8] M.B. Casey, M.M. Brabeck: *Exception to the male advantage on a spatial task: Family handedness and college major as factors identifying women who excel.* Neuropsychologia **27**, 689–696 (1989).

[9] M.B. Casey, M.M. Brabeck: *Women who excel on a spatial task: Proposed genetic and environmental factors.* Brain and Cognition **12**, 73–84 (1990).

[10] J. Eliot: *Models of psychological space.* Springer, New York 1987.

[11] S.E. Embretson: *Improving the measurement of spatial aptitude by dynamic testing.* Intelligence **11**, 333–358 (1987).

[12] A. Feingold: *Cognitive gender differences are disappearing.* American Psychologist **43**, 95–103 (1988).

[13] G.H. Fischer: *The linear logistic test model as an instrument in educational research.* Acta Psychologica **36**, 359–374 (1973).

[14] G.H. Fischer: *Probabilistic test models and their applications.* German Journal of Psychology **2**, 298–319 (1978).

[15] G.H. Fischer: *Logistic latent trait models with linear constraints.* Psychometrika **48**, 3–26 (1983).

[16] G.H. Fischer: *The linear logistic test model.* In G.H. Fischer, I.W. Molenaar (eds.): *Rasch Models — Foundations, Recent Developments, and Applications*, Springer, New York 1995, 131–155.

[17] G.H. Fischer: *Linear logistic models for change.* In G.H. Fischer, I.W. Molenaar (eds.): *Rasch Models — Foundations, Recent Developments, and Applications*, Springer, New York 1995, 157–180.

[18] G.H. Fischer, I.W. Molenaar (eds.): Rasch Models – Foundations, Recent Developments, and Applications. Springer, New York 1995.

[19] G.H. Fischer, E. Seliger: *Multidimensional linear logistic models for change.* In W.J. van der Linden, R.K. Hambleton (eds.) *Handbook of Modern Item Response Theory*, Springer, New York 1997, 323–346.

[20] G. Gittler: *Dreidimensionaler Würfeltest (3DW). Ein Rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens [Three-dimensional cube comparison test (3DC). A Rasch-scaled spatial ability test]*. Beltz Test, Weinheim 1990.

[21] G. Gittler: *Testpsychologische Aspekte der Raumvorstellungsforschung — Kritik, Lösungsansätze und empirische Befunde [Test-psychological aspects of research on spatial ability: critique, tentative solutions, and empirical findings]*. Unpublished habilitation thesis, University of Vienna 1992.

[22] D.F. Halpern: *Sex differences in cognitive abilities.* 2nd ed., Erlbaum, Hillsdale, NJ, 1992.

[23] D. Kimura: *Weibliches und männliches Gehirn [Female and male brain]*. Spektrum der Wissenschaft, November 1992, 104–113.

[24] A.G. Krasnoff, J.T. Walker, M. Howard: *Early sex-linked activities and interests related to spatial abilities.* Personality and Individual Differences **10**, 81–85 (1989).

[25] P.C. Kyllonen, D.F. Lohman, R.E. Snow: *Effects of aptitude, strategy training, and task facets on spatial test performance.* Journal of Educational Psychology **76**, 130–145 (1984).

[26] W.J. van der Linden, R.K. Hambleton (eds.): *Handbook of Modern Item Response Theory.* Springer, New York 1997.

[27] M.C. Linn, A.C. Petersen: *Emergence and characterization of sex differences in spatial ability: A meta-analysis.* Child Development **56**, 1479–1498 (1985).

[28] D.F. Lohman: *Spatial ability: A review and reanalysis of the correlational literature.* Tech. Rep. No. **8**, Aptitude Research Project, Stanford University School of Education, Stanford, CA, 1979.

[29] D.F. Lohman, P.D. Nichols: *Training spatial abilities: Effects of practice on rotation and synthesis tasks.* Learning and Individual Differences **2**, 69–95 (1990).

[30] E.E. Maccoby, C.N. Jacklin: *The psychology of sex differences.* Stanford University Press, Stanford, CA, (1974).

[31] W.F. McKeever: *Hormone and hemisphericity hypotheses regarding cognitive sex differences: Possible future explanatory power, but current empirical chaos.* Learning and Individual Differences **7**, 323–340 (1995).

[32] N. Newcombe, J.S. Dubas: *A longitudinal study of predictors of spatial ability in adolescent females.* Child Development **63**, 37–46 (1992).

[33] D.M. Olson, J. Eliot: *Relationships between experiences, processing style, and sex-related differences in performance on spatial tests.* Perceptual and Motor Skills **62**, 447–460 (1986).

[34] G. Parameswaran: *Gender differences in horizontality performance before and after training.* Journal of Genetic Psychology **156**, 105–113 (1995).

[35] J.W. Pellegrino, R. Glaser: *Cognitive correlates and components in the analysis of individual differences.* Intelligence **3**, 187–214 (1979).

[36] G. Rasch: *Probabilistic models for some intelligence and attainment tests.* The Danish Institute of Educational Research, Copenhagen 1960 (Expanded Edition, The University of Chicago Press, Chicago 1980).

[37] J.T. Richardson: *Gender differences in mental rotation.* Perceptual and Motor Skills **78**, 435–448 (1994).

[38] D.F. Sherry, E. Hampson: *Evolution and the hormonal control of sexually-dimorphic spatial abilities in humans.* Trends in Cognitive Sciences **1**, 50–56 (1997).

[39] M.L. Signorella, W. Jamison, M.H. Krupa: *Predicting spatial performance from gender stereotyping in activity preferences and self-concept.* Developmental Psychology **25**, 89–95 (1989).

[40] L.L. Thurstone: *The differential growth of mental abilities.* Research Report No. **14**, Psychometric Laboratory, University of North Carolina, Chapel Hill 1955.

[41] R. Vasta, C.E. Gaze: *Can spatial training erase the gender differences on the water-level task?* Psychology of Women Quarterly **20**, 549–567 (1996).

[42] D. Voyer, S. Voyer, M.P. Bryden: *Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables.* Psychological Bulletin **117**, 250–270 (1995).

[43] M. Wilson (ed.): *Objective Measurement: Theory into Practice.* Ablex, Norwood, NJ, 1992.

Received May 4, 1998