Detecting bimodality in the analogical reasoning performance of elementary school children

Hosenfeld, B.; van der Maas, H.L.J.; van den Boom, D.C.

DOI
10.1080/016502597385261

Publication date
1997

Published in
International Journal of Behavioral Development

INTERNATIONAL JOURNAL OF BEHAVIORAL DEVELOPMENT, 1997, *20* (3), 529–547

Detecting Bimodality in the Analogical Reasoning Performance of Elementary Schoolchildren

Bettina Hosenfeld
Leiden University, The Netherlands

Han L.J. van der Maas
University of Amsterdam, The Netherlands

Dymphna C. van den Boom
University of Amsterdam, The Netherlands

This paper reports on modelling six frequency distributions representing the analogical reasoning performance of four different samples of elementary schoolchildren. A two-component model outperformed a one-component model in all investigated data sets, discriminating accurate performers with high success probabilities and inaccurate performers with low success probabilities, whereas for two data sets a three-component model provided the best fit. In a treatment-control group data set, the treatment group comprised a larger proportion of accurate performers than the control group, whereas the success probabilities of the two latent classes were nearly identical in both groups. In a repeated-measures data set, both the success probabilities of the two latent classes and the proportion of accurate performers increased from the first to the second test session. The results provided a first indication of a transition in the development of analogical reasoning in elementary schoolchildren.

The study presented here was a re-analysis of several data sets containing analogical reasoning performance data of elementary schoolchildren, in search of a discontinuity in the development of analogical reasoning. The aim of the study was to detect bimodality, one necessary indicator of

Requests for reprints should be sent to Bettina Hosenfeld, Department of Developmental and Educational Psychology, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands.

We gratefully acknowledge the assistance of Bernice Andeweg, Nicolet Berben, Nicole Berger, Marga Boorsma, Martine van Hunksloot, Katja Kappelhof, Wendy Kettenis, Liesbeth Kooijman, Albert Lauwer, Ria Mast, Greet Moerland, Sascha den Outer, Digna Schijffelen, Jacqueline van der Veen, Ineke Vermeijs, Esther Westenberg, and Oglaya Woerdings in the data collection.

developmental discontinuities (van der Maas & Molenaar, 1992). This indicator can be investigated in simple cross-sectional data sets, whereas the examination of other indicators, as for example sudden jumps in a growth curve, requires longitudinal data. Bimodality has already been found in performance data from task domains like conservation, classification, and the understanding of horizontality and verticality (Thomas, 1989; Thomas & Lohaus, 1993; Thomas & Turner, 1991; van der Maas, Walma van der Molen, & Molenaar, 1993). In analogical reasoning performance, however, bimodality has not yet been detected.

Analogical reasoning (i.e. establishing a correspondence between sets of relations), is considered to be a skill essential for knowledge acquisition and a fundamental component of intelligence (Goswami, 1991). Several cross-sectional studies have consistently reported large age differences in children's analogical reasoning performance (Alexander, Willson, White, & Fuqua, 1987; Goldman, Pellegrino, Parseghian, & Sallis, 1982; Goswami, 1991; Sternberg & Nigro, 1980; Sternberg & Rifkin, 1979; Wagner, as cited in Case, 1985). In general, the probability of solving almost any analogy problem increases with age. Accurate age boundaries for mastering the analogical principle reported by the different authors, however, vary with the type of task administered.

Analogical reasoning seems to be absent in infants and to emerge for the first time during childhood. Some evidence for this hypothesis was provided by several authors who compared the solution processes of children with low and high analogy performance (Alexander et al., 1987; Goldman et al., 1982; Sternberg & Rifkin, 1979; cf. Crisafi & Brown, 1986; Goswami, 1991). In general, children with low analogy performance use free associations, which lead to correct analogy solutions only coincidentally. This was most convincingly demonstrated by Alexander et al. (1987). Low-performing individuals solving geometric analogies verbalised predominantly lower-order relations, that is, relations between the third term of the analogy and the solution. In addition, their erroneous responses conformed to a matching-by-similarity rule on the dimensions colour, shape, and size. In contrast, high-performing analogical reasoners verbalised mainly higher-order relations (i.e. relations between the two pairs of the analogy). Hence, developmental changes in analogical reasoning might be described as the replacement of the similarity rule by a more sophisticated rule, which includes considering the constraint of parallel relations.

If there are two qualitatively different solution rules for analogy problems leading to different success probabilities, this two-rule structure should be reflected in a bimodal frequency distribution of analogy performance data. Bimodality in a frequency distribution results when there are two different sets of behavioural values with high probabilities of occurring, the modes, whereas there is also a set of behavioural values with a low probability of occurring, the inaccessible region. Under each of the two modes, a class of subjects can be found. These two classes represent the accurate performers, who obviously master the principle of the task, and the inaccurate performers, who apply rules that are inadequate for the solution of the task (Thomas & Turner, 1991).

Although bimodality has been detected in performance data from several Piagetian tasks (Thomas, 1989; Thomas & Lohaus, 1993; Thomas & Turner, 1991; van der Maas et al., 1993), in analogical reasoning performance data it has not yet been demonstrated. On the contrary, the unimodal distribution of verbal analogy performance data has even been presented twice as a contrasting example to bimodality in spatial task performance (Lohaus, Kessler, Thomas, & Gediga, 1994; Thomas & Lohaus, 1993). Nevertheless, we expected to find bimodality in analogical reasoning performance, because we presumed the two qualitatively different solution rules reported by other researchers (e.g. Alexander et al., 1987). In general, switches between solution rules are expected to be discontinuous, which brings about bimodality (van der Maas, in press). The two solution rules in analogical reasoning, matching by similarity and considering parallel relations, were expected to lead to low and high success probabilities, respectively.

The phenomenon of bimodality does not need to occur in the performance of any age group, but may be an indicator of a local transition in the development of analogical reasoning. In such a case, unimodality would sufficiently describe analogy performance before and after the shift from one rule to the other, whereas during the rule shift, two distinct modes would emerge (van der Maas & Molenaar, 1992).

Evidence of multimodality, or bimodality as a special case, can be found when mixture distribution analyses are applied to frequency data, based on the assumption that frequency distributions are mixtures composed of a finite number of components (Everitt & Hand, 1981; Titterington, Smith, & Makov, 1985). If the variable that splits the observations into groups is known, each of the components can be described separately. If the splitting variable, however, is unknown, only the combined distribution can be studied. The first purpose of modelling mixture distributions is to determine whether or not a data set represents a mixture of several distributions. Furthermore, models with different numbers of parameters can be specified and their fit to the observations can be compared. The parameters to be estimated are the number of components and the proportion of observations under each component. For example, a simple case is a two-component mixture distribution of normal components:

$$F = \pi\, N(\mu_1, \sigma_1) + (1 - \pi)\, N(\mu_2, \sigma_2),$$

where π and $(1 - \pi)$ specify the proportions of the normally distributed components, which are each determined by the means (μ_1, μ_2) and standard deviations (σ_1, σ_2). It is also possible to assume other distributions or more components. The following, for example, is a three-component mixture distribution of binomial components:

$$F = \pi_1 \, B(\theta_1) + \pi_2 \, B(\theta_2) + (1 - \pi_1 - \pi_2) \, B(\theta_3),$$

where π_1, π_2, and $(1 - \pi_1 - \pi_2)$ specify the proportions of the binomial components, which are determined by the probabilities of a correct response ($\theta_1, \theta_2, \theta_3$). Furthermore, the application of mixture models is not restricted to investigations concerning performance data, but is always useful, if hypotheses about qualitative categories are examined, as, for example, in temperament research (Stern, Arcus, Kagan, Rubin, & Snidman, 1995).

Each data set can be fitted by a model, if the number of components is large enough. In order to find a reasonably parsimonious model, the number of components should be chosen on the basis of theoretical reasons. Moreover, the fit of a model with several latent classes does not necessarily imply bi- or multimodality, as was pointed out by Thomas and Lohaus (1993). A mixture of two normal distributions with equal means and different variances, for example, results in a unimodal distribution. Therefore, in addition to the mixture analyses, the model plot should be visually examined in search of bimodality. Only if the model plot displays two clearly separable peaks, it can be concluded that bimodality is present.

In the study reported here, we examined the frequency distributions of analogy performance from four different samples in search of bimodality as a possible indicator of a rule shift. We chose the binomial model, because we assumed that the test scores were binomially distributed. Two of the data sets contained cross-sectional data, the third one consisted of data from a treatment-control group design, and the fourth consisted of data from a repeated-measures design.

METHOD

Subjects

Four different samples of kindergarten and elementary schoolchildren, ranging in age between 5 and 11 years were drawn from several schools. Sample 1 comprised the total population of grades 2, 4, and 6 in three different schools (222 children). Sample 2 consisted of 101 kindergarten children and children from grades 1 and 2 in five different schools. In this sample, from each grade of each school, eight children were randomly selected. Nineteen incomplete records had to be discarded. Sample 3 consisted of 119 children from grades 2, 3, and 4 in five different schools; four girls and four boys were randomly selected from each grade. The record of one child was discarded, because it was incomplete. Sample 4 consisted of 36 children from grades 2 and 4 in two different schools. Five girls and five boys from each grade were randomly selected and tested twice. Because 4 of the 40 children completed only one of the two test sessions, their records were discarded. Table 1 displays the number of boys and girls for each grade in each sample and the mean age of the children in each group.

Material

The geometric analogies test (Hosenfeld, van den Boom, & Resing, submitted) consisted of 36 open-ended items, which were designed out of six basic geometric shapes and five transformations by means of a facet design (see Fig. 1). The level of difficulty of each item could be predicted satisfactorily by the number of elements and the number of transformations the item contained. A Mokken scale analysis (Mokken, 1971) revealed monotone homogeneity and double monotonicity. Hence, the items and the subjects could be ordered reliably on a common dimension. Furthermore, genuine parallel test items were constructed by means of the same construction rules as were the original items. The level of difficulty of the original and the parallel test items corresponded highly. The same was true for the interrelations of both test versions with external variables.

TABLE 1
Mean Age in Months of the Subjects by Sample and Grade

				Sex		*Age*	
Sample	*N*	*Grade*	*n*	*Girls*	*Boys*	*M*	*SD*
1	222	2	85	43	42	92	4.5
		4	67	27	40	119	6.1
		6	70	36	34	141	5.5
2	101	0[a]	38	20	18	70	3.6
		1	29	14	15	84	3.9
		2	34	15	19	97	6.7
3a	59	2	20	10	10	93	4.7
		3	19	10	9	104	4.5
		4	20	10	10	116	4.2
3b	60	2	20	10	10	93	3.9
		3	20	10	10	103	4.2
		4	20	10	10	115	3.9
4	36	2	18	9	9	96	3.6
		4	18	10	8	122	6.5

Note. Sample 3 was divided into a control group (3a) and a treatment group (3b).
[a] Grade 0 = kindergarten class.

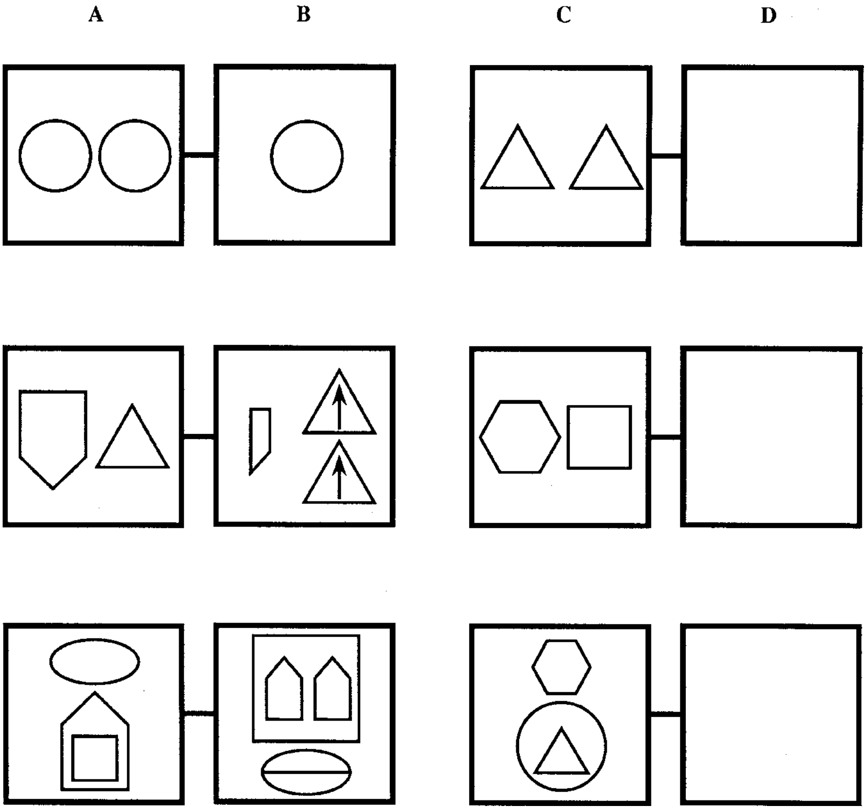

FIG. 1. Three examples of geometric analogies of different levels of item difficulty.

According to the taxonomy by Halford (1993), the geometric analogies administered can be characterised as analogies of the missing-element missing-relation type, which require relational mapping. For example, in order to solve the geometric analogy “small triangle : large triangle :: small circle : ?”, the relation from A to B (“small” changes to “large”) has to be discovered and to be applied to C (“small circle”) in order to find the missing element D (“large circle”). Difficult items, as demonstrated in Fig. 1, contained more information than easy items, but could be solved in small steps, so that the amount of information that had to be processed in parallel remained small.

Procedure

The testing procedures that were followed for each of the four samples varied in the number of items administered and the kind of instruction given. In Sample 1, the complete test with 36 items was administered to whole classes with a standard instruction including the presentation of three examples of correct task solutions. Furthermore, the six basic geometric shapes were named by the test instructor and then copied by the children. In Samples 2, 3, and 4, a selection of 20, 15, and 14 items, respectively, was administered individually with the same standard instruction, but this time at the beginning of and twice during the test session some practice items were provided instead of the example items. Together with the practice items the standard instruction was repeated in order to remind the children of what they were supposed to do, but no feedback or training was provided. Moreover, in the individual test session, the children themselves named the six geometric shapes and then copied them.

Whereas the data from Samples 1 and 2 were cross-sectional, the data from Samples 3 and 4 were collected within an experimental design. The subjects of Sample 3 were matched by grade, sex, and pre-test score on the Standard Progressive Matrices (Raven, 1958) and were assigned to either an experimental group, which solved the test items following a thinking-aloud procedure, or a control group, which solved the items according to the standard procedure described earlier. In Sample 4, every subject was tested twice. One week after the first test session, a parallel version of the original test items was administered. In all samples, the test items were presented in an open-ended format, that is, the children were required to draw their item solutions in an open space in the test booklets.

Scoring

For each item one standard solution was defined. The composed item solutions were compared with these standard solutions and were scored as either correct or incorrect. The number of items correctly solved constituted the variable test score, the binomial variable in the subsequent analyses.

RESULTS

The results will be presented in three sections. First, a description of the six data sets, which were analysed separately, will be given. Second, the key parameters of the binomial mixture analyses, some evaluation statistics, and the decision rule we used for interpreting our results will be explained. Third, the plots of the frequency distributions for the six data sets will be presented together with the results of the binomial mixture analyses.

Before running the analyses, we formed six separate data sets. For Samples 1 and 2, the analogy performance of the total samples was analysed; for Sample 3, the analogy solutions of the treatment group and the control group were analysed separately; for Sample 4, analogy performance was analysed for the two test sessions separately. In order to create homogeneous item sets for subsequent use in the binomial mixture analyses, ten items were selected from the data sets of Samples 1 and 2. These ten items were also part of the item set administered to Sample 3. Five of the ten items overlapped with the item set administered to Sample 4. The level of item difficulty ranged from .37 to .83 for Sample 1; from .08 to .54 for Sample 2; from .15 to 1.00 for Sample 3, and from .42 to .96 for Sample 4.

In order to assess the fit of several models and to compare models with different numbers of modes, we used mixture techniques (Everitt & Hand, 1981), which deliver a set of statistics. We applied binomial mixtures, because the binomial model is the simplest one that can be applied to test scores, which can be assumed to be binomially distributed. We also fitted mixture models of normal distributions, but found the same trends in the results as for the binomial mixtures. Hence, only the results of the binomial mixture models will be presented.

Because we assumed two different rules for analogy solution in elementary schoolchildren, the two-component model was the focus of model testing. For each data set, a one-, a two-, and a three-component binomial mixture model was fitted to the frequency distributions of the test scores and the success probabilities of the latent classes (θ_1, θ_2, and θ_3) and the proportions of subjects belonging to each class (π_1, π_2, and π_3) for each model were estimated. Next, the respective goodness-of-fit indices of the three models were compared. The two-component model was selected, when the evaluation statistics indicated that it outperformed the one-component model and when the three-component model did not enhance the model fit of the two-component model substantially. Overall, we considered our hypothesis of bimodality rejected, if in the majority of the data sets the one- or the three-component model fitted the observations convincingly. If, however, the three-component model fitted best, while in the model plots only two peaks were discernable, we nevertheless retained the two-component model.

Three evaluation statistics were computed for each model in order to find the most parsimonious model with the best fit (Thomas, 1989; Thomas & Turner, 1991). First, Akaike's information criterion (AIC), discussed by Thomas and Lohaus (1993), served as a selection criterion within the family of models fitted to the same data set. It is defined (Thomas & Turner, 1991, p. 182) as: "minus two times the loglikelihood function plus a 'penalty factor' equal to twice the number of parameters estimated from the data." Hence, the smallest AIC points out the best fitting model in connection with parameter parsimony. Second, the Pearson χ^2 statistic is an indication of the goodness-of-fit of the model to the observations. Third, the proportion of variance accounted for (VAF) by the model intuitively seems to be the clearest indicator of a model fit. It can be used for the comparison of different model solutions for the same data set as well as for the comparison of the model solutions of different data sets. It ignores, however, the issue of parameter parsimony.

When interpreting the results of the mixture analyses, we took into account that the assumptions of the mixture analyses are rather strict. Not only has the number of components to be fitted, but also the form of the distributions has to be specified. Because our hypothesis about bimodality did not include any assumptions about the form of the distribution, but only about the number of components, we decided to interpret the indicators of goodness-of-fit liberally and to compare these interpretations with the plots of the observed test scores and the model plots.

Table 2 displays the means and the standard deviations of the test scores as well as the results of the model estimations for each of the six data sets. For Sample 1, the estimates of the success probabilities were: $\theta_1 = .62$; $\theta_1 = .11$, and $\theta_2 = .80$; $\theta_1 = .07$, $\theta_2 = .67$, and $\theta_3 = .94$, for the one-, the two-, and the three-component model, respectively, whereas the corresponding proportions of subjects in the latent classes were: $\pi_1 = 1.00$; $\pi_1 = .26$, and $\pi_2 = .74$; $\pi_1 = .22$, $\pi_2 = .48$, and $\pi_3 = .30$. When we compared the AICs, the χ^2 statistics, and the VAFs of the one- and two-component model (see Table 2), the two-component model represented a great improvement over the one-component model on all three fit indices, whereas the improvement from the two- to the three-component model was definitely smaller: The AIC decreased from 1709 through 1064 to 1018, χ^2 decreased from 67163 ($df = 9$) through 91 ($df = 7$) to 29 ($df = 5$), and the variance accounted for increased from 20% through 91% to 97%.

Furthermore, in the frequency plot (see Fig. 2), bimodality was visible. A large number of subjects solved either none of the items or 6 or more of the 10 items. There were two discernable peaks in the plot of the two-component model, whereas the second and the third peak in the three-component model formed one unit. Moreover, the inaccessible region between the first and the second mode remained almost stable in both the two- and the three-component solutions.

For Sample 2, the three-component model was found to be linked with the lowest AIC, the smallest χ^2, and the largest amount of explained variance, whereas the increase in explained variance from the two- to the three-component model was small. In this case, however, the χ^2 statistic indicated a sufficient model fit of the three-component model. The frequency plot of the observed test scores (see Fig. 3) showed that a large part of the sample did not solve more than 1 of the 10 items, a smaller group solved about 4 items, and even less subjects solved 6 and more items. The plot of the two models on the observed data confirmed that the three-component model fitted the observations more convincingly than the two-component model (see Fig. 3). The model plot of the two-component model, however, displayed the two peaks we expected and an almost empty region in between.

TABLE 2
Binomial Mixture Model Estimates with 1, 2, and 3 Components for the 6 Data Sets from Samples 1–4

Sample	*N*	i^a	*M*	*SD*	θ_1	θ_2	θ_3	π_1	π_2	π_3	VAF	AIC	χ^2	*df*	*P*
1	222	10	6.18	3.4	.62			1.00			.20	1709	67163	9	.00
					.11	.80		.26	.74		.91	1064	91	7	.00
					.07	.67	.94	.22	.48	30	.97	1018	29	5	.00
2	101	10	2.44	2.82	.24			1.00			.23	626	12139	9	.00
					.07	.58		.66	.34		.88	432	40	7	.00
					.03	.31	.78	.50	.35	.15	.99	409	5	5	.57
3a	59	15	9.08	2.84	.61			1.00			.44	317	111	14	.00
(Control					.48	.76		.54	.46		.95	293	10	12	.66
group)					.34	.54	.78	.11	.53	.36	1.00	296	7	10	.69
3b	60	15	10.22	2.61	.68			1.00			.48	304	76	14	.00
(Treatment					.43	.75		.21	.79		.99	284	10	12	.59
group)					.43	.74	.99	.20	.77	.03	1.00	286	7	10	.74
4	36	14	9.56	4.18	.68			1.00			.17	305	342427	13	.00
(Test 1)					.19	.82		.22	.78		.87	195	31	11	.00
					.07	.59	.89	.14	.31	.55	.97	183	6	9	.72
4	36	14	11.25	3.04	.80			1.00			.22	216	152157	13	.00
(Test 2)					.32	.87		.13	.87		.90	161	11	11	.55
					.30	.83	.98	.12	.65	.23	.97	161	5	9	.85

Note: θ_1, θ_2, and θ_3 are the estimations of the success probabilities of each component; π_1, π_2, and π_3 are the proportions of the sample under each component; VAF is the variance accounted for by the model, and AIC is the Akaike's information criterion.

[a] *i* is the number of items administered to the sample.

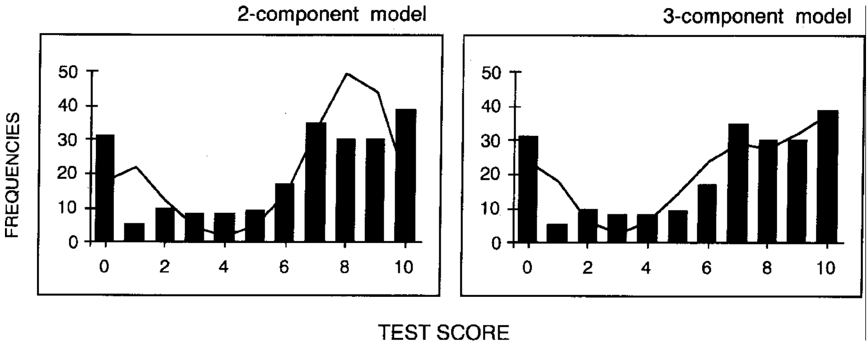

FIG. 2. Frequency distributions of the observed test score for Sample 1 together with the estimations for the 2- and 3-component model.

For Sample 3, we first compared the test performance of the treatment group with that of the control group. A Mann-Whitney U-test revealed a significant treatment effect, $z = 2.21$, $P < .05$. Children following the thinking-aloud procedure received a higher test score ($M = 10.22$, rank $= 66.89$) than children following the standard procedure ($M = 9.08$, rank $= 52.99$). Next, we obtained the model estimates for the treatment and the control group separately. In both groups, the AIC indicated the best fit for the model with two components. The nonsignificant χ^2-statistics and the high VAF-indices confirmed this conclusion. Hence, bimodality provided the most parsimonious description of the analogy performance distributions of both the treatment and the control group. The estimates for the success probabilities of the two groups corresponded highly, $\theta_1 = .48$ and $.43$, and $\theta_2 = .76$ and $.75$, respectively, whereas the proportion of analogical reasoners was significantly larger in the treatment group than in the control group, $\pi_2 = .46$ and $\pi_2 = .79$, $z = 3.71$, $P < .001$. Obviously, the experimental procedure

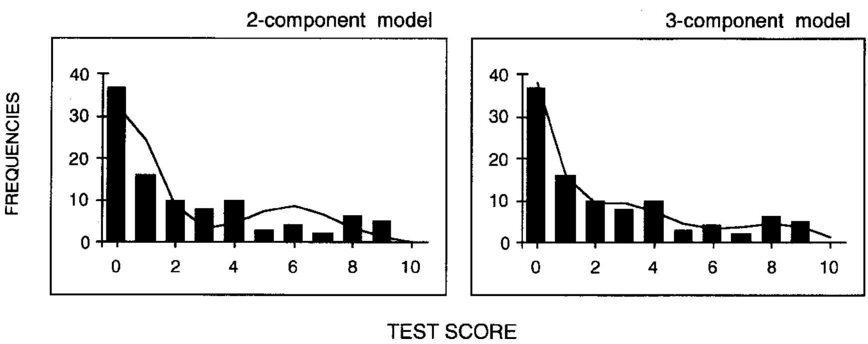

FIG. 3. Frequency distributions of the observed test score for Sample 2 together with the estimations for the 2- and 3-component model.

did not influence the performance of the whole treatment group, but increased the probability of belonging to the high-performing class for about one-third of the subjects.

The frequency plots of the analogical reasoning performance of both groups in Sample 3 (see Fig. 4), display two separate peaks in the observed data, but not in the curve representing the estimated model. Because the number of subjects in the sample was small, the plot of the estimated model is less clear than the plot of the observed data. Furthermore, bimodality is more clearly visible in the frequency distribution of the treatment group than in that of the control group. This observation points to the phenomenon of divergence, which means that the influence of a splitting control variable, in our case the thinking-aloud procedure, can amplify bimodality. Moreover, the estimated modes were equivalent in both groups, whereas the

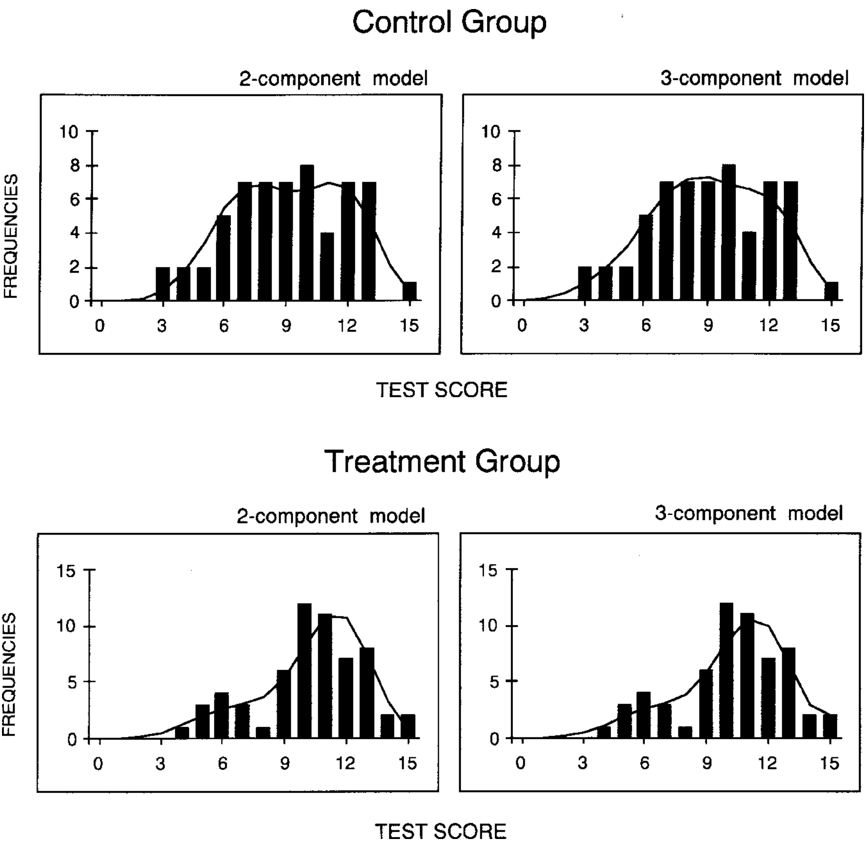

FIG. 4. Frequency distributions of the observed test score for the control and the treatment group of Sample 3 together with the estimations for the 2- and 3-component model.

proportion of analogical reasoners was larger in the treatment group than in the control group. Probably, the transitional subjects, who cannot be identified in a cross-sectional design, belong to the group of nonanalogical reasoners in the control condition, whereas they belong to the group of analogical reasoners in the treatment condition.

Furthermore, the verbal data obtained from the treatment group attested to the validity of the two-component structure. After estimating the posterior probabilities of belonging to the first or the second mode of the frequency distribution as suggested by Thomas and Lohaus (1993), we determined a cut-off point of eight items correct and defined each subject as either a nonanalogical or an analogical reasoner. Following a categorisation procedure similar to the one developed by Alexander et al. (1987), we obtained the number of spontaneously uttered lower-order and higher-order relations for each subject. A lower-order relation was defined as a relation between the third term of the analogy and the solution, whereas a higher-order relation implied both a relation between the first and the second term and a relation between the third term and the solution. The cross-tabulation revealed a significant association between type of reasoning and category of verbalisation. The 12 nonanalogical reasoners together uttered 29 (16% of 180 instances) lower-order and 30 (17%) higher-order relations; the 46 analogical reasoners, on the other hand, uttered 54 (8% of 690) lower-order and 210 (30%) higher-order relations [$\chi^2(2, N = 870) = 20.77, P < .001$].

A similar pattern as in Sample 3 was found for the two measurement occasions in Sample 4. A Wilcoxon test revealed a significant repeated-measurement effect. Overall, the analogy performance increased from the first ($M = 9.56$, rank $= 4.50$) to the second ($M = 11.25$, rank $= 17.30$) test session, $z = 4.77$, $P < .001$. Moreover, in neither of the two test sessions did a one-component model fit the frequency distributions satisfactorily. For Test Session 1, the three-component model provided the best fit, whereas for Test Session 2, the two-component model turned out to be the best fitting model. The choice between the two- and the three-component model, however, seemed arbitrary, because the AICs and the VAFs of the two models hardly differed. When comparing the parameter estimates of the two-component models for the two test sessions, we found that the success probability of both classes changed slightly, $\theta = .19$ and $.32$, $z = 1.56$, $P = .06$ and $\theta_2 = .82$ and $.87$, $z = 1.85$, $P < .05$. Furthermore, the proportion of accurate performers also increased by 8%, $z = 2.31$, $P < .05$, from the first to the second test session. Apparently, repeated testing caused two effects, an improvement in the performance of the whole sample and a change in class membership of a few subjects. Again, in Sample 4, bimodality, which was obtained by the model estimations, was more clearly visible in the data than in the model (see Fig. 5). Most probably, the small number of subjects in the sample accounted for this phenomenon. In accordance with our expectation that repeated testing within one week should not induce a large number of subjects to move from one mode to the other, the proportions of subjects in the two classes hardly changed.

In sum, the results provided some support for our hypothesis about bimodality in the analogical reasoning performance of elementary school-children. In all six data sets, the two-component mixture model outperformed the one-component model. A good fit of the two-component model, however, was demonstrated in only three of the six data sets. For one of the remaining data sets, none of the three models tested could be fitted sufficiently, for the remaining two, the three-component model provided the most convincing description of the data.

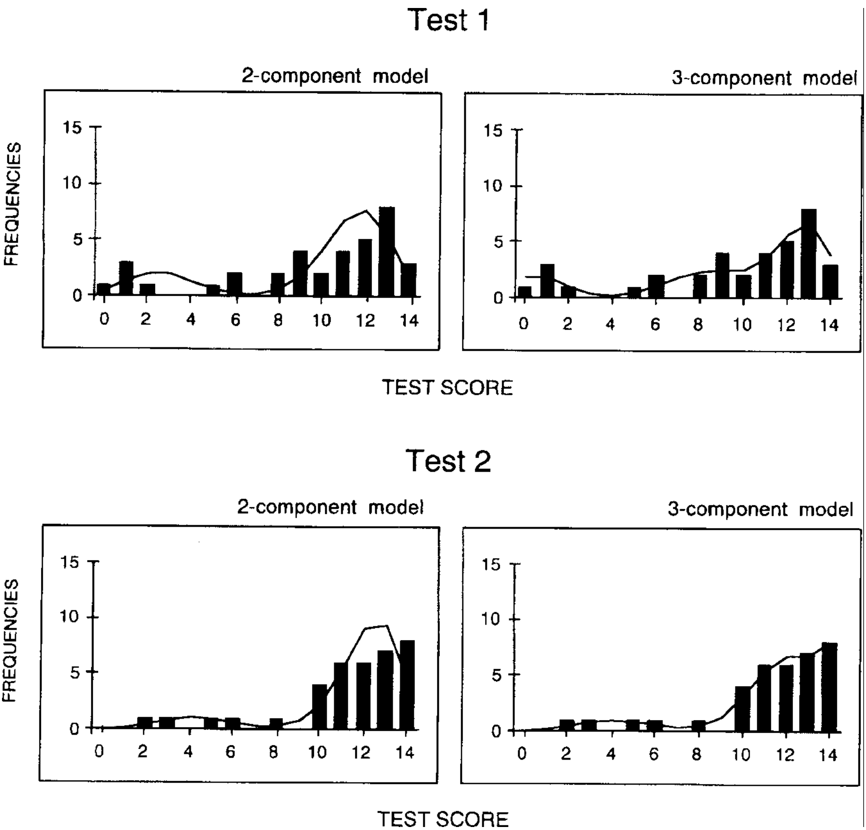

FIG. 5. Frequency distributions of the observed test score for Sample 4 at two measurement occasions, together with the estimations for the 2- and 3-component model.

DISCUSSION

This paper reported on a re-analysis of six data sets containing the analogy performance of children aged 5 through 11 years. Although two qualitatively different, age-related solution rules (i.e. matching by similarity and considering parallel relations), have been observed before in analogical reasoning, in none of these inquiries (Alexander et al., 1987; Goldman et al., 1982; Sternberg & Rifkin, 1979) has bimodality in the frequency distributions of the test scores been examined. If formal evidence of a discontinuity in the development of analogical reasoning is sought, the detection of bimodality is a necessary first step (van der Maas et al., 1993).

Our results suggest that the two-component model outperformed the one-component model in all six data sets. Therefore, the hypothesis of unimodality was rejected. Moreover, the three-component model outperformed the two-component model in only two data sets, whereas in one data set more than three components were necessary to fit the distribution. Because in the first place we were looking for deviations of the distributions from the unimodal model, we interpreted the three-component solution as a confirmation of the rule-shift hypothesis as well. Overall, it can be concluded that in each of the data sets at least two distinct latent classes, analogical reasoners with low and high performance, respectively, were identified. Additional support for the two-class solution can be derived from the literature on analogical reasoning. Alexander et al. (1987) displayed the freqency plots of the analogical reasoning performance of 4- and 5-year-old children collected in two studies. Both distributions show two clearly discernible peaks, indicating bimodality. Moreover, both samples were subdivided at the chance level for multiple-choice performance into two groups, nonanalogical and analogical reasoners, whereas the cut-off point lay precisely between the two peaks of the distribution. Most probably, a re-analysis of the data collected by Alexander et al. (1987), by means of a mixture decomposition technique, would reveal a good fit of the two-component model and confirm the division of the sample.

Why did the three-component model provide the best fit for the data of Sample 2 and why did none of the three models apply to the data of Sample 1? Possibly, the age of the subjects was not evenly distributed in these samples, so that age effects may have masked the developmental discontinuity (Fischer, Pipp, & Bullock, 1984). Furthermore, because both samples consisted of children from three different grades, grade also may have accounted for the partial success of the three-component model. In future studies, the age and grade distributions of the subjects should be strictly controlled. Another possible interpretation of the relative success of the three-component model is that there were indeed three different performance classes, nonanalogical reasoners and two groups of analogical reasoners, one with low and one with high information-processing capacities. In contrast to the nonanalogical reasoners, both groups of analogical reasoners master the analogical principle. The analogical reasoners with low information-processing capacity, however, are not able to solve the more complex analogies completely, which require the successive processing of several transformations. The difference between the nonanalogical reasoners and the analogical reasoners refers to the qualitative shift we are looking for, whereas the difference between the two classes of analogical reasoners might refer to a quantitative shift.

There were three more findings in our study pointing to a discontinuity in the development of analogical reasoning, which went beyond the detection of bimodality. First, the two-class structure of the subsample following a thinking-aloud procedure was confirmed by the verbal utterances of the children. Bimodality in this data set, at least, was related to the two postulated solution rules. An alternative explanation that different complexity levels in the item sets might be responsible for the two modes in the distribution can, therefore, be excluded. Second, the thinking-aloud procedure amplified bimodality. Hence, thinking-aloud seems to be related to one of the variables that may explain the rule shift in analogical reasoning, for instance, increasing monitoring skills (Goswami, 1991). Another possible interpretation is that the thinking-aloud procedure might have increased consistent responding in both inaccurate and accurate performers (Russo, Johnson, & Stephens, 1989). Further experiments are needed to test these hypotheses. Third, in both the treatment-control group design and the repeated-measures design, group differences and measurement effects were largely explained by the difference in the proportion of accurate performers. Bimodality, therefore, was not simply an incidental anomaly, but a meaningful characteristic of the frequency distributions. Repeated testing, however, also increased the success probabilities of both latent classes, probably because the subjects became familiar with the test material, which enhanced encoding processes. Elaborate encoding of the features of an analogy and relatively fast processing of the attribute-comparison processes had already been demonstrated to be a successful strategy in adults (Sternberg, 1977).

The proportion of children belonging to the high-performing class was larger in the older samples than in the youngest sample including kindergarten children. This phenomenon might be truly developmental in nature. It is likely that children move from a state of nonanalogical reasoning to a state of analogical reasoning because they acquire a new solution rule. In contrast to Thomas and Lohaus (1993), who investigated spatial abilities and found bimodality in the performance of both children and adults, we believe that bimodality in analogy performance is not a universal characteristic at every age, but a peculiarity of the transition from associative to analogical reasoning. If the analogy performance of more extreme age groups were examined, we would expect simple unimodal models to fit the data. For toddlers, who are not able to observe the constraint of parallel relations, a one-component model with success probabilities near zero should suffice to describe the frequency distribution of analogy performance. For adolescents, who apply the principle of analogical reasoning almost faultlessly, a one-component model with success probabilities near one should be adequate.

Several authors (e.g. Brown, 1989; Crisafi & Brown, 1986; Goswami, 1995; Goswami & Brown, 1989) would disagree with the assumption that toddlers are not able to solve open-ended analogies, when there is no possibility to rely on association. We admit that there is research demonstrating that 3- and 4-year-old children master analogies, if sufficient domain knowledge is provided, and that even infants are able to recognise relational similarity (Goswami, 1992). The tasks administered and the procedures used in these studies, however, are not totally convincing. Often, multiple-choice items were presented, which can be partly solved by sheer guessing (Goswami, 1995; Goswami & Brown, 1989), sometimes items were administered, for which the correct solution can be found by analogical reasoning as well as by free association (Goswami & Brown, 1989) or by direct application of domain knowledge (Brown, 1989). Sometimes, the relations between the A- and the B-term or between base and target were presented explicitly by the test instructor (Crisafi & Brown, 1986), or feedback about correct or incorrect solutions was given (Goswami & Brown, 1989). Because the geometric analogies administered in our study demanded little domain knowledge or verbal skills and, on the other hand, provided little opportunity to guess the correct answer, they probably measured analogical reasoning directly. Geometric analogies, therefore, should be used again in subsequent studies.

The fact that the definition of a successful performer depends on the sort of task administered constitutes a problem in other domains of cognitive development as well. In most children, for example, the conservation of volume occurs later than the conservation of number. Although one might already label the child who masters the conservation of number a conserver, several indicators of a discontinuity have been demonstrated for the development of volume conservation as well (van der Maas et al., 1993). Hence, in the development of analogical reasoning, there might also be several domain-specific discontinuities. We believe that the question at which age exactly analogical reasoning is acquired, is less important than the questions: what kind of changes occur in the development of analogical reasoning?, how can they be detected?, and how can they be explained?

Overall, the results of our study gave a first indication that the proposed rule shift in the development of analogical reasoning can be observed in children attending the first grades of elementary school. One must, however, be wary of a new trend or relation that has been discovered in only one study. What we need next is the replication of our results under more ideal conditions including samples with uniform age distributions, an individual testing procedure for each sample, and the administration of a highly homogeneous item set. Later, a replication of the findings for different tasks, as for instance verbal analogies or problem analogies, should be carried out.

Finally, further exploration of the rule shift in analogical reasoning should be guided by the question whether it represents a genuine discontinuity in the development of children's thinking. For this purpose, it might be useful to adapt the model of a transition as defined by catastrophe theory (Gilmore, 1981), which can be applied to issues of cognitive development (van der Maas & Molenaar, 1992). A transition in catastrophe theory is a discontinuity in a behavioural variable that is related to continuous change in a control variable. Bimodality in cross-sectional performance data, which we found in this study, is one of the necessary indicators of a transition. In order to find sufficient evidence of such a transition in the development of analogical reasoning and to determine the control variable, extensive longitudinal and experimental studies are needed. For now, our data suggest that the search for a discontinuity in analogical reasoning could be a fruitful area of research giving a whole new dimension to inquiries into this domain of cognitive development.

Manuscript received March 1995
Revised manuscript received February 1996

REFERENCES

Alexander, P.A., Willson, V.L., White, C.S., & Fuqua, J.D. (1987). Analogical reasoning in young children. *Journal of Educational Psychology*, *79*, 401–408.

Brown, A.L. (1989). Analogical learning and transfer. What develops? In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 369–412). Cambridge: Cambridge University Press.

Case, R. (1985). *Intellectual development. Birth to adulthood.* London: Academic Press.

Crisafi, M.A., & Brown, A.L. (1986). Analogical transfer in very young children: Combining two separately learned solutions to reach a goal. *Child Development*, *57*, 953–968.

Everitt, B.S., & Hand, D.J. (1981). *Finite mixture distributions.* London: Chapman & Hall.

Fischer, K.W., Pipp, S.L., & Bullock, D. (1984). Detecting developmental discontinuities. Methods and measurement. In R.N. Emde & R.J. Harmon (Eds.), *Continuities and discontinuities in development* (pp. 95–121). New York: Plenum.

Gilmore, R. (1981). *Catastrophe theory for scientists and engineers.* New York: Wiley.

Goldman, S.R., Pellegrino, J.W., Parseghian, P., & Sallis, R. (1982). Developmental and individual differences in verbal analogical reasoning. *Child Development*, *53*, 550–559.

Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development*, *62*, 1–22.

Goswami, U. (1992). *Analogical reasoning in children.* Hove, UK: Lawrence Erlbaum Associates Ltd.

Goswami, U. (1995). Transitive relational mappings in three- and four-year-olds: The analogy of Goldilocks and the three bears. *Child Development*, *66*, 877–892.

Goswami, U., & Brown, A.L. (1989). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition*, *35*, 69–95.

Halford G. (1993). *Children's understanding. The development of mental models.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Hosenfeld, B., van den Boom, D.C., & Resing, W.C.M. (submitted). *Constructing geometric analogies for the longitudinal testing of elementary school children.*

Lohaus, A., Kessler, T., Thomas, H., & Gediga, G. (1994). Individuelle Unterschiede bei räumlichen Fähigkeiten im Kindesalter. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *26*, 373–390.

Mokken, R.J. (1971). *A theory and procedure of scale analysis with applications in political research.* The Hague: Mouton.

Raven, J.C. (1958). *Standard progressive matrices.* London: Lewis & Co.

Russo, J.E., Johnson, E.J., & Stephens, D.L. (1989). The validity of verbal protocols. *Memory and Cognition*, *17*, 759–769.

Stern, H.S., Arcus, D., Kagan, J., Rubin, D.B., & Snidman, N. (1995). Using mixture models in temperament research. *International Journal of Behavioral Development*, *18*, 407–423.

Sternberg, R.J. (1977). Component processes in analogical reasoning. *Psychological Review*, *84*, 353–378.

Sternberg, R.J., & Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child Development*, *51*, 27–38.

Sternberg, R.J., & Rifkin, B. (1979). The development of analogical reasoning processes. *Journal of Experimental Child Psychology*, *27*, 195–232.

Thomas, H. (1989). A binomial mixture model for classification performance: A commentary on Waxman, Chambers, Yntema, and Gelman (1989). *Journal of Experimental Child Psychology*, *48*, 423–430.

Thomas, H., & Lohaus, A. (1993). Modeling growth and individual differences in spatial tasks. *Monographs of the Society for Research in Child Development*, *58*, (9, Serial No. 237).

Thomas, H., & Turner, G.F.W. (1991). Individual differences and development in water-level task performance. *Journal of Experimental Child Psychology*, *51*, 171–194.

Titterington, D.M., Smith, A.F.M., & Makov, U.E. (1985). *Statistical analysis of finite mixture distributions.* New York: Wiley.

van der Maas, H.L.J. (in press). The dynamical and statistical properties of cognitive strategies: Relations between strategies, attractors, and latent classes. In K.M. Newell & P.C.M. Molenaar (Eds.), *Dynamical systems in development.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

van der Maas, H.L.J., & Molenaar, P.C.M. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review*, *99*, 395–417.

van der Maas, H.L.J., Walma van der Molen, J., & Molenaar, P.C.M. (1993). Discontinuity in conservation acquisition: a longitudinal experiment. In H.L.J. van der Maas, *Catastrophe analysis of stagewise cognitive development. Model, methods and application.* Doctoral dissertation, University of Amsterdam, The Netherlands.