Available online at www.sciencedirect.com

SCIENCE DIRECT®

Developmental Review 24 (2004) 289–321

DEVELOPMENTAL REVIEW

www.elsevier.com/locate/dr

Rules and development in triad classification task performance[☆]

Maartje E.J. Raijmakers,[*] Brenda R.J. Jansen, and Han L.J. van der Maas

Department of Developmental Psychology, University of Amsterdam, The Netherlands

Received 15 January 2004; revised 28 May 2004

Abstract

Rule use in perceptual classification was investigated in adults and in 4- to 12-year-old children. Two studies of performance on triad classification tasks with large samples ($N=226$ and $N=328$) are presented to (a) contrast theoretical predictions from the holistic-to-analytic-shift theory (Smith & Kemler, 1977) and the differential-sensitivity account (Cook & Odom, 1992), and (b) to contrast findings with Thompson's (1994) rule analysis technique with latent class analysis (LCA). The findings, which demonstrate the value of LCA in this context, support the differential-sensitivity account, but not the holistic-to-analytic-shift theory. All resulting latent class models included two one-dimensional classification rules, but no holistic and no identity rules. In addition, a small, but significant number of children and adults failed to show systematic rule use, even with increased salience of dimensional differences. Furthermore, developmental effects were found primarily with respect to the type of dimensions, to which the children attended. As children develop from 4 to 12 years, the dominant dimension is first brightness, then size, and finally, towards the age of 12, orientation.
© 2004 Elsevier Inc. All rights reserved.

Keywords: Triad classification task; Latent class analysis; Perceptual development; Perceptual classification; Free classification; Individual differences

[☆] The research of Dr. M.E.J. Raijmakers has been made possible by a fellowship of the Royal Netherlands Academy of Art and Sciences. We thank Ch. De Liagre Böhl, M. Lammers, M. Ooijevaar, D. Span, L. Swaab, and E. Verlinden for their valuable contribution to the experimental studies.

[*] Corresponding author. Fax: +31-206390279.

E-mail address: m.e.j.raijmakers@uva.nl (M.E.J. Raijmakers).

0273-2297/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.dr.2004.06.002

The holistic-to-analytic shift theory predicts a qualitative change in children's perception of multi-dimensional objects during development. The general idea is that young children judge similarity by considering all stimulus compounds together, i.e., holistically (e.g., Smith & Kemler, 1977; Smith, 1989). As they mature, children increasingly apply analytic methods to assess the similarity of multi-dimensional stimuli. Given that the stimulus dimensions are separable, they consider similarity on different stimulus dimensions separately. In terms of rule use, this theory predicts that young children apply holistic rules in perceptual classification tasks. That is, they judge the stimuli with the smallest overall differences to go together best. Older children and adults are expected to apply identity or dimensional rules. According to an identity rule,[1] as suggested by Smith (1989), stimuli that share a value on one or more dimensions go together best. The application of a dimensional rule results in the judgment that two stimuli go together best if their values on one particular dimension, which is designated beforehand, are most similar.

However, Smith (1989) noted that pre-school children sometimes follow dimensional strategies, and that adults sometimes classify holistically. The latter is observed particularly when the adults are under time pressure, when they have to perform a concurrent task, or when they are responding impulsively (Kemler Nelson, 1989; Smith & Kemler Nelson, 1984; Ward, 1983). Hence, the holistic-to-analytic shift theory embodies two main assumptions: (1) There are two, qualitatively different, processing modes: holistic and analytical processing. (2) There is a developmental sequence such that processing is holistic during early childhood, and is gradually supplanted by analytical processing as children mature.

An alternative theory is the differential-sensitivity account (Cook & Odom, 1992), which is based on ideas of Gibson (1979). According to this theory, young children, similarly to adults (e.g., Ashby, Queller, & Berretty, 1999), generally focus on a distinct dimension. Humans of all ages generally use dimensional rules. What develops according to this theory is (1) the sensitivity to differences in dimensional values; (2) the consistency of behavior on a task; (3) the positions of dimensional differences in the salience hierarchy. This approach to the perception of multi-dimensional objects is consistent with some aspects of information-processing theories of proportional reasoning. In particular, Siegler (1981) found, in proportional reasoning tasks, that young children consider only the most dominant dimension, when confronted with an item with at least 2 relevant stimulus dimensions. Children as young as 5 years have been found to perform in a rule-governed fashion (Siegler, 1978).

The triad classification task has been used widely to investigate developmental changes in the perception of multi-dimensional objects (see e.g., Cook & Odom, 1992; Smith & Kemler, 1977; Smith, 1989; Thompson, 1994; Thompson & Markson, 1998; Ward, 1983; Wilkening & Lange, 1987). In the standard version of the task, people are shown three stimuli that are characterized by values on two dimensions, e.g., brightness and size. The stimulus compounds are depicted schematically in the left panel of Fig. 1. The three stimuli are designed in such a manner that two stimuli

[1] To avoid confusion of dimensional-identity rules with dimensional rules, we refer to Smith's (1989) dimensional-identity rule as the identity rule.

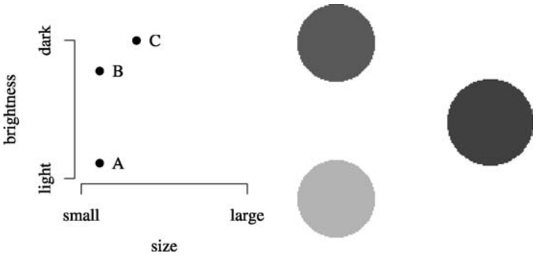


Fig. 1. (Left) Schematic representation of a standard item of the triad classification task consisting of three stimuli A, B, and C with different values on two dimensions. (Right) The corresponding item as presented to the participant. The stimuli are, clockwise starting at the top, B C A.

(A and B in Fig. 1) have the same value on one dimension, but differ greatly on the other dimension. The third stimulus (C in Fig. 1) differs only slightly from one of the other stimuli (B in Fig. 1) on both dimensions.[2] The right panel of Fig. 1 shows an example. People are asked to judge which two (of the three) objects go together best. In this task the behavioral variable typically studied is the number of BC classifications relative to the number of AB classifications. A BC classifications is consistent with a holistic rule, while an AB classification is consistent with an identity rule. According to a dimensional-brightness rule B and C go together best, and according to a dimensional-size rule A and B go together best. AC classifications are called haphazard or erroneous responses, and are expected to occur infrequently.

The most frequently applied method to test predictions of the holistic-to-analytic shift theory is a group-wise analysis of the association between age and the number of AB classifications relative to the number of BC classifications on triad classification items. The developmental trend is studied by comparing group averages. The main result of such analyses is that the frequency of AB classifications increases, whereas the frequency of BC classifications decreases with age. However, Cook and Odom (1992) stressed the importance of analyzing individual response patterns instead of analyzing group averages of AB responses, because salience hierarchies for dimensional differences are subject to individual differences.

On the basis of a systematic analysis of individual response patterns on the triad classification task, Thompson (1994) presented evidence for the differential-sensitivity account, and against the holistic-to-analytic shift theory. She found evidence for the application of dimensional rules, but little evidence for the application of holistic and identity rules, in the youngest age groups (kindergarten). Furthermore, she showed that the consistency of rule use increases with age. Wilkening and Lange (1987) also found little evidence for the application of holistic rules in perceptual classification.

Thompson's rule analysis methodology largely resembles Siegler's (1981) rule-assessment methodology (see next Section). Jansen and van der Maas (1997) proposed latent class analysis (LCA; e.g., Clogg, 1994) as a statistical approach to

[2] Because different stimulus dimensions are not directly comparable, the difference between stimuli B and C is very small on both dimensions, compared to the one-dimensional difference between A and B (see Fig. 1).

Siegler's rule-assessment methodology for a proportional reasoning task, the balance scale task. The value of LCA in this context is that rules need not to be predefined as in Siegler's and Thompson's methods. Hence, using LCA rules that are not expected (or patterns that do not resemble a rule) can also be detected. Moreover, LCA allows for statistical testing of the goodness of fit of rule models. This provides an objective criterion by which one may determine whether a given rule is present in the best (well fitting, parsimonious) model. As argued below, in the analysis of the triad classification task, LCA offers a statistically reliable method for assessing rules.

In the present article, we apply LCA to assess rules on the triad classification task. We follow Thompson's (1994) approach in the design of our test. First, we briefly discuss Thompson's rule analysis methodology. Second, we discuss the value of LCA in analyzing rules in triad classification performance. Third, in two cross-sectional studies, we establish which rules were applied in the triad classification task by children of 6–10 years of age (Experiment 1), and by children of 4–12 years of age, and adults (Experiment 2). Finally, we investigate the age-related changes in the application of rules, including age related shifts in the salience hierarchy of dimensional differences.

Thompson's rule analysis

Thompson (1994) and Thompson and Markson (1998) designed a rule-testing framework to systematically study individual differences in rule-use on the triad classification task. Unlike the classical group-wise analysis, Thompson's rule-testing framework can accommodate individual differences in rule use. The rule-testing framework consists of a well-developed approach to the design of suitable tests, and analysis techniques that focus on individual response patterns.

Test design is based on a systematic categorization of types of triad classification items, in which each proposed rule gives rise to a distinct response pattern to these items. Schematic displays of the applied item types are shown in Fig. 2. A type I triad is the most commonly used item type (cf. Smith & Kemler, 1977). It consists of two stimuli, denoted A and B, that share identical values on one dimension, but differ considerably on a second dimension. The third stimulus, C, differs from stimulus B slightly on both dimensions. A and C differ slightly on one dimension, and considerably on the other dimension. Type II triads resemble type I triads, in that B and C still have the closest over-all similarity, and A and B share a common value on one dimension. The difference is that dimensional values of stimulus C are intermediate to dimensional values of A and B. A type III triad consists of two stimuli, A and C that each share one dimension with stimulus B. Moreover, A and B differ considerably on one dimension, whereas B and C differ only a little on the other dimension. Type IV triads (not used in Thompson, 1994) resemble type III triads, but the difference between A and B is comparable to the difference between C and B. The difference between subtype a and b triads concerns the dimension, on which A and B share identical values. Specifically, in Fig. 2, stimuli A and B of types Ia, IIa, and IIIa (Ib, IIb, and IIIb) have identical size (brightness).

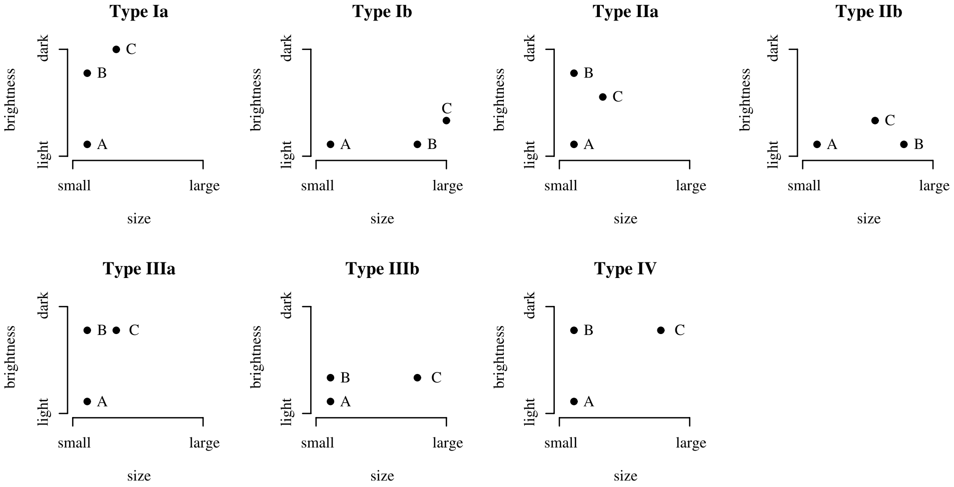
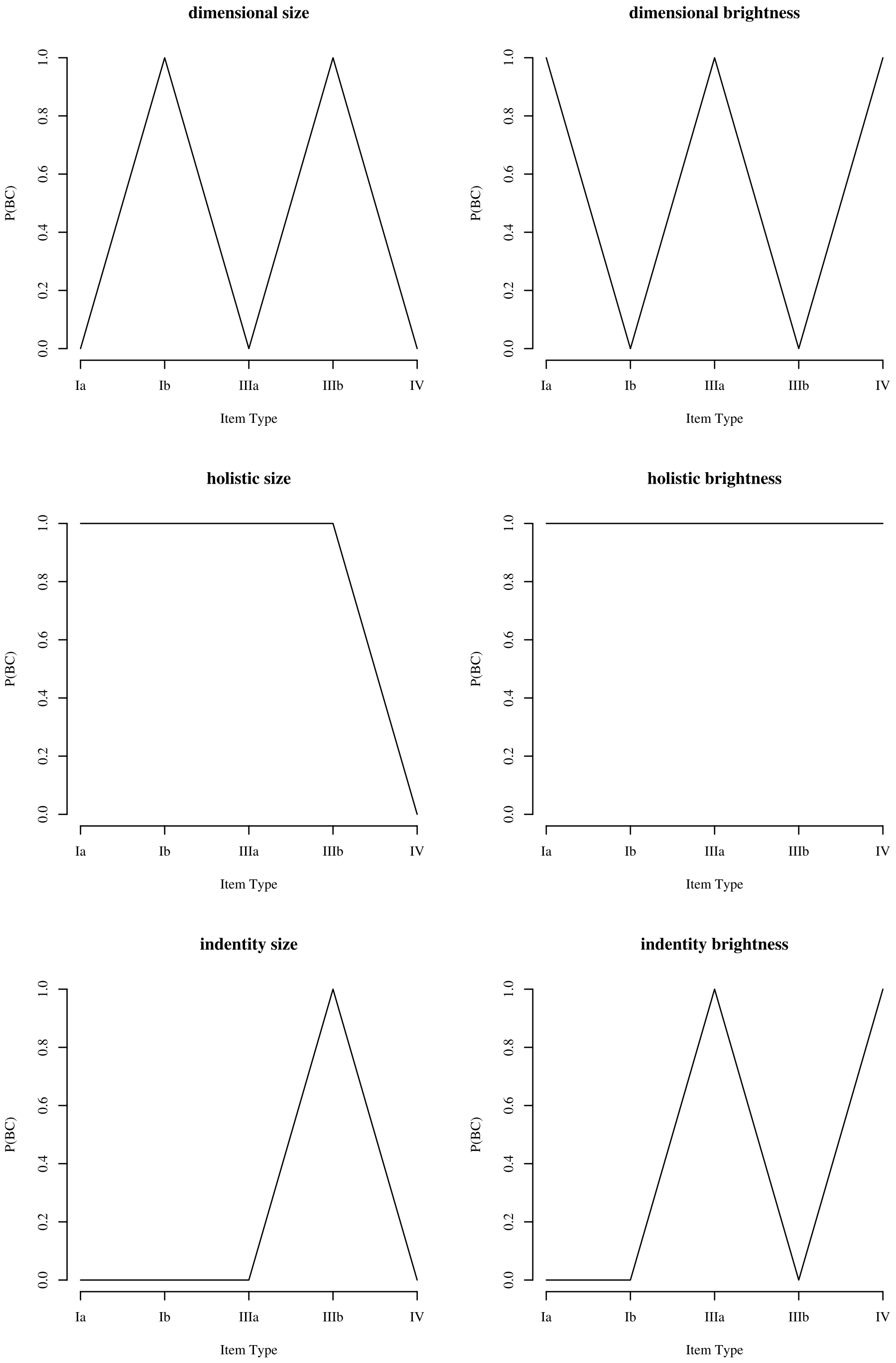


Fig. 2. Item types used by Thompson (1994) and Thompson and Markson (1998). The axes, as they are displayed here, agree with stimuli presented in Experiment 1. In Experiment 2 horizontal and vertical axes display values of size and values of orientation, respectively.

Fig. 3 shows the expected response patterns associated with classification rules on different types of items. On basis of the three main rules, namely the holistic, identity, and dimensional rules, and two subtypes, a and b, six rules can be distinguished. These are holistic-brightness and holistic-size, i.e., holistic rules, which are associated with a preference for, respectively, brightness and size on type IV items; identity-brightness and identity-size, i.e., identity rules, which are associated with a preference for, respectively, brightness and size on type III and IV items; and dimensional-brightness and dimensional-size, which result in classifications on basis of, respectively, the brightness and size dimensions. For example, Thompson's (1994) test in Experiment 1 included 10 items each of types Ia, Ib, IIa, and IIb presented a given day, and the same number of items presented a subsequent day.

To match an individual response pattern comprising 40 items (i.e., all items of day one) with all six rules, Thompson (1994) calculated the proportion of responses that were consistent with a given rule (i.e., the PC-value). The rule with the highest PC-value is then designated the dominant classification rule. The PC-value provides an indication of the consistency of the individual's rule use. We interpret Thompson's (1994) proposed test for the reliability of a classification as follows. She calculates a reliability of a classification from the number of responses that are consistent with the dominant rule. To this end, she tests the null-hypothesis that the number of consistent responses are generated by chance. Generated by chance means that the prior probability of an AB response equals the prior probability of a BC response equals 0.5. To establish that a subject is generating responses with a probability larger than .5, one defines a critical area of the binomial distribution associated with the number of items, the value of the probability associated with the null hypothesis (.5), and a given α value. For instance, given 20 items and an α of .06, any number of AB responses greater than 14 (probability .058) would result in a rejection of the hypothesis



dimensional size

dimensional brightness

holistic size

holistic brightness

indentity size

indentity brightness

that the AB responses are generated with a probability of .5. For a rule classification to be reliable Thompson requires that the classification of every subset of items (i.e., Ia and IIa versus Ib versus IIb) is reliable. This systematic approach of rule testing allows for assessment of individual rules without the requirement of verbal justifications. Thompson showed that rule analysis of triad classification data resulted in conclusions that are inconsistent with those obtained from group-averaging methods.

In the present article, we focus on two aspects of the rule-testing framework that are open to improvement. First, the method used in this framework can only detect predefined rules. If a child applies an alternative rule, the response pattern is either misclassified, or it appears as an unreliable classification of the dominant rule. Hence, the use of unexpected rules passes unnoticed. Second, Thompson's statistical analyses were conducted on the level of individual response patterns. For each individual response pattern a statistical test was done, which defines the reliability of the classification. This procedure is subject to the problem of chance capitalization. For instance given an α of .05, one expects a priori that one out of 20 tests will result in a significant p value. If 4 out of 56 conducted tests indicate a reliable holistic rule classification (as in Experiment 2 of Thompson, 1994), one may wonder whether the inclusion of the holistic rule in the rule model is statistically justified. We prefer a method that is less susceptible to this chance capitalization. In addition to the classification of individual response patterns, we would like to determine the most parsimonious set of rules that can account for the data as a whole. LCA is a statistical technique for rule analysis of categorical data that satisfies these requirements. In the next Section we briefly discuss LCA in the application to triad classification data.

Latent class analysis for detecting rules

The assessment of rules on the basis of a series of responses on triad classification items involves the analysis of categorical observed variables and a categorical latent variable (Clogg, 1994). In the present context, the observed variables are the responses to the triad classification items with AB, BC, and AC as nominal categories. From these responses we want to infer rules, which are not directly observable. To accommodate the use of rules, we introduce a latent variable with nominal categories, each representing a given rule, such as holistic rule, dimensional-brightness rule, etc. LCA is highly suited in the present case, because LCA defines models in which a categorical latent variable (the applied rule) explains the observed associations among categorical observed data (responses on triad classification items). As LCA is a standard statistical technique, and many good introductions are available

Fig. 3. Diagrams of the expected rule patterns for different item types. The horizontal axes display the item type, the vertical axes display the expected probability of BC responses according to each rule. *Note.* In Experiment 2, the response patterns of the brightness rules agree with the equivalent orientation rules. For example, the response pattern of the dimensional-orientation rule is equivalent to the response pattern of the dimensional-brightness rule.

(Clogg, 1994; Heinen, 1996; McCutcheon, 1987; Rindskopf, 1987), we limited our presentation to a brief description of this method.

In LCA we model the frequency table of all possible response patterns. If we want to model 20 items each with 3 response categories, the frequency table consists of 3^{20} cells. A latent class model (LC model) represents a mixture of distributions, in which each class forms one component distribution in the mixture distribution. Each class of an LC model is defined by two kinds of probabilities: one unconditional probability, and several conditional probabilities. Unconditional probabilities represent the size of the classes, i.e., the proportion of participants whose behavior is attributable to the use of a given rule, i.e., the proportion of individuals in the class that represents the rule. The probabilities with which individuals *within a given class* produce particular responses to a given item are the conditional probabilities. The probabilities are *conditional* on membership of a given class. By careful inspection of the configuration of conditional probabilities associated with a latent class, one may infer the rule that the members of the class use. For example, suppose that a well-fitting LC model contains (among other classes) one class with an unconditional probability of .3. This suggests that 30% of the cases are in this class. Furthermore, suppose that within this class the conditional probabilities producing a BC response to items of type Ia, IIIa, and IV are near 1.0. This implies that the conditional probabilities of responding AB and of responding AC are close to zero. So the members of this class will very often produce a BC response to items of type Ia, IIIa, and IV. Finally, suppose that the conditional probabilities of responding AB to items of type Ib and IIIb are near one. This implies that the conditional probabilities of BC and AC responses are close to zero. From this configuration of conditional probabilities (see Fig. 3) one would infer that 30% of the participants follow a dimensional-brightness rule. Note that we do not assume that the conditional probabilities of the predicted responses (BC given item types Ia, IIIa, and IV; AB given items types Ib, IIIb) are exactly 1.0. Although this is theoretically possible, it is unlikely that children will behave deterministically according to a given rule. They may make errors of commission for a variety of reasons, including boredom, fatigue, inattentiveness, etc. LCA accommodates this by allowing the conditional probabilities to be less than one. Note finally that the conditional probabilities may assume intermediate values, which are not necessarily easy to interpret in terms of the application of a rule.

In fitting an LC model one specifies the number of latent classes. Given this number, the conditional probabilities and unconditional probabilities are estimated from the data by means of some method of estimation. Here we use the method of maximum likelihood to estimate the probabilities (Azzelini, 1996). We mainly performed exploratory LCA. This means that models with an increasing number of classes are fitted to the data until a model is found that fits the data reasonably well. The ultimate choice for a model depends both on its goodness of fit, its parsimony, and on interpretability of the probability estimates. Clearly a well fitting, but uninterpretable, model is of little use.

The fit of an LC model may be assessed by a likelihood ratio χ^2 statistic, which indicates the discrepancy between the observed and expected response frequencies. Given several well-fitting models, we select the most parsimonious model by means of the Bayesian Information Criterion, BIC (Schwarz, 1978). BIC considers the likelihood ratio χ^2 statistic in relation to the number of participants and the number of degrees of freedom. A relatively low BIC indicates a more parsimonious, relatively well-fitting model.

Advantages and disadvantages of LCA

As in Thompson's rule matching analysis, the focus of LCA is on individual differences in rule use. By using LCA, we avoid some criticism of other rule-analysis techniques (Jansen & van der Maas, 1997; Strauss & Levin, 1981; Wilkening & Anderson, 1982; see also Siegler & Chen, 2002). First, we do not need to define the nature of the expected rules a priori (although this is possible in LCA). Second, by means of exploratory LCA, we determine the number of rules (i.e., the number of latent classes) that is required to achieve the best and the most parsimonious description of the data by means of statistical, objective, fit measures. The interpretation of the conditional probabilities of each class in terms of a rule is done a posteriori. In principle we may encounter unanticipated rules. As mentioned above, the conditional probabilities of the latent classes may be difficult to interpret in terms of the application of a rule. Note that the type of rules that can be distinguished depend strongly on the items in the test. Evidently, the items used in the test limit the types of rules that can be detected in an LCA. For instance, if one expects to detect three rules, one requires items that give rise to clear rule-related differences in conditional probabilities of the responses to the items. An item that elicits the same responses, regardless of which of the three rules is used, cannot help to establish the use of the three hypothesized rules. Items therefore have to be chosen carefully. The goodness of fit tests address the fit of the LC model as a whole.

A disadvantage of LCA is that the asymptotic results associated with maximum likelihood estimation and testing, such as standard errors of estimates and the χ^2 goodness of fit index, are unreliable if the data are sparse. Sparseness implies that the frequency table of the data includes many cells with few or zero observations. In the present case, sparseness is a problem due to the relatively large number of items (24 items with 3 response choices give rise to a frequency table of 3^{24} cells). To reduce this problem, we decided to analyze subsets of the items to address specific aspects of the theory (see Boom, Hoijtink, & Kunnen, 2001, for a different approach). Nevertheless, as sparseness may still be a problem in the subsets, we applied Monte Carlo bootstrapping procedure to obtain reliable test statistics (Langeheine, Pannekoek, & Van de Pol, 1995). We analyzed two kinds of subsets of items. First, we tested for homogeneity of item types (type Ia, type Ib, etc.) by fitting LC models to responses to items of the same type. Inconsistent, or heterogeneous, responding may occur, which cannot be interpreted in terms of simple rule use. (Jansen & van der Maas, 1997). This may occur when children switch between rules during the test. For instance, Thompson (1994) detected switches of rules on two successive days. In addition, it may be due to the unreliability of particular items (i.e., producing many erroneous AC responses). Second, we test for the presence of rules by analyzing a carefully selected subset of items, which comprised a well-balanced mix of items types. From each item type we selected the most reliable items, based on the number of AC (i.e., erroneous) responses.

A second, related disadvantage of LCA is that often large sample sizes are needed to obtain reliable model estimates. Therefore, to facilitate data collection, we designed paper-and-pencil tests of the triad classification task, which are amenable to classroom-wise administration. One might expect paper-and-pencil tests to be less reliable than individual administrations, which are common in triad classification tests. However, other task domains such as the balance scale task, show high levels of agreement between paper-and-pencil tests and individual administrations (Chletsos, 1986). We investigated the reliability of our data by comparing results obtained with these data to Thompson's (1994) results concerning the number of AC responses (i.e., erroneous or haphazard responses) and the consistency with the expected rules (i.e., PC-values).

General data analytic procedure

Data obtained in the two experiments that are described below were subjected to LCA in the following manner. We first fitted LC models per item type to test for homogeneity. Second, we fitted LC models on a set of selected items of several types. After identifying the simplest, best fitting LC model for this set of items, we assigned individual participants to one of the latent classes by means of posterior probabilities of class membership (Langeheine et al., 1995). The rule assignments were then used in a multinomial logistic regression analysis (Hosmer & Lemeshow, 1989) to model age-related changes in rule use. We conducted two experiments to detect rule use in triad classification task performance. The aim of the first experiment, which is a partial replication of Thompson's (1994) study, is to evaluate Thompson's findings using LCA of the data. The second experiment is a replication of Experiment 1 using different stimulus dimensions, an extended age range, and additional experimental manipulations to test hypotheses that were based on the results of Experiment 1.

Experiment 1

We based the experimental setup on the design of Thompson's (1994) Experiment 1. The major difference between the present and Thompson's study is our use of a paper-and-pencil test instead of a computerized procedure. We analyzed the data in three ways: (a) group-wise analysis, (b) Thompson's rule matching analysis, and (c) LCA. We present only basic findings obtained in the first two types of analyses. We present the results obtained with LCA in more detail, because LCA is statistically most advanced.

Method

Participants

We tested 226 participants from four successive grades of two primary schools in the Netherlands. The breakdown of the sample by age is as follows: 26 6-years of age ($M = 6$ years 8 months, $SD = 2.64$ months), 46 7-years of age ($M = 7$ years 6 months, $SD = 3.96$ months), 63 8-years of age ($M = 8$ years 6 months, $SD = 3.12$ months), 61

9-years of age ($M = 9$ years 5 months, $SD = 3.72$ months), 29 10-years of age ($M = 10$ years 4 months, $SD = 3.00$ months), and 1 11-years of age (age = 11 years).

Design and stimuli

The test consisted of twelve type I items and twelve type III items. Relevant stimulus dimensions were brightness and size. Thompson's type II items are not expected to discriminate other rules than do type I items. Moreover, Thompson (1994) showed that type II items generated more AC responses. This means that they are less reliable than type I items. We therefore excluded type II items from the test. We also excluded type IV items from the test (as in Thompson, 1994). The exclusion of type IV items makes it impossible to distinguish the holistic-brightness rule from the holistic-size rule. With type I and type III items we can distinguish the following rules: holistic rule, identity-brightness rule, identity-size rule, dimensional-brightness rule, and dimensional-size rule.

Stimuli consisting of circles were designed with 7 levels of size and 7 level of brightness. The size was determined by the diameter, which ranged from 1.6 to 2.5 cm in steps of .15 cm, Brightness was determined by the percentage of black dots per cm^2, which ranged from 20 to 80% in steps of 10%. Stimuli were generated on a computer, and all the booklets were directly printed from file. Evidently, brightness levels on screen, as Thompson presented them, are difficult to compare with brightness levels on paper. We chose the brightness levels that resembled those in Thompson's test as closely as possible on the basis of visual inspection. We return to the importance of exact dimensional values below.

Type I items were constructed such that the difference between stimuli A and B was three steps on one dimension, and the difference between stimuli B and C was one step on both dimensions. Type III items were constructed such that the smallest difference was two steps and the largest difference was three steps. Stimulus circles, satisfying these constraints, were randomly selected from the set of 49 unique circles. The three circles that constituted an item were presented in a triangle with the centers 5 cm apart. Four configurations of triangles were used, which were chosen randomly. The positions of the stimuli A, B, and C within the triangle, i.e., the triangle configuration, were randomized, as was the order in which the stimuli were presented. The test consisted of 24 items, i.e., six items of each of the 4 subtypes (Ia, Ib, IIIa, IIIb). To ensure that children could not copy each others work in the classroom, and to test the effect of sequence of presentation and the effect of the configuration of the three stimuli, four different versions of the paper-and-pencil test were produced.[3]

Procedure

Participants were tested in a classroom setting. They collectively received instructions from a student before they individually completed their test booklets. The instructions incorporated an example item composed of three flowers that was shown on a large poster in front of the classroom. This same item featured on the front page

[3] Detailed information about the items and different versions of the test can be obtained from the first author.

of the test booklet. The stimulus dimensions of the example item (patterns filling in the heart and the leaves of the flowers) differed from that of the test items. Moreover, there was no incorrect choice, as response AC in standard items. The instruction was to identify the two that best go together, which is the standard instruction for triad classification items (see e.g., Thompson, 1994). No instruction was given about separate stimulus dimensions, or about rules one can apply in solving the items. The children had to connect the two flowers that they thought went together best, with a straight line. Children could, if they wished, correct a response by crossing out the line and drawing a new one. Following the example item with flowers, two example items with circles were shown, and the three possible responses were pointed out. Again, no explanation was given why a given combination was better than others. After the instruction, each child individually solved the 24 test items in his or her test booklet. Students attended the children during the test to make sure that they did not copy each others work.

Results

Group-wise analysis

On the paper-and-pencil test, the proportion of AC classifications on type I items was small ($M = 0.064$, $SD = .078$). In contrast to standard findings for type I and type II items, the total number of AC classifications showed no decrease with age, $F(4, 221) = .14, p = .97$. The second standard finding in literature is that the number of AB classifications observed on the type I items increases with age. This implies that the number of identity responses, in contrast to the number of holistic responses, increases with age. Hence, the number of AB classifications relative to the number of valid responses (AB and BC) is of actual interest ($M = .31$, $SD = .14$). Here, age differences were not significant, $F(4, 221) = 2.26, p = 0.06$. In the discussion of Experiment 1 and 2, we offer simple explanations for the discrepancy between the present findings and those in literature.

Thompson's rule matching analysis

In rule matching analysis the observed response patterns of each participant were compared to the expected response patterns associated with the rules (Fig. 3). A response pattern was classified by the rule with the maximum PC-value. If two rules had the same maximum PC-value, the response pattern was classified as the rule that was used most frequently, since the a-priori chance of being classified as that rule is highest. Table 1 shows the rule classification and the accompanying average PC-values for each age group. Most children used the dimensional rules, but 17% of the response patterns were classified as most consistent with the holistic rule. Little evidence was found for the identity rule in any of the age groups.

In spite of differences in experimental design, on average, PC-values in the present experiment did not differ significantly from the PC-values of Thompson's (1994) first experiment ($F(1, 251) = 2.71$ $p = .10$). In contrast with Thompson's findings, however, average PC-values did not significantly increase with age ($F(4, 221) = 1.12$, $p = .35$). PC-values did vary with rule ($F(4, 221) = 6.2$, $p < .001$). Post hoc analyses showed

Table 1
Best fitting rules and corresponding average PC-values for participants in Experiment 1

Rule	Percentage of subjects						PC-value					
	Age						Age					
	6	7	8	9	10	Tot	6	7	8	9	10	Tot
H	23	15	17	16	17	17	.71 (.11)	.64 (.07)	.66 (.10)	.67 (.08)	.74 (.10)	.68 (.09)
IB			2	3		1			.79 (–)	.56 (.03)		.64 (.13)
IS				3	3	1				.77 (.09)	.63 (.13)	.72 (.10)
DB	50	54	30	18	30	34	.78 (.11)	.80 (.12)	.77 (.13)	.84 (.12)	.79 (.13)	.79 (.12)
DS	27	30	51	59	50	46	.76 (.12)	.73 (.18)	.83 (.09)	.79 (.13)	.80 (.13)	.79 (.13)
N	26	46	63	61	30	226	.76 (.11)	.75 (.14)	.78 (.12)	.77 (.14)	.78 (.13)	.77 (.13)

Note. H, holistic rule; IB, identity-brightness rule; IS, identity-size rule; DB, dimensional-brightness rule; DS, dimensional-size rule; PC, proportion consistent; standard deviations of PC-values are between brackets.

that the PC-values of the dimensional rules were higher than the PC–value of the holistic rule. Interaction effect of rules by age was not significant ($F(8, 211) = 1.28$, $p = .25$; tested without identity rules). The results of this analysis support Thompson's evidence for the dominant choice of dimensional rules among all age groups.

Latent class analysis

LCA was performed in two steps. First, we fitted LC models to the responses to each item type to test the homogeneity of the responses to these item types. If all participants responded consistently according to one of the rules, we would expect two classes for each item type: one class of AB responders (i.e., response patterns with consistently high probabilities of AB responses) and one class of BC responders (i.e., response patterns with consistently high probabilities of BC responses). Table 2 shows the selection of the simplest, best fitting LC model for each item type. Fig. 4 shows the parameter estimates of the most parsimonious, and best fitting LC models.

A 2-class model provided the best description of the responses to the type Ia items (see Fig. 4). The first class, comprising 45% of the cases (i.e., the estimate of the class probability is .45), is characterized by a relatively high conditional probability of responding AB (and thus a low probability of BC responses) on items 1–3. Items 4 to 6, however, gave rise to more ambiguous response patterns: There was an increased probability of responding AC, and there was little difference in probability of responding either AB or BC. The second class of the LC model (probability .55) shows a clear pattern of BC responses to all items 1–6.

The model selected in the analysis of the responses to the type Ib items has similar characteristics: It includes one class consisting of BC responders (probability .57). The other class (probability .43) is characterized by high probabilities of AB responses for three items (4, 5, and 6), and a more ambiguous configuration of response probabilities on the other three items. In this class again the probability of AC responses is somewhat increased on the more ambiguous items of the AB responders.

Table 2
Goodness-of-fit and selection of latent class models per item type of participants in Experiment 1

Itemtype	Model	L^2	pb (L^2)	*df*	BIC
Ia	1 class	516.50	<.001	716	2351.72
	2 class*	272.90	.11	704	2167.74
	3 class	240.91	.13	697	2173.69
Ib	1 class	393.25	<.001	716	2043.23
	2 class*	205.54	.70	706	1909.73
	3 class	177.96	.36	697	1930.94
IIIa	2 class	239.23	<.001	705	1802.17
	3 class*	179.30	.07	698	1780.19
IIIb	2 class	169.38	<.001	708	1567.90
	3 class*	104.50	.11	703	1524.71
Combined	2 class	317.19	<.001	1272	2118.10
	3 class	254.42	.33	1263	2125.39
	3 class-a*	297.47	.09	1275	2071.44

Note. *Indicates the most parsimonious, best fitting model; L^2, likelihood ratio statistic; pb(L^2), p value of L^2 obtained by Monte Carlo bootstrap; *df*, degrees of freedom; BIC, Bayesian Information Criterion. We do not report the fit of 4-class models, because they resulted in unreliable parameter estimates. Model 3a is a 3-class model with equality constraints for type Ia and Ib items and equality constraints for type IIIa, and IIIb items.

A 3-class LC model provided the best description of the responses to type IIIa items. The first class of the model (probability .43) is characterized by relatively high probabilities of responding AB. The second class (probability .28) is characterized by high probabilities of responding BC. Finally, in the third class (probability .29) the probabilities of either AB or BC responses were about equal. The same pattern was present in the LC model of the responses to the type IIIb items: one class (probability .30) is characterized by relatively high probabilities of responding AB, the second class (probability .42) is characterized by relatively high probabilities of responding BC, and the third class (probability .28) is not characterized by any clear response pattern, although there was a preference for responding BC for items 4, 5, and 6.

In conclusion, overall the homogeneity of item types was supported. That is, within classes the conditional probabilities of all item were similar. However we observed the following exceptions: Some type Ia and Ib items showed deviating conditional probabilities in one group of participants (the mostly AB responders). The LC models of the type IIIa and IIIb items included a group of participants that did not manifest a clear and homogeneous response pattern. Note that the items with the predicted response patterns were also characterized by the lowest probabilities of AC responses. Homogeneity of item types was sufficient to obtain a representative selection of items for each item type.

The second prediction of the rule model that we tested concerns distinguishable response patterns over different item types, which correspond to different rules. We selected two items of each item type, and we analyzed these eight items together in a

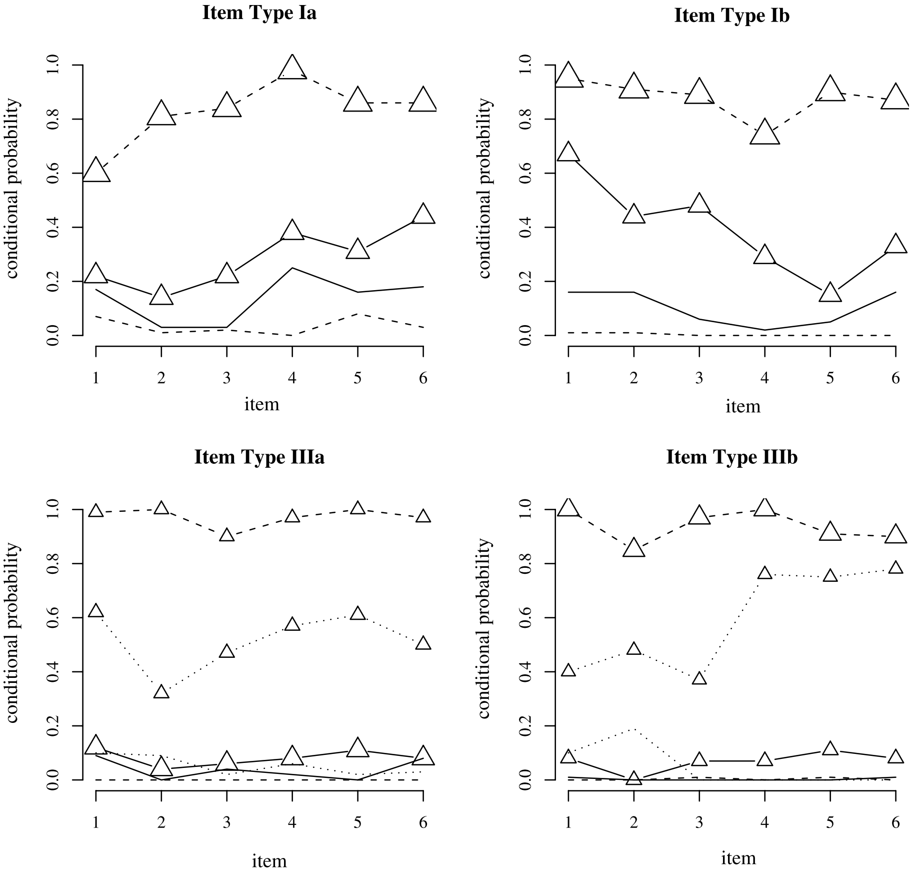


Fig. 4. Diagrams of LC models that are selected in Table 2 (i.e., marked with an *) in Experiment 1. The lines with triangles denote the conditional probabilities of BC responses. The lines without triangles denote the conditional probabilities of AC responses. The plain lines represent the class that agrees best with consistent AB responses; The dashed lines represent the class that agrees best with consistent BC responses; The dotted lines (for type IIIa and type IIIb items) represent the residual class. The size of the triangles denote the unconditional parameters, that is, the proportion of participants that is modeled by that class.

single LC model.[4] The selection of items was based on the number of AC-responses. Assuming that AC responses are indeed errors, we chose the two items from each item type with the fewest AC responses. On the two type III items with the fewest AC responses, only 7 participants chose the response AC. These participants were removed from the sample (resulting in $N = 219$). This reduced the number of response categories for type III items to two. In turn, this greatly reduced the size of the frequency table, and consequently increased the reliability of the estimates.

[4] The selected items were: Ia-2, Ia-3, Ib-4, Ib-5, IIIa-3, III-a5, IIIb-3, and IIIb-4, as they are displayed in Fig. 4. We conducted a second series of LCAs with different items to check whether the results depended on the selection of items. This analysis resulted in comparable LC models.

Although the number of AC responses was also small on type I-items, we retained the AC responses for these types, because removing participants with AC responses would result in an unacceptable decrease in sample size.

As shown in Table 2, exploratory analyses resulted in a 3-class LC model. In this model two classes corresponded to the dimensional-size and the dimensional-brightness rules. To reduce the number of freely estimated parameters, we constrained certain parameters to be equal. Specifically, we constrained the conditional probabilities of responses to items of the same type (i.e., type I items versus type III items) within a class to be equal in the classes that corresponded to one of the dimensional rules, such that the equality constraints are consistent with the dimensional rules. The latent class that did not correspond to any expected rule did not include any parameter constraints.

As the constrained model 3a has the lowest BIC, we considered this to be the most parsimonious, best fitting model. Fig. 5 shows the parameter estimates of this model. Comparison with Fig. 3 reveals that one class (probability .26) is characterized by a dimensional-size rule. Within this class, type Ia items, type Ib items, type IIIa items and type IIIb items are characterized by high probabilities of AB responses, BC responses, AB responses, and BC responses, respectively. A second class (probability .34) is characterized by a dimensional-brightness rule. The remaining class (probability .40) does not appear to be characterized by any rule in particular. Within this

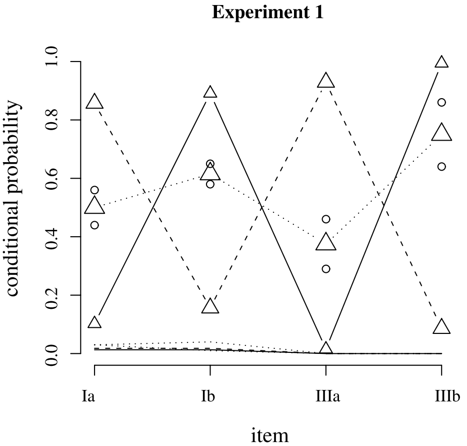


Fig. 5. Diagram of LC model 3-a of Table 2, which is an LC model of a combination of item types in Experiment 1. The lines with triangles denote the conditional probabilities of BC responses. The lines without triangles denote the conditional probabilities of AC responses. The plain lines denote the dimensional-size (DS) class; The dashed lines denote the dimensional-brightness (DB) class; The dotted lines denote the residual class (Res). For the DS class and the DB class the triangles denote the conditional probabilities of responding BC (the conditional parameters of the two modeled items per item type are constrained to be equal). For the Res class the open circles denote the conditional probabilities of responding BC for each item (the conditional parameters of the two modeled items per item type are not constrained to be equal); the triangles of the Res class denote the mean of the probabilities of the two items of the same type. The size of the triangles denote the unconditional parameters, that is, the proportion of participants that is modeled by that class.

Table 3
Cross table of number of participants classified by rule matching analysis and LCA in Experiment 1

Latent class analysis	Rules matching analysis					
	H	IB	IS	DB	DS	Total
Removed	2			1	4	7
Muddling through	32	3	3	7	41	86
Dim. Brightness	3			69	1	73
Dim. Size	2				58	60
Total	39	3	3	77	104	219

Note. H, denotes holistic rule; IB, denotes identity-brightness rule; IS, denotes identity-size rule; DB, denotes dimensional-brightness rule; DS, denotes dimensional-size rule. Note that 7 participants were removed from the LCA; they are not included in this table.

class, we could not discern any clear preference for either AB or BC responses on most items, except for 3 items.[5] The response tendency on these items was towards the dimensional-size rule. The standard errors of the parameter estimates in the residual class were higher (between .055 and .104) than those in the other two classes (between .023 and .047), which indicates that the estimates were less precise in this class.

On basis of the posteriori probabilities of class membership, we assigned individual participants to one of the three classes representing the rule that they most probably used. We subsequently investigated the agreement in assignment to rules based on Thompson's rule-matching analysis and on the LCA. Table 3 shows the resulting cross-tabulation, from which it emerges that 58% of the classifications match. The most striking differences between the results is the evidence for a holistic rule in rule matching analysis and a residual group in LCA. As it is clear in the cross-tabulation, most participants who were assigned to the holistic rule in Thompson's classification, were assigned to the residual group in the LC model. Conversely, participants who were assigned to the residual group either followed the holistic rule or the dimensional-size rule according to the rule matching analysis. The latter group includes 19% of the participants.

We can also investigate the relation between age and rule use by assigning individuals to classes of the LC model on the basis of their posterior probabilities of class membership. Fig. 6 shows the observed probabilities of rule use by age and the predicted probabilities of a multinomial logistic regression model (Dobson, 2002). We applied the multinomial regression procedure to model the dependence of a nominal categorical response (rule use with nominal categories: dimensional-size rule, dimensional-brightness rule, and residual) on a continuous predictor variable (age). Multinomial logistic regression analysis showed that rule-use depended on age ($\chi^2(2) = 14.2$, $p < .001$). In particular, the probability of using the dimensional-size rule increased with age relative to the probability of using the dimensional-brightness

[5] The conditional probabilities differ significantly between responses within an item for items Ib-5 (*SE*'s are .069 and .075 for responses AB and BC, respectively), IIIa-3 (*SE*'s: .061 and .061), and IIIb-4 (*SE*'s: .055 and .055).

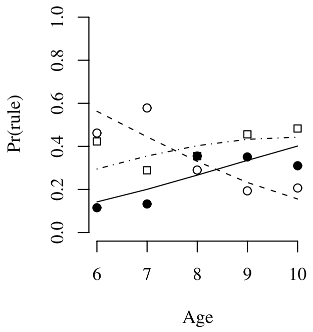


Fig. 6. Results of multinomial logistic regression analysis of the probability of rule use. Solid line is the regression line of the probability of using the dimensional-size rule, solid dots are the corresponding observed probabilities. Dashed line and open dots correspond to the use of the dimensional-brightness rule. Dot-dashed line and squares correspond to the probability of muddling through.

rule ($B = .58$; $\chi^2(1) = 13.0, p < .001$). The probability of being in the residual group as opposed to using one of the two-dimensional rules, did not change with age ($B = -.28$; $\chi^2(1) = 1.2, p = .27$).

Discussion

Group-wise analysis showed that the proportion of AC classifications, i.e., erroneous or haphazard responses, is comparable to standard findings in literature. In addition, individual response patterns were as consistent with rules as response patterns in Thompson's (1994) first experiment. Hence, the reliability of the group-wise paper-and-pencil test appears to be comparable to the reliability of Thompson's individual test. In contrast to findings in literature, we did not observe an age-related decrease of AC responses for type I items. However, a closer look at the results of Thompson's first experiment shows that for type I items separately, the age-related decrease of the number of AC responses was slight (.04, .03, and .02 in kindergartners, second graders, and fourth graders, respectively). In contrast to standard findings in the literature, we did not find an age-related increase of AB responses for type I items. Moreover, we did not find an effect of age on PC-value. These discrepancies may be explained by the difference in age groups between our and Thompson's (1994) first study: we did not include a group of 4-year-old children. This issue is addressed in the second study. More striking, however, is that the overall percentage of AB responses in type I items was low ($M = .31$, $SD = .14$, $N = 226$) compared to Thompson's Experiment 1 ($M = .47$, $SD = .07$, $N = 27$) and to Smith and Kemler's (1977) studies ($M = .54$, $SD = .19$, $N = 30$). This result is consistent with results of Thompson's (1994) second experiment, in which the average proportion of AB responses was low, and age-related increase of the number of AB responses was absent. Note that according to the group-wise analysis this would indicate a large number of subjects using a holistic rule. Nevertheless, Thompson's rule analysis showed very little evidence, and the LCA showed no evidence, in support of the use of holistic rules.

A striking result of the LCA of type Ia items and of type Ib items (Fig. 4) is that the expected AB responses were less consistent than the expected BC responses. Moreover, overall there were fewer AB responses than BC responses (note that these two results from the LCA are mutually independent). Hence, it appears that in applying dimensional rules, which were the only rules that we detected, type I items in Experiment 1 with expected AB responses were more difficult than items with expected BC responses. Assuming an AB response is generated by the application of the dimensional rule, this requires the comparison of a difference of zero steps with a difference of one step. Assuming a BC response is generated by application of the dimensional rule, this requires the comparison of a difference of three steps with a difference of one step. Apparently, with the type I items in Experiment 1, the latter comparison is easier to make (i.e., for those who apply the dimensional-size rule, items of type Ia are more difficult to respond consistently than Ib items). This conclusion is confirmed in the group of subjects who used rules (i.e., the participants modeled by the residual group of the LC model (Fig. 5) are disregarded). The consistency of expected AB responses is higher than the consistency of the expected BC responses ($F(1, 128) = 253.8, p < .001$).

The results of the LCA per item type largely confirmed the first hypothesis of the rule model in that the participants were found to respond consistently within item type. This enabled us to select a representative set of items to perform the LCA of a combination of item types. Moreover, the responses to the most reliable items (i.e., those that elicited the fewest AC responses) were most consistent with our predictions.

The LCA of a combination of item types, aimed at assessing rule use, resulted in a 3-class model: two classes agreed with dimensional rules (the dimensional-size and dimensional-brightness rules). Importantly, no evidence was obtained to support the use of the holistic or identity rules. An additional class, comprising 40% of the participants, could not be related to any expected response pattern. Generally, these participants alternated between AB and BC responses, but they displayed a small tendency towards the dimensional-size rule. They did not appear to guess blindly, because the number of their AC responses was low. Possibly, they switched strategies in an unsystematic way. We denote this class as the “muddling-through” group. Note that this class could not have been identified using Thompson’s (1994) method, because muddling-through is understandably not a recognized response mode in Thompson’s rule model. It is unlikely that this class can be identified with rule use, as it is not clear what these children actually do. Nonetheless, this class is large (40%), and does account for behavior that cannot be interpreted in terms of the use of either dimensional rule. An age effect was found among the rule users: The proportion of participants following the dimensional-size rule increased with age relative to the proportion of participants following the dimensional-brightness rule.

To replicate the results of the LCA, we conducted a second experiment. A possible explanation for the large group of participants who muddled through, is that the differences between stimuli were too small to systematically evoke rule-guided behavior. To explore this hypothesis, we systematically varied the size of stimulus differences in Experiment 2. To replicate age-related changes that were found in earlier studies, but not in Experiment 1, we increased the age range of the participants in Experiment 2. Finally, we considered a different combination of stimulus dimensions to establish that the absence of evidence for other than dimensional rules does not depend on this aspect of the items.

Experiment 2

Experiment 2 was again conducted using a paper-and-pencil test. However, there were two important differences with Experiment 1. First, the test consisted of two parts, each comprising 24 items. In the first part, the differences between the dimensional values of stimuli were relatively small. In the second part, the differences were relatively large. With this manipulation, we want to establish whether the number of participants who muddled through depends on the ease with which the differences between the stimuli can be distinguished. The second difference with Experiment 1 concerned the use of stimulus dimensions. We chose stimulus dimensions with known just notable differences, namely size and orientation. Finally, the age range was increased to be comparable to the age range in earlier studies (e.g., Thompson, 1994).

Method

Participants

We tested 357 participants: 311 children from a primary school, and 46 adults from an adult education center in the Netherlands. Due to their large number of missing values 24 participants were excluded from the analysis. These participants were 7-years of age or younger. Moreover, information about the age of 5 participants was missing. This resulted in a sample of 328 participants, distributed over age as follows: 22 4-years of age ($M = 4$ years 6 months, $SD = 2.9$ months), 23 5-years of age ($M = 5$ years 6 months, $SD = 2.9$), 33 6-years of age ($M = 6$ years 5 months, $SD = 3.4$), 33 7-years of age ($M = 7$ years 7 months, $SD = 3.5$), 34 8-years of age ($M = 8$ years 6 months, $SD = 2.8$), 40 9-years of age ($M = 9$ years 5 months, $SD = 3.7$), 33 10-years of age ($M = 10$ years 5 months, $SD = 3.1$), 32 11-years of age ($M = 11$ years 4 months, $SD = 3.0$), 37 12-years of age ($M = 12$ years 5 months, $SD = 5.3$), 26 18 to 50-years of age ($M = 30$ years 11 months, $SD = 11$ years and 0.1 months), 15 50-years of age or more ($M = 60$ years 11 months, $SD = 5$ years 3.2 months).

Design and stimuli

The test used in the second experiment was similar to the test of the first experiment. The differences are limited to the stimulus dimensions used and the length of the test. Stimuli were constructed with two varying dimensions: size and orientation. Each stimulus consisted of a circle with a 1 cm long line pointing from the circumference of the circle towards the center (or crossing it if the circle radius is smaller than 1). Steps between two values on a dimension are based on just noticeable differences as established by Fernandez (1976) in kindergarten children.

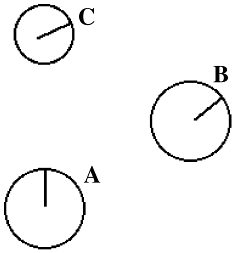


Fig. 7. A type Ia item of part 2 of the test in Experiment 2. The stimuli are, clockwise starting at the top, C B A.

Stimulus circles were designed in 7 sizes (diameters ranged from 1.76 to 4.00 cm, in steps of .28 cm, which is twice as large as the step values in Experiment 1), and 9 levels of orientation (orientations ranged from 90°, vertical line from the top downwards, and 157,36°, rotated clockwise in steps of 8.42°). Fig. 7 shows some of these stimuli.

We constructed items of type I, III, and IV. Type IV items were included, as stimulus dimensions were made comparable by using information on the just noticeable differences. The test consisted of two parts: stimuli with small, and stimuli with large dimensional differences. The first part of the test included type I items in which the one-dimensional difference between stimuli A and B was varied in three steps. The two-dimensional difference between stimuli B and C was varied in one step on both dimensions. The type III items differed in one step with respect to one dimension, and in three steps with respect to the other dimension. The type IV items varied in three steps with respect to both dimensions. In the second part of the test, all stimulus differences were twice as large. Both the first and the second part of the test included four items of each subtype Ia, Ib, IIIa, and IIIb, and eight type IV items.

The test booklet started with the same set of instruction items as in Experiment 1, which was followed by the 24 items of part 1, and the 24 items of part 2. Due to an error in the production of the booklets, three items were not identical in all the versions of the test, and were therefore excluded from the analysis (Part 1: one of type Ib, one of type IIIa, and part 2: one of type Ib). As in the first experiment, different versions of the test were created by varying triangle configuration, position of A, B, and C, and sequence of the stimuli (see Footnote 3). However, part 1 of the test always preceded Part 2.

Procedure

All participants were tested group-wise in the classroom setting. The average group size was 25. Instruction was equivalent to the instruction of Experiment 1. Immediately after the instruction, participants individually solved the 48 items in the test booklet. The test session was attended by students to make sure that the children did not copy each others work, or communicate during the completion of the booklet.

Results

Group-wise analysis

Considering only the 7 type I items, the proportion of AC classifications was higher in part 1 of the test ($M = .10$, $SD = .14$) than in part 2 of the test ($M = 0.07$, $SD = .11$; Wilcoxon: $Z(1) = -4.05, p < .001$). In both parts of the test, the number of AC responses decreased with age (Kruskal–Wallis: part 1: $\chi^2(10) = 42.40$, $p < .001$; part 2: $\chi^2(10) = 48.30$, $p < .001$). Since the exclusion of AC responses of participants amounts to an age-dependent selection of responses, we do not focus on the number of AB responses relative to the number of BC responses. The proportion of AB responses in type I items was lower in part 1 ($M = .34$, $SD = .19$) than in part 2 ($M = .45$, $SD = .20$; $Z(1) = -6.93, p < .001$). In part 1, the number of AB classifications did not vary significantly between age groups ($M = .34$, $SD = .19$). In part 2 differences between age groups were significant ($M = .44$, $SD = .20$; $\chi^2(10) = 44.34$, $p < .001$; results retain significance for the 4- to 12-year-old children only; $\chi^2(8) = 20.10, p = .01$) with the older participants giving more AB responses.

Thompson's rule matching analysis

Given type I, type III, and type IV items, one can distinguish six different rules, as shown in Fig. 3. Table 4 shows the number of participants that was assigned to each rule by rule matching analysis. The response patterns of most participants were quite consistent with a dimensional rule. In both part 1 and 2, PC-values were higher for the dimensional rule than for the other rules (Part 1: $F(5, 327) = 30.55, p < .001$; part 2: $F(5, 327) = 26.14, p < .001$). There was a main effect of part of the test on PC values ($F(1, 317) = 80.86$, $p < .001$), indicating that the second part of the test resulted in higher PC-values. As expected, the rules were adhered to more consistently when the stimuli were easier to distinguish (i.e., in part two). In contrast to the first experiment, there was also a main effect of age on PC values ($F(10, 317) = 6.69$, $p < .001$). Contrast analysis shows that the mean PC value of age groups 4, 5, 6, and 18 were lower than the mean PC values of respective older age groups. There was no interaction effect of age by part of the test.

Table 4
Percentage of participants classified with rule-matching analysis and corresponding average PC-values for participants in Experiment 2

Rule	Part 1		Part 2	
	%	PC	%	PC
Hol. orientation	12	.63 (.11)	5	.65 (.15)
Hol. size	8	.64 (.15)	2	.58 (.07)
Iden. orientation	2	.59 (.12)	9	.76 (.18)
Iden. size	4	.62 (.15)	5	.70 (.16)
Dim. orientation	41	.75 (.15)	43	.81 (.14)
Dim. size	32	.77 (.17)	35	.83 (.17)
N	333	.72 (.16)	333	.79 (.17)

Note. Standard deviations are between brackets.

Latent class analysis—part 1

We conducted LCA separately for each item type. The results were comparable to Experiment 1, and are therefore not discussed in detail. The majority of participants produced homogeneous responses to most of the items. As observed in Experiment 1, a subgroup of participants did not respond homogeneously to some item types. To model rule-use using LCA, we selected, from each item type, the item that elicited the smallest number of AC responses. Table 5 shows the goodness-of-fit and the selection of LC models for the chosen combination of items. Without parameter constraints, 4-classes were needed to model the data adequately. This model included two classes that resembled dimensional rules, and two classes that were characterized by ambiguous patterns of conditional probabilities. To further simplify the models, we subjected parameters in both 3-class and 4-class models to equality constraints similar to equality constraints applied in the LC models of Experiment 1. The resulting simplest, best fitting model, according to the BIC, is a 3-class model with equality constraints (described in the note of Table 5). We do not discuss the other poorly fitting models.

The resulting model (see Fig. 8, left panel) includes 3 classes. The comparison with Fig. 3 shows that one class (unconditional probability is .48) is characterized by the dimensional-orientation rule; a second class (probability .34) is characterized by the dimensional-size rule; and the third class (probability .18) does not appear to be characterized by any discernable rule. Moreover, conditional probabilities of giving AC responses in the latter class were relatively high, and standard errors of conditional probabilities were relatively large (in the range of .12 and .14, compared to the range of .02 and .06 in the other classes). We conclude that the first part of Experiment 2 provides support only for the use of dimensional rules. Moreover, a fairly large group of participants muddled through.

Latent class analysis—part 2

Results of LCA of part 2 were largely equivalent to the results of part 1 of the test. Despite the better distinguishable stimulus values, and despite the higher average

Table 5
Goodness-of-fit and selection of latent class models on the first part of the test of participants in Experiment 2

# Classes	L^2	pb(L^2)	N	df	BIC
2	174.65	<.001	249	221	1985.64
3	131.74	.039	249	214	1981.35
4	113.19	.107	249	208	1995.90
4a	133.25	.140	249	211	1999.41
3a	161.56	.150	249	221	1972.55
3b*	149.91	.060	249	219	1971.93

Note. The asterisk indicates that this model is the most parsimonious, best fitting model. Other symbols are explained in the note of Table 2. Model 4a is a 4-class model with equality constraints for the two-dimensional rule classes for items III and IV. Model 3a is a 3-class model with equality constraints for items Ia and Ib in one class implementing the dimensional-orientation rule and equality constraints for items IIIa and IV in the two classes implementing dimensional rules. Model 3b is a 3-class model with equality constraints for items IIIa and IV.

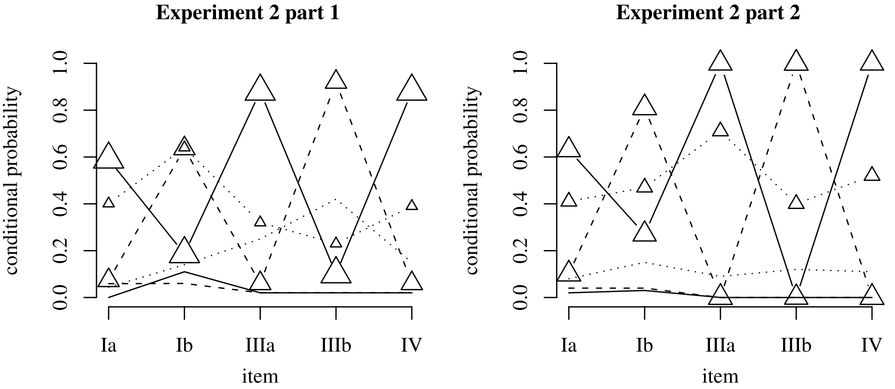


Fig. 8. (Left panel) Diagram of LC model 3-b of Table 5, which is an LC model of a combination of item types in part 1 of Experiment 2. (Right panel) The LC model of a combination of item types in part 2 of Experiment 2, which is equivalent to the selected model of part 1 (same number of classes and same equality constraints on conditional parameters). The lines with triangles denote the conditional probabilities of BC responses. The lines without triangles denote the conditional probabilities of AC responses. The plain lines denote the dimensional-orientation (DO) class; The dashed lines denote the dimensional-size (DS) class; The dotted lines denote the residual class (Res). The size of the triangles denote the unconditional parameters, that is, the proportion of participants that is modeled by that class.

PC-values item types in this part of the test gave rise to homogeneous responding in most, but not all, participants. Moreover, the LC models for each item type deviated from the anticipated two-class structure, as they included one or more additional classes. To fit a model to the responses of a combination of items, items of different types were selected on basis of the number of AC responses. The three-class model with equality constraints that were equivalent to the best LC model of part 1 (cf. Table 5) again had a relatively low BIC, but it did not fit the data very well ($L^2(221) = 148.36, p = .034$[6]). Nevertheless, we based further analyses on this model to ease comparisons with the result obtained with part 1. LC models with more than three classes were also estimated. These models consistently included two classes, which were characterized by dimensional rules. These classes were characterized by unconditional probabilities that were comparable to that of the three class model. Remaining subclasses could not be interpreted in terms of any expected rule.

Parameter estimates were comparable to the best LC model of part one (see Fig. 8, right panel). The class identified with the dimensional-orientation rule had a probability of .37, and the class identified with the dimensional-size rule had a probability of .38. The class that was not associated with a clear preference for any response (except for the responses on the item of type IIIa) had a probability of .25. Moreover, relatively high conditional probabilities for AC responses were observed in this last class, and standard errors of conditional probabilities were relatively large (in the range of .06 and .07 compared to the range of .00 and .048 in the other classes).

[6] Note that the *p* value is calculated by a Monte Carlo bootstrapping procedure (Langeheine et al., 1995).

Table 6
Cross table of number of participants classified by LCA in part 1 and part 2 of Experiment 2

Part 1	Part 2			
	DO	DS	MT	Total
Dim. orientation	84	11	20	115
Dim. size	4	69	11	84
Muddling through	6	14	15	35
Total	94	94	46	234

Note. DO, denotes dimensional-orientation rule; DS, denotes dimensional-size rule; MT, denotes muddling through.

The results of LCA support the conclusions of part 1 and Experiment 1, as there is no evidence for other than dimensional rules. Moreover, a considerable number of participants was found to muddle through. There was no significant difference between part 1 and part 2 of the test with respect to the probability of belonging to the class that was not characterized by a rule (part 1: probability is .18, $SE = .086$; part 2: probability is .25, $SE = .030$; $z = 1.32, p = .094$).

The assignment of participants on the basis of posterior probabilities to the classes in part 1 and in part 2 of Experiment 2 agreed reasonably well. Table 6 shows the cross-tabulation of the assignments of participants. The numbers in the off-diagonal positions indicate the number of rule switches between the two parts of the test. These results suggest that the application of the dimensional-orientation rule and the dimensional-size rule was most stable between part 1 and part 2 of the test. Furthermore, participants who muddled through during one part of the test mostly applied a dimensional rule in the other part.

Age-related changes

Fig. 9 shows the observed probabilities of rule use by age, and the predicted probabilities of the multinomial logistic regression model for children up to age 12. The

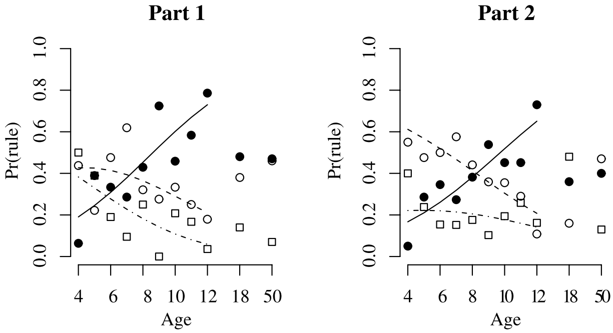


Fig. 9. Results of multinomial logistic regression analysis of the probability of rule use by age (up to 12 years of age). Solid line is the regression line of the probability of using the dimensional-orientation rule, solid dots are the corresponding observed probabilities. Dashed line and open dots correspond to the use of the dimensional-size rule. Dot-dashed line and squares correspond to the probability of muddling through.

multinomial logistic regression analyses indicated that the probability of using a rule did generally depend on age (Part 1: $\chi^2(2) = 22.5$, $p < .001$; part 2: $\chi^2(2) = 24.6$, $p < .001$). In particular, among the rule users, i.e., those who applied the dimensional-size and dimensional-orientation rule, the probability of using the dimensional-orientation rule increased with age (Part 1: $B = .26$; $\chi^2(1) = 10.9$, $p < .001$; part 2: $B = .31$; $\chi^2(1) = 23.3$, $p < .001$). The probability of using a rule, as opposed to muddling through, changed with age in part 1 ($B = .55$; $\chi^2(1) = 11.6$, $p < .001$), but did not change with age in part 2 ($B = .15$; $\chi^2(1) = 1.3$, $p = .24$). Apparently, the small differences in stimulus values in the first part of the test hindered consistent judgment of the stimuli in the young participants. Just noticeable differences are known to vary with age (Fernandez, 1976). Participants over 18 years of age deviated from the observed pattern: they used the dimensional-orientation rule relatively less often.

Discussion

Standard findings in literature were replicated in Experiment 2. We found an age-related decrease of the number of AC responses for type I items in both part 1 and part 2 of the test. We observed an increase in the number of AB responses only in part 2 of the test. Hence, the standard finding of a decreasing number of holistic responses (i.e., AB responses) with age was replicated only with sufficiently distinguishable stimuli. This finding is not directly predicted by a rule model that only incorporates dimensional rules. However, it may be due to the relative difficulty of producing consistent AB responses compared to producing consistent BC responses in applying a dimensional rule. As in Experiment 1, we compared the consistency of expected AB responses with the consistency of expected BC responses to type I items. We divided all the participants in two groups by matching each individual response pattern with the dimensional-size and the dimensional-orientation rule. Each participant is assigned to the rule with the best match. We applied a MANOVA with part of the test (i.e., part 1 or part 2) and response type (consistency of AB responses and consistency of BC responses) as within subject factors, and age as the between subject factor.[7] We obtained the following results: there was a main effect of response type, which means that BC responses were more consistent than AB responses ($F(1,317) = 77.1, p < .001$). There was an interaction effect between response type and age ($F(10,317) = 3.6$, $p < .001$), meaning that the difference in consistency between AB and BC responses decreased with age. There was an interaction between response type and part ($F(1,317) = 35.9$, $p < .001$), meaning that the difference in consistency between AB and BC responses was largest in part 1. Finally, there was a three-way interaction between part, response type, and age ($F(10,317) = 2.7$, $p = .004$). This means that in part 1 the age-related decrease of the difference in response type was larger than in part 2. These results suggest that the relative difficulty of consistent AB responses (i.e., relative to consistent BC responses) depended on the salience of dimensional differences. With greater salience, comparing identity to one step

[7] We also analyzed the participants modeled by the two-dimensional rule classes of the LC models depicted in Figs. 8 and 9. Results are consistent with the reported MANOVA, except for the three-way interaction, which was not significant.

difference became easier (i.e., the number of consistent AB responses increased) relative to comparing one step to three steps. The salience of dimensional differences increases with age (e.g., Cook & Odom, 1992; Fernandez, 1976). This would explain the classical finding of an increase of the number of AB responses with age. However, it also shows that the effect depends on the size of differences of dimensional values used in the test and on the age range. This might explain the fact that this classical result was not always found (Experiment 1, Experiment 2 part 1, and Thompson's (1994) Experiment 2).

Following Thompson's rule analysis, the response patterns of the majority of the participants were classified as consistent with a dimensional rule. Furthermore, PC-values increased with age. However, in these analyses the PC-values were calculated on basis of the six different rules. Since we only found evidence for two interpretable rules, we recalculated PC values of response patterns only in relation to the dimensional-size and the dimensional-orientation rule. Subsequently, we again found an effect of part of the test (PC values of part 1 are lower than for part 2, $F(1,317) = 130.9$, $p < .001$), and an effect of age (PC values increase with age; $F(10,317) = 7407, 1, p < .001$). The interaction of part by age was found to be absent.

LCA showed that most participants followed dimensional classification rules. In addition, the response patterns of a considerable number of participants were not consistent with the application of a rule, i.e., these participants appeared to muddle through. The size of this group was equal for both parts of the test. In contrast to the participants who applied the dimensional rules, the participants in this residual group did not display very consistent behavior. Specifically, the participants in the residual group in one part of the test mostly followed a rule in the other part. In part 1 of the test, but not in part 2, the size of the residual group decreased with age. The salience hierarchy of dimensional differences was related to age: relative to the dimensional-size rule, the probability of using the dimensional-orientation rule increased with age for both parts of the test.

Discussion

Thompson (1994) and others (e.g., Wilkening & Lange, 1987) assessed rule-use in perceptual classification on the triad classification task by considering individual response patterns. Thompson (1994) developed an ingenious experimental design to assess rule-use in triad classification task performance using only the responses on triad classification items, i.e., without the requirement of verbal justifications. These innovations have resulted in radically different conclusions about the development of perceptual classification than those based on the group-averaging procedures (e.g., Smith & Kemler, 1977). Thompson (1994) found evidence for the differential-sensitivity theory of Cook and Odom (1992). Her main conclusions are that (1) children mostly use dimensional rules in perceptual classification, and that there is very little evidence that children apply holistic or identity rules; (2) consistency in rule-use increases with age; (3) there is some evidence for changes in the relative salience of dimensional relations. Adults readily adopt a dimensional-brightness rule in favor of a dimensional-size rule, whereas children between 4 and 8 years of age (using the same stimuli) do the opposite.

Thompson's statistical analysis of individual response patterns leaves some important questions unanswered. First, how can one identify the best fitting, but parsimonious, rule model that can account for the children's classification task performance? Second, do children and adults use other rules or response strategies than those expected on the basis of theoretical considerations? Third, how reliable are the statistical results obtained by Thompson? These issues appeared to be important in applying the rule-assessment methodology of Siegler (1981), which is similar to Thompson's rule-matching analysis. Jansen and van der Maas (1997) showed that with Siegler's rule-assessment methodology rules are wrongfully detected in computer generated data relating to the balance scale task. Although these findings were obtained using the balance scale task, and Thompson's and Siegler's methods differ in some respects, the issue of reliability requires careful consideration. Finally, the results in most triad classification task studies are based on small numbers of participants. The study of age-related changes in individual differences, especially changes in relative salience of dimensional relations, requires larger samples.

To address these issues, we conducted two empirical studies based on the experimental design of Thompson. The two studies involved two different combinations of stimulus values, size–brightness (Experiment 1) and size–orientation (Experiment 2). In contrast to Thompson, we used LCA to analyze the rules underlying responding to triad classification items. LCA provides a statistically reliable way to assess rules in categorical data (Jansen & van der Maas, 1997). The fit of a rule model is tested on the whole data set, such that the best fitting, most parsimonious rule model can be selected on basis of statistical criteria. Moreover, if LCA is applied in an exploratory fashion, no rules have to be specified beforehand, i.e., we can let the data "speak for themselves." Hence, using LCA one can detect unanticipated rules, and/or identify groups of participants who do not follow a clear rule. Finally, LCA models latent structures in the data, which avoids possible problems of other rule assessment methodologies (Jansen & van der Maas, 1997; Strauss & Levin, 1981).

Apart from these advantages, LCA does place certain demands on the data. First, a relatively large number of participants is needed to obtain reliable parameter estimates. To ascertain adequate samples, we administered a paper-and-pencil test group-wise, instead of testing subjects individually, as is common practice in triad classification tests. Second, one is necessarily limited to a relatively small number of items to fit LC models (but see De Soete, 1993). Hence we performed two kind of analyses. First, we fitted models per item type to examine the homogeneity of item types. Second, in the detection of rules, we fitted models on a combination of the most reliable items of each type. In addition, we reduced the number of item types by excluding the redundant and less reliable type II items.

Replication and explanation of earlier findings

In both experiments the data obtained with the paper-and-pencil test proved to be comparable to the data obtained using an individually administrated test

(Thompson, 1994; Thompson & Markson, 1998) in terms of the average PC-values (i.e., the consistency with six expected rules) and the number of AC responses (i.e., erroneous responses). Moreover, we largely replicated the basic findings observed in triad classification task performance. In Experiment 2, the number of AC responses decreased with age and the number of AB responses increased with age in part 2 of the test. Moreover, the average proportion of AB responses in part 2 of the test was similar to, for example, Thompson's (1994) findings of Experiment 1. However, the average proportions of AB responses in Experiment 1 and in part 1 of Experiment 2 were smaller than the proportions reported by Thompson (1994) and others (e.g. Smith & Kemler, 1977). This might be explained by a low salience of dimensional differences of stimuli. We found that in applying a dimensional rule to type I items, the consistency of expected AB responses differed from the consistency of expected BC responses. Moreover, the consistencies and the differences in consistencies were dependent on the salience of dimensional differences and, as is consequently expected, also on age. This would explain the standard finding that the number of AB responses on type I items increases with age. Although we tried to replicate Thompson's Experiment 1 stimuli as closely as possible in our Experiment 1, it is possible that differences in brightness and size as presented on a computer screen appeared to be different than on paper. Importantly, in spite of the relatively high number of BC choices, response patterns were inconsistent with holistic rules.

Usefulness of LCA

The results of the LCA demonstrated the importance of using stringent statistical criteria in attributing individual response patterns to the use of rules. In general, the agreement in the identification of individuals as specific rule users based on LCA and based on Thompson's rule analysis was moderate: 58% of the assignments to the dimensional rules in Experiment 1 matched. If we equate holistic and identity rules with muddling through, the match between these methods is 75%. In our two experiments we found evidence for the use of dimensional rules, but not for the use of either holistic or identity rules. In spite of the presence of individual response patterns that matched best with holistic or identity rules according to rule matching analysis, none of the selected LC models contained latent classes that could be interpreted in terms of holistic or identity rule use. The best fitting, most parsimonious rule models included three classes. In both Experiment 1 and Experiment 2, two classes of response patterns were consistent with the dimensional rules: the dimensional-size and the dimensional-brightness rule in Experiment 1, and the dimensional-size and the dimensional-orientation rule in Experiment 2. The conditional probabilities in the third, residual class were hard to interpreted in terms of rule use. In this class, the probability of choosing an AB or BC response was about equal, whereas the probability of AC responses was relatively low. Possibly, participants in this class switched erratically between rules, but they may also have engaged in guessing in choosing between AB and BC responses. We denoted this group as the muddling through class. This third strategy (if the term is appropriate) could not have be detected using averaging techniques or rule analysis. The presence of this class appears to be robust, because it includes between 18 and 40% of the participants, and it was detected under different task conditions and in different age groups.

Developmental trends

Given, the relatively large samples, we were able to consider age-related changes in classification. The most important developmental phenomenon concerned the stimulus dimension on which participants based their classifications. In Experiment 1, the proportion of participants using the dimensional-size rule increased with age, relative to the proportion of participants who used the dimensional-brightness rule. In Experiment 2, the proportion of participants using the dimensional-orientation rule increased between 4 and 12 years of age, relative to the proportion of participants who used the dimensional-size rule. Thompson (1994) found that adults used the brightness rule more often than the size rule, whereas in children the reverse was observed. These age differences do not agree with our findings. However, in contrast to us, Thompson did not find differences in the salience hierarchy of dimensional differences between children of different ages. In our Experiment 2, adults deviated from the developmental pattern of children between 4 and 12 years of age, in that they relatively more often favored the dimensional-size rule to the dimensional-orientation rule. Thompson also observed qualitative differences between adults and children, which makes it difficult to compare results of children and adults. Our results agree with the literature, in which the salience of dimensional differences is found to be related to increasing age in the following order: brightness, size, and orientation (Cook & Odom, 1992; Fernandez, 1976; Lipák, Szombati, & Kleininger, 1976).

A second developmental trend concerned the size of the muddling through group. It appeared that given small differences in dimensional values of the three stimuli, the size of the muddling through group decreased with age. This is in agreement with Fernandez's (1976) finding that just noticeable differences decrease with age. With larger differences of dimensional values the proportion of participants that did not follow a clear rule, did not vary with age between 4 and 12 years of age. Therefore, it is unlikely that the members of the muddling through group did not perceive the dimensional differences.

Muddling through

We found that, under all test conditions, the observed behavior of a considerable group of participants showed no consistency with expected behavior (i.e., rule use). We called this group the *muddling through* group. Consistency is an important criterion to establish rule use (Reese, 1989). However, the lack of consistency does not mean that the behavior in this group is random. Notably, this group produced relatively few AC responses (i.e., erroneous responses). One possible explanation for the behavior of this group is that individuals unsystematically switched between rules during the course of the experiment. This would result in unsystematic patterns of responses AB and BC. Switching between rules or processing strategies during development is predicted by fuzzy-trace theory (Brainerd and Reyna, 2001). Brainerd and

Reyna (2001) argue that both children and adults have two reasoning systems that operate in parallel: an analytical reasoning system and an intuitive reasoning system. Both these systems improve with age, but they continue to coexist, i.e., the intuitive system is not replaced by the analytical system. Hence, according to this theory, a switch between reasoning systems during the course of the experiment is possible in adults and children alike. Moreover, switching strategy is especially likely to occur in the task setting of the triad classification task, because there is no optimal strategy. The holistic-to-analytic shift theory also allows for switching between strategies, but views this as a temporary consequence of the immaturity of the analytic reasoning system. In contrast to the holistic-to-analytic shift theory, the fuzzy-trace theory would not necessarily predict an developmental decrease of switching strategies. A developmental decrease in the size of the muddling through group was found only with low salience stimulus differences. The latter suggests a developmental trend in the perception of stimulus differences, and not in possible rule switches. However, the existence of two qualitatively different reasoning systems is not confirmed by the present studies, because evidence for the application of holistic rules is lacking. Hence, in the current studies it is more plausible that participants switched between two-dimensional rules (i.e., two analytical rules), instead of between a holistic rule (i.e., an intuitive rule) and a dimensional rule (i.e. an analytical rule). Moreover, in Experiment 2 about half of the subjects who muddled through in one part of the test applied a dimensional rule in the other part. Note that this does not question the existence of an intuitive reasoning system in general. Rather lacking evidence for a holistic reasoning system is limited to the perception of multi-dimensional objects in the context of the triad classification task.

A second possible explanation for the existence of a muddling through group lies also in the nature of a free classification task. Wilkening and Lange (1987) stressed that in the triad classification task the purpose of the classification is not well-defined for participants. This may result in other behavior than the expected rule-like behavior. Specifically, they found that participants are less likely to exhibit rule-like behavior, if their response time is not limited. Also in our study, participants might have entertained meta-cognitive considerations about the aims of the experiment, which interfered with purely perceptual classifications. This could possibly also explain why adults deviated from the developmental pattern. This hypothesis can be investigated using a speeded, free classification task, such as in Smith and Kemler Nelson (1984), and Wilkening and Lange (1987). However, the experimental set-up should allow for the analysis of individual response patterns by LCA, as in the present article. The application of a statistically advanced and stringent method of rule assessment, such as LCA, is required to obtain a reliable description of individual rule use and development thereof.

Differential-sensitivity account

Our empirical results support the differential-sensitivity theory of Cook and Odom (1992). First, we showed that a considerable number of participants, both adults and children, ranging in age from 4 to 12 years, used one-dimensional classification rules.

No evidence was found for the use of holistic rules or identity rules. Also the dependency on age of the proportion of responses that are consistent with rule use is consistent with the differential-sensitivity theory. The results of the present study indicate that the participants became increasingly more consistent in following rules between 4 years of age and adulthood. Finally, the relative salience of dimensional differences changed with development. The dominant dimension over the period of 4–12 years was first brightness, subsequently size, and finally, towards the age of 12, orientation.

References

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception and Psychophysics, 61*, 1178–1199.

Azzelini, A. (1996). *Statistical inference based on the likelihood.* London: Chapman and Hall.

Boom, J., Hoijtink, H., & Kunnen, S. (2001). Rules in the balance: Classes, strategies, or rules for the balance scale task? *Cognitive Development, 16*, 717–735.

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. In: H. W. Reese & R. Kail (Eds.), *Advances in child development and behavior* (pp. 41–100). San Diego, CA, US: Academic Press.

Chletsos, P. N. (1986). *A paper-and-pencil test replicating Siegler's rule-assessment approach on Piaget's balance beam task. Instruction Manual.* New York: Rutgers University.

Clogg, C. C. (1994). Latent class models. In G. Arminger (Ed.), *Handbook of statistical modeling in the behavioral and social sciences.* New York: Plenum.

Cook, G. L., & Odom, R. D. (1992). Perception of multidimensional stimuli: A differential-sensitivity account of cognitive processing and development. *Journal of Experimental Child Psychology, 54*, 213–249.

De Soete, G. (1993). Using latent class analysis in categorization research. In I. vanMechelen, J. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: theoretical views and inductive data analysis.* London: Academic Press.

Dobson, A. J. (2002). *An introduction to generalized linear models* (2nd ed.). London: Chapman and Hall.

Fernandez, D. (1976). Dimensional dominance and stimulus discriminability. *Journal of Experimental Child Psychology, 21*, 175–189.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton-Mifflin.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Thousand Oaks, CA: Sage Publications.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression.* New York: Wiley.

Jansen, B. R. J., & van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321–357.

Kemler Nelson, D. G. (1989). The nature and occurrence of holistic processing. In B. E. Shepp & S. Ballesteros (Eds.), *Object perception: Structure and process.* Hillsdale, NJ: Erlbaum.

Langeheine, R., Pannekoek, J., & Van de Pol, F. (1995). *Bootstrapping goodness-of-fit measures in categorical data analysis.* The Netherlands: CBS Statistics.

Lipák, J., Szombati, G., & Kleininger, O. (1976). Preference of visual discrimination factors in childhood. *Studia Psychologica, 18*, 292–306.

McCutcheon, A. L. (1987). *Latent class analysis.* Beverly Hills: Sage.

Reese, H. W. (1989). Rules and rule-governance: Cognitive and behavioristic views. In: S. C. Hayes (Ed.), *Rule governed behavior: Cognition, contingencies, and instructional control* (pp. 3–84).

Rindskopf, D. (1987). Using latent class analysis to test developmental models. *Developmental Review, 7*, 66–85.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Siegler, R. S. (1978). The origins of scientific reasoning. In R. S. Siegler (Ed.), *Children's thinking: what develops?* (pp. 109–149). Hillsdale, NJ: Lawrence Erlbaum.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46*.

Siegler, R. S., & Chen, Z. (2002). Development of Rules and Strategies: Balancing the Old and the New. *Journal of Experimental Child Psychology, 81*, 446–457.

Smith, J. D., & Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology, 24*, 279–298.

Smith, J. D., & Kemler Nelson, D. G. (1984). Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology, General, 24*, 279–298.

Smith, L. B. (1989). A model of perceptual classification in children and adults. *Psychological Review, 96*, 125–144.

Strauss, S., & Levin, I. (1981). Commentary. *Monographs of the Society for Research in Child Development, 46*.

Thompson, L. A. (1994). Dimensional strategies dominate peceptual classifications. *Child Development, 65*, 1627–1645.

Thompson, L. A., & Markson, L. (1998). Developmental changes in the effect of dimensional salience on the discriminability of object relations. *Journal of Experimental Child Psychology, 70*, 1–25.

Ward, T. B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 103–112.

Wilkening, F., & Anderson, N. H. (1982). Comparison of two rule-assessment methodologies for studying cognitive development and knowledge structure. *Psychological Bulletin, 92*(1), 215–237.

Wilkening, F., & Lange, K. (1987). When is children's perception holistic. Goals and styles in processing multidimensional stimuli? In T. Globerson & T. Zelniker (Eds.), *Cognitive style and cognitive development*. Norwood, NJ: Ablex.