The Roles of Labelling and Abstraction in the Development of Cognitive Flexibility

by

Sophie Jacques

A thesis submitted in conformity with the requirements

for the degree of Doctorate of Philosophy

Graduate Department of Psychology

University of Toronto

0-612-63777-8

The Roles of Labelling and Abstraction in the Development of Cognitive Flexibility

Doctorate of Philosophy. 2001

Sophie Jacques

Department of Psychology. University of Toronto

Three experiments were conducted to determine whether language contributes to the emergence of cognitive flexibility in preschoolers. The Flexible Item Selection Task (FIST). a measure of flexibility. was developed in Exp. 1 for use with preschoolers. On each trial. children were shown 3 items (e.g.. a big red shoe. a big blue shoe. a small blue shoe) and asked to select 2 items that matched on one dimension (Sel. 1: e.g.. size). and then to select a second pair that matched on another dimension (Sel. 2: e.g.. colour). The results revealed that 2-year-olds did not understand task instructions and 3-year-olds performed poorly on both selections. In contrast. 4-year-olds did well on Sel. 1. but they did significantly worse than 5-year-olds on Sel. 2. suggesting that they had specific problems with switching flexibly between dimensions. It was then hypothesized that changes in flexibility between 4 and 5 years might be due to underlying changes in language. Exp. 2 was conducted to test this claim by using label manipulations on the FIST. The results revealed that Sel. 2 performance was not only correlated with receptive language development. but it was also influenced by labelling on Sel. 1. When 4-year-olds were asked to label the dimension on Sel. 1 that was relevant to that selection (the dimension on which items matched; e.g.. size). their Sel. 2 performance improved significantly compared to children who were not asked to label. or who were asked to label the irrelevant dimension (the dimension that did not vary across the 3 items; e.g.. shape). Also, children in the relevant-label condition who tended to label correctly on Sel. 1 also did better on Sel. 2 than children in that condition who made many labelling errors. Exp. 3 was then conducted to determine which kinds of labels improve 4-year-olds' performance on Sel. 2. In Exp. 3, the experimenter selected items on Sel. 1 and labelled them in some predetermined way. Results showed that labels that referred to the relevant dimension helped 4-year-olds on Sel. 2 whereas labels that referred to the irrelevant dimension did not. This pattern held whether the experimenter labelled the dimension (e.g., "size") or the cue (e.g., "big"). In general the results are consistent with the hypothesis that the emergence of flexible thinking in the preschool years may be mediated by language development.

Acknowledgments

I must first thank my supervisor, Phil Zelazo, for his guidance, support, and commitment throughout these years. Phil, you have taught me the skills that I will need to develop my own research program, and I sincerely appreciate it. I also thank my committee members, Bob Lockhart and Charles Helwig, for their patience and understanding, and for going well beyond the call of duty by reading and making very useful comments on drafts with little advanced notice.

Many thanks go to Olena Bartkiw, Jodie Burton, and Stella Lourenco for their invaluable assistance with data collection and participant recruitment. I am also grateful to Adrian Liu, Laura McGrath, Victoria Orekhovsky, and Alexandra Sutherland, as well as to several others in the lab for their assistance on various aspects of these experiments. Thanks also to friends in the CSC (Duane, Janet, Keith, Norma, Stella, Ulrich, & particularly, Stuart, my "small" brother), friends at Knox, both "the old boys" (Brad, Daniel, David, Darren, Joe, John, Louisa, Mike, Sean, & Wael) and "the new boys" (Ajay, Bill, & Walter), and friends at home (Chantal, Julie, Fanny, & Louise). Thank you all for doing the impossible job of attempting to keep me relatively sane throughout the process (at least that is what I think you were trying to do. . .). Thanks for forcing me to have somewhat of a social life and for all the laughs!

Finally, I extend my warmest thanks to my family: to my parents, Magella and Paulette, my brother, Denis, his wife, Karine, and their children, Anthony, Myriam, and Gabriel, for their constant support and kindness throughout these years. Un gros merci à vous tous! Sans votre aide je n'aurais jamais pu finir mes études. Je vous dois beaucoup.

~~**Dedication**~~

To my father,

Magella Jacques

(1941 - 1998)

Cher papa,
La maladie t'a empêché de réaliser ton rêve et
la mort t'a empêché de me voir réaliser le mien.

Mais où que tu sois, papa, j'espère que tu sais
que cette fois la maladie n'a pas triomphé.

J'ai pu réaliser enfin notre rêve et
je sais que tu serais fier.
Sophie

Table of Contents

CHAPTER I

CHAPTER II

List of Tables

List of Figures

List of Appendices

CHAPTER I

INTRODUCTION

LANGUAGE, ABSTRACTION, AND COGNITIVE FLEXIBILITY

> *Thus, the same colour being observed to-day in chalk or snow, which the mind yesterday received from milk, it considers that appearance alone, makes it a representative of all of that kind; and having given it the name whiteness, it by that sound signifies the same quality wheresoever to be imagined or met with; and thus universals, whether ideas or terms, are made.*
>
> **- John Locke (1690/1875; Book II, Chapter XI, Section IX)**

1.1. Development of Abstraction and Cognitive Flexibility

Several researchers have argued that the preschool period is marked by fundamental changes in abstraction, the ability to consider part properties or features of objects without regard to their instantiation within particular concrete stimuli (e.g., Smith, 1989a, Werner, 1948). By definition, abstract representations are detached from the external representations on which they are based, and as a result, the thought processes that depend upon such representations should be relatively flexible compared to processes that rely upon concrete representations. In other words, a direct consequence of forming abstract representations of objects and events is that such representations allow individuals to distance themselves from the information available in the immediate concrete environment (i.e., the "here and now"), thereby permitting them to represent multiple alternate realities simultaneously and to manipulate information cognitively with relative ease. On this account, then, abstract representations should be a prerequisite for the development of cognitive flexibility, the ability to consider simultaneously multiple conflicting representations of one object or event, and to be able to act differentially on the basis of each representation of the same object or event. In a related vein, Werner (1948) held that cognitive rigidity (i.e., the opposite of cognitive flexibility) occurs in instances in which there is a failure to form abstract representations. In his words, "rigidity and lack of plasticity in motive and goal are grounded in a comparative lack of polarity between the subject and the world" (p. 211).

Perhaps changes in cognitive flexibility, then, are due to an increase in abstraction that occurs at around the same point in development. In fact, it may not be a coincidence that the end of the preschool period is also marked by widespread changes in several domains on tasks that require children to entertain and manipulate multiple representations

~~simultaneously. For example, at around 4 or 5, children improve significantly on executive~~ function tasks such as the Dimensional Change Card Sort (Frye, Zelazo, & Palfai, 1995), the day-night Stroop task (Gerstadt, Hong, & Diamond, 1994), and the windows task (Russell, Mauthner, Sharpe, & Tidswell, 1991); as well as on standard theory of mind tasks such as false belief (e.g., Wimmer & Perner, 1983), appearance-reality (e.g., Flavell, Flavell, & Green, 1983), and representational change (e.g., Gopnik & Astington, 1988) tasks. In order to perform well on all of these executive function and theory of mind tasks, children must be able to switch flexibly between multiple perspectives; specifically, they must be able to withhold responding on the basis of a more salient or a more direct representation of reality, and instead respond in a manner consistent with a less salient or less direct representation. Hence, they must be flexible in how they represent reality and be able to govern their behaviour accordingly, even if it means acting in a way that contradicts how they might normally respond to a particular situation.

1.2. Language, Abstraction, and Cognitive Flexibility

Undoubtedly, language has a significant impact on human cognition. Whorf (1956) went so far as to affirm that the way in which humans analyse and understand their environment, and the way in which they act in relation to this environment are a direct consequence of the language they speak. More relevant to the development of cognition itself, Vygotsky (1929, 1934/1986, 1978) proposed that human cognitive development occurs as a result of two independent lines of development–namely, the natural line of development in which maturation of the nervous system leads to developmental changes in basic cognitive processes, such as memory or attention, on the one hand; and the cultural line of development

~~in which children not only learn about culturally~~ invented tools from their elders. but. by a process of internalization. they also learn how to appropriate these tools for themselves to control their own thoughts and behaviours internally. Language. on Vygotsky's account. is one of the most important of these cultural tools for the development of the human mind and it makes possible the development of unique cognitive processes that are unavailable to other animals. More recently. Katherine Nelson (1996) has elegantly explicated her own views on the critical role that language plays both in the development of early cognitive development. in general. and in the development of autobiographical memory. in particular.

Although most researchers agree that language has some impact on cognition. it is not entirely clear how best to characterize the precise nature of this impact and the specific processes responsible for its involvement. Some. like Vygotsky (1934/1986) and Whorf (1956). argue that language has an instrumental role in the development of cognition. whereas others. like Piaget (1964/1967). argue that language can certainly facilitate specific kinds of cognitive processes. but it is not essential for their emergence. For instance. Piaget concluded that.

> language and thought are linked in a genetic circle where each necessarily leans on the other in interdependent formation and continuous reciprocal action. In the last analysis. both depend on intelligence itself. which antedates language and is independent of it (p. 98).

Irrespective of the position that one takes in regards to the relation between language and cognition. however. there is no doubt that language relates to abstraction. Indeed. language and abstraction are sometimes treated as the same. Although it is true that language–by its very nature as a symbolic medium–conveys information abstractly. it does not necessarily follow that all abstract representations are linguistic in nature (cf. Werner &

Kaplan, 1963), a point I shall return to in Chapter V. Perhaps, then, changes in language–or more precisely, in the use to which children put language (i.e., by using linguistic representations to represent information abstractly)–account for the development of cognitive flexibility. Bruner (1973), in fact, has suggested just that:

> In effect, language provides a means, not only for representing experience, but also for transforming it. . . . Once the child has succeeded in internalizing language as a cognitive instrument, it becomes possible for him to represent and systematically transform the regularities of experience with far greater flexibility and power than before (p. 330).

Even Piaget (1964/1967) admitted a role for language in the expression of flexible thought, "language confines itself to profoundly transforming thought by helping it to attain its forms of equilibrium by means of a more advanced schematization and a more mobile abstraction" (pp. 91-92).

Some support–albeit only indirect–exists for the hypothesis that language may play a determining role in the development of cognitive flexibility. First, Glucksberg and Weisberg (1966) introduced a labelling manipulation with adults on Duncker's (1945) candle problem, a classic measure of cognitive flexibility in adults. In the candle problem, adults are presented with a candle, matches, and a box containing tacks, and asked to affix the candle vertically against a wall, light it, and ensure that it not drip wax on the table or on floor. To succeed on the candle problem, adults must first empty the box, affix it on its side to the wall using one of the tacks, and then place the candle on top of the box and light it. Adults often fail to solve the problem, or at least require a substantial amount of time to do so, because they fail to use the box as a platform for the candle. The oversight in using the box is believed to result from a "functional fixedness" problem: Adults perceive the box only in terms of its current function as a tack container, and because of their overwhelming tendency to fixate on the box's current function, they fail to entertain the possibility that it could serve another function (i.e., serve as a platform for the candle).

In each of three experiments, Glucksberg and Weisberg (1966) introduced the candle problem by showing adult participants a sheet of paper with pictures of the four objects involved in the task. For some participants, one of the objects (i.e., the tacks) or all four objects were identified by their respective written label (e.g., the word "tacks" appeared by the picture of the tacks, the word "box" appeared by the picture of the box). Adults who were provided with the written label for the box (i.e., the functionally fixed object) found the correct solution more rapidly and were less variable in their performance than those who were shown either no written labels or only the label for the tacks. On the basis of their findings, Glucksberg and Weisberg interpreted the inflexibility experienced by adults on this task somewhat differently from the commonly held view. That is, they argued that because adults failed to label the box spontaneously, they also failed to notice it as an object in its own right. It is not that they were fixed on the box's current function, and as a result, they could not see any other use for it. Rather, because of the box's current function, they failed to represent it as an object in its own right, and as a result, they failed even to consider it when attempting to solve the problem. In their own words, Glucksberg and Weisberg state that "it is not a function that is unavailable to S, but rather the functionally fixed object itself" (p. 659). On their account, the label for the box served simply to make it available to participants for them to use.

Irrespective of the specific interpretation of the candle problem that one favours, Glucksberg and Weisberg's results nicely demonstrate how labelling can influence

~~participants' tendency to demonstrate cognitive flexibility~~. ~~However interesting the results of~~ Glucksberg and Weisberg's (1966) experiments are, their experiments were done with adults–not with children. Hence, it remains to be determined not only whether labelling manipulations help induce flexibility in children as well, but also whether language might be responsible for the development of flexibility in children, more generally.

There exists some correlational support in at least two domains for a link between language and the development of flexible thought in preschool children. First, Bruner and Kenny (1966) conducted an experiment in which they assessed children on their ability to reproduce nontransposed and transposed versions of a matrix of nine clear-plastic beakers that varied on three possible heights and on three possible diameters. The beakers were arranged in a 3 x 3 matrix so that the beakers in the first row were all the same height, but were shorter than those in the middle row (which were all the same height), which in turn, were shorter than those in the last row. Similarly, the beakers in the first column were all the same diameter, but they were thinner than those in the middle row, which were in turn thinner than those in the last row. As a result, the beaker in the northeast corner was the tallest and thickest beaker, while the beaker positioned in the southwest corner was the shortest and thinnest beaker. After the beakers were scrambled, children either had to replace the beakers in their original position (i.e., the nontransposed version) or the experimenter put one of the beakers (e.g., the shortest and thinnest beaker) in a new position on the array (e.g., the northwest corner) and asked children to produce a transposed version of the original matrix. The transposed version was more difficult than the nontransposed version. This is perhaps not surprising given that the transposed version required that children determine the new correct positions of each of the individual beakers relative to the one(s) already in place

~~while simultaneously suppressing their tendency to place each beaker back in its original~~ position. In other words, to perform correctly on the transposed version, children needed to be flexible in how they represented the position of the beakers. Of particular interest were findings that the authors reported relating children's performance on the transposed version with the labels that children used to refer to the beakers. That is, prior to performing the task, the experimenter asked children to explain how beakers in different rows (and columns) differed from each other. Children referred to differences between beakers using either (a) precise dimensional terms that referred to specific dimensions (e.g., "short" vs "tall" for height or "skinny" vs. "fat" for width), (b) global, undifferentiated terms like "big" and "little" that did not differentially apply to one dimension or the other, or (c) a confounded description of the beakers that combined dimensional and global terms (e.g., "fat" vs "little" or "big" vs. "short"). The way in which children referred to the beakers in this preliminary phase of the experiment predicted how they did on the transposed version of the task. Those who used dimensional terms performed well, and according to the authors, if they used dimensional terms to refer to both height and width at the same time, then they performed even better. In contrast, children who confounded global and dimensional terms by using them together performed worse overall.

More recently, in the theory of mind literature, Astington and Jenkins (1999) found significant relations between language development and performance on standard theory of mind tasks including false belief and appearance-reality tasks, tasks that arguably require children to reason simultaneously from two points of view. More important, however, they also found that language development–particularly syntax development–at 40 months of age predicted children's ability to reason on standard theory of mind tasks at 44 and 47 months

~~even when the variance due to age and theory of mind performance at 40 months was taken~~ into account. However, the reverse was not true: Performance on theory of mind tasks at 40 months did not predict scores on later language measures. Although these results are interesting and are consistent with the hypothesis that language directly affects cognitive flexibility, there are several other reasons why measures of language development might predict later performance on false-belief tasks. For example, children with more developed language skills might also be more social than less language-proficient children, and as a result of their increased exposure to social experiences, these more language-proficient children might learn earlier to consider other people's mental states both in their everyday interactions and on tasks assessing this kind of understanding. Moreover, standard theory of mind tasks, such as the false belief task, are not ideal measures of cognitive flexibility, even if cognitive flexibility is required to perform well on these task, because several other processes may also be needed to perform successfully on these tasks (e.g., good understanding of mental states).

For all the reasons outlined above, the existing evidence for the role of language in the development of cognitive flexibility is open to question. Clearly, to determine more convincingly whether language plays a role in the development of flexible thinking in preschoolers, experiments with preschoolers are needed in which language manipulations are introduced experimentally–as Glucksberg and Weisberg (1966) did in their studies with adults–and children are assessed on relatively uncomplicated measures of cognitive flexibility. The overall aim of the current experiments was to do just that. First, however, a suitable task needed to be devised for use with preschoolers in which other cognitive demands were minimized.

1.3. Assessing Abstraction and Cognitive Flexibility

In recent years, developmental neuropsychologists have adapted several traditional neuropsychological tests of abstraction and flexible thinking for use with school-aged children (e.g., Chelune & Baer, 1986; Chelune & Thompson, 1987; Welsh, Pennington, & Groisser, 1991), but there is a paucity of convenient and suitable tests of these abilities for use with preschoolers. A classic neuropsychological test of these abilities, the Wisconsin Card Sorting Task (WCST), was developed by Berg and Grant (Berg 1948; Grant & Berg, 1948) for the specific purpose of assessing in normal adults "[t]he phenomena of 'abstract behavior' and 'shift of set' in thinking" (Berg, 1948, p. 15). The WCST has been an important tool for differentiating between individuals with various types of brain dysfunction (e.g., Milner & Petrides, 1984). For example, in a well-known study, Milner (1963) found that individuals with lesions to dorsolateral prefrontal cortex were significantly impaired on the WCST relative to individuals with lesions to other cortical areas. Moreover, the WCST has been used to assess neuropsychological functioning in school-aged children with a variety of developmental psychopathologies (see Pennington & Ozonoff, 1996, for a review) and normative developmental data on the WCST are also available (e.g., Chelune & Baer, 1986; Chelune & Thompson, 1987; Welsh et al., 1991).

More recently, however, researchers have expressed some dissatisfaction with the WCST (e.g., Delis, Squire, Bihrle, & Massman, 1992; Levine, Stuss, & Milberg, 1995; Pennington & Ozonoff, 1996), in part because failures on the WCST are difficult to interpret given the large number of cognitive processes that need to be intact in order to perform well on this task. For example, in addition to assessing participants' ability to detect a correct dimension and their ability to switch flexibly between dimensions, successful performance on

~~the WCST depends upon the ability to benefit from feedback. the ability to keep the correct~~ dimension in mind over several trials, and the ability to inhibit prepotent responses. Moreover, recent evidence indicates that difficulties on the WCST may not be specific to individuals with lesions to the frontal lobes (see Stuss, Eskes, & Foster, 1994, for a review). Finally, the numerous cognitive and instructional demands of the WCST undermine its utility with preschoolers.

1.3.1. The Visual-Verbal Task

A less well-known neuropsychological test of abstraction and cognitive flexibility is the Visual-Verbal Test developed by Feldman and Drasgow (1951; Drasgow & Feldman, 1957). The original version of the task (Feldman & Drasgow, 1951) involved 43 cards, each of which depicted four objects. On each card, three of the four objects could be grouped according to one dimension, and either the same or a different trio could be grouped according to a second dimension. For example, on one stimulus card, objects consisted of a small white circle, a large white circle, a large black circle, and another large white circle. Participants were asked to group three objects that were alike in some way (e.g., the three large circles) and then to group three objects that were alike in a different way (e.g., the three white circles). Objects could be grouped according to "color, form, size, structural similarities, naming, and function" (p. 56). The first grouping that participants made provided a direct measure of their ability to detect a single dimension (the abstraction component of the task), and the second grouping provided a measure of participants' ability to switch flexibly between dimensions (the cognitive flexibility component), although it also required abstraction. Feldman and Drasgow found that, relative to a comparison group consisting of normal adults, individuals with schizophrenia had particular difficulty detecting a correct second grouping, suggesting that they have difficulty with flexible thinking. This finding has been replicated several times (e.g., Drasgow & Feldman, 1957; Siegel, 1957; Stuss et al., 1983).

As a measure of abstraction and cognitive flexibility, the Visual-Verbal Test possesses several advantages over the WCST. Whereas the WCST requires that participants respond according to a specific dimension across several trials before switching and responding according to another dimension, the Visual-Verbal Test requires that participants select items according to one dimension and immediately switch and select items according to another dimension. For this reason, the Visual-Verbal Test places fewer demands on working memory. In addition, unlike the WCST, performance on the Visual-Verbal Test does not depend upon participants' ability to benefit from feedback. Because of the simplicity of the instructions and of the task, poor performance on this task can be more easily interpreted. Furthermore, because of its simplicity, a modified version of the Visual-Verbal Test can even be used with preschoolers.

1.3.2. The Flexible Item Selection Task

To assess abstraction and cognitive flexibility in preschoolers, I developed a new task, the Flexible Item Selection Task (FIST), based on the Visual-Verbal Test (Feldman & Drasgow, 1951). On each trial of the FIST, children are shown three **items**.[1] This item trio constitutes a **trial set** (e.g., one small yellow teapot, one small blue teapot, one medium blue teapot; see Figure 1). In each trial set, one pair of items (e.g., the small yellow teapot and the small blue teapot) match each other on a **cue** of one **relevant dimension** (e.g., small for size).

[1] Throughout the dissertation, on the first appearance of an important term, it appears in bold font. In addition, to assist the reader, a glossary of these terms also appears in Appendix A.

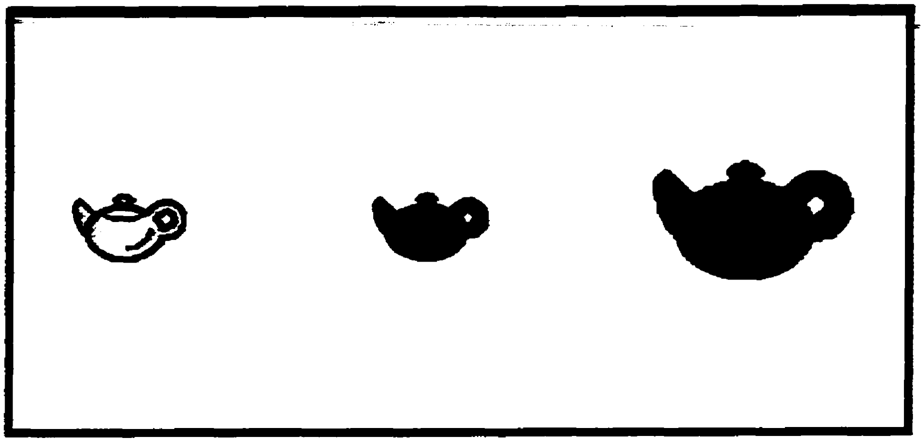

Figure 1. Example of a trial set presented in the Flexible Item Selection Task.

but differ from the third item on that dimension (e.g., medium for the medium blue teapot). A different pair of items (e.g., the small blue teapot and the medium blue teapot) match each other on a cue of a different relevant dimension (e.g., blue for colour), but differ from the remaining item on that dimension (e.g., yellow for the small yellow teapot). The cue of a third (and a fourth in Experiment 1) **irrelevant dimension**, is constant across the three items (e.g., teapot for shape; one for number). Thus, on each trial, one item (e.g., the small blue teapot) always matches a second item on one dimension (e.g., the small yellow teapot), and at the same time, matches the third item on another dimension (e.g., the medium blue teapot). This particular item can be referred to as the **pivot item** because it needs to be selected twice. On each trial, participants are first asked to select two items that are alike in some way (**Selection 1**; e.g., the small ones). Once they have made an unambiguous response, they are then asked to select a second pair of items that are alike in some other way (**Selection 2**; e.g., the blue ones). Like the Visual-Verbal Test, it is assumed that Selection 1 responses provide an index of children's ability to abstract a single dimension, whereas Selection 2 responses provide an index of children's ability to switch flexibly between dimensions.[2]

Consequently, although the FIST is equivalent to the Visual-Verbal Test in essential characteristics, it differs from its predecessor in several respects. For example, it uses child friendly stimuli, it has fewer trials than the Visual-Verbal Test, it uses fewer and more clearly defined dimensions, and it was developed with more tightly controlled counterbalancing procedures (see Section 2.2.2. and 3.2.2.).

1.4. Purpose of Current Study

The overall aims of this series of experiments were to determine whether language plays a pivotal role in the development of cognitive flexibility in children and if so, how it might come to exert its influence. Experiment 1 was conducted to assess whether the FIST is an appropriate task for use with preschoolers as a means of exploring the development of abstraction and cognitive flexibility and to determine whether there are meaningful age-related differences in performance on this task within this particular age range. Experiment 2 had two purposes. First, it was conducted to determine whether the findings from Experiment 1 could be replicated with a modified and improved computerized version of the FIST. Second, and more important, labelling manipulations were also included to determine whether or not language affects performance on this task. Experiment 3 was conducted to further specify the exact nature of the role that labelling plays on this task by examining whether particular types of labels differ in their effectiveness in improving performance.

[2] The validity of this assumption is tested empirically in Experiment 3.

CHAPTER

EXPERIMENT 1

PRESCHOOLERS' PERFORMANCE ON AN INITIAL VERSION OF THE FLEXIBLE ITEM SELECTION TASK

2.1. Introduction

The primary purposes of Experiment 1 were to determine whether the FIST is appropriate for use with preschoolers as a test of abstraction and cognitive flexibility and whether meaningful age-related differences in performance exist on this task. Children at four ages (i.e., 2-, 3-, 4-, and 5-year-olds) participated in the experiment. To ensure that any differences between age groups were due to differential difficulties with abstraction, cognitive flexibility, or both, and not to possible age differences in understanding and following basic task instructions, criterial trials were also included. In order to succeed on these criterial trials, children only needed to understand the instructions themselves, which were identical to the instructions used in the FIST; children were not required to abstract or represent flexibly dimensional information on these trials because it was possible for them to use overall perceptual similarity information instead of dimensional information to perform well on these criterial trials (i.e., they could use a simple perceptual-matching strategy). If children erred on these trials, they did not receive the task proper.

2.2. Method

2.2.1. Participants

The sample consisted of 60 two-year-olds (M = 30.6 months, SD = 1.6 months, range = 27.4 to 35.8 months), 53 three-year-olds (M = 43.0 months, SD = 1.9 months, range = 38.7 to 46.6 months), 49 four-year-olds (M = 54.7 months, SD = 2.0 months, range = 51.1 to 58.5 months) and 35 five-year-olds (M = 66.3 months SD = 2.5 months, range = 60.7 to 69.6 months), including 97 girls and 100 boys. The girl:boy ratios were 32:28 for 2-year-olds, 29:24 for 3-year-olds, 24:25 for 4-year-olds, and 12:23 for 5-year-olds. Eight children (6 two-year-olds and 2 three-year-olds) were excluded from the analyses because they refused to perform the task in its entirety. Children were recruited from several local daycare centres or from a database of children whose parents had expressed an interest in participating in research. Informed consent was obtained from the parents of all children who participated in the experiment. Children with suspected or known developmental or medical disorders that might affect their performance did not participate in the experiment (1 child in a daycare centre was not tested because he was suspected of having a developmental delay, whereas another was not tested because he was suspected of having a hearing impairment). Likewise, children in daycare centres with a poor grasp of the English language were also not tested (n = 3).

2.2.2. Task Design

Forty-eight 21.5 x 5.5 cm laminated white cards were used. Each card[3] depicted a set of objects that were derived from the combination of four dimensions (colour. shape. size. and number). Each dimension was represented by one of three cues. The colour dimension was represented by the colours pink. purple. and orange; the shape dimension was represented by a phone. a pair of socks. and a fish; the size dimension was represented by small (the mean rectangular area of each object was approximately 7 cm^2). medium (approximately 13 cm^2). and large (approximately 23 cm^2) objects; and the number dimension was represented by one, two. or three objects. Thus. a specific card might depict one large pink phone. whereas another card might depict three small purple fish. When only one object was depicted on a card. it was positioned in the centre of the card. When two objects were depicted. one was positioned at the extreme left of the card and the other at the extreme right. Finally. when three objects were depicted. one was located at the extreme left end. another at the centre. and the other at the extreme right end of the card.

2.2.2.a. Demonstration and criterial trials. The demonstration trial and the two criterial trials always consisted of sets of four cards. Two of these cards were identical on all four dimensions (i.e.. colour. shape. size. and number). whereas the other two cards. which were also identical to each other on all four dimensions. differed from the first pair on all dimensions (see Figure 2). For example. for the demonstration trial. two cards depicted one small pink pair of socks and two cards depicted two medium orange fish.

These three preliminary trials (demonstration trial and criterial trials) were always

[3] Note that in Experiment 1. the term "card" is used instead of "item" consistently to refer to one of the three items in a trial set because of the confusion that may result from the fact that number was included as a possible relevant dimension in Experiment 1, but in subsequent experiments. the term "item" is adopted again.

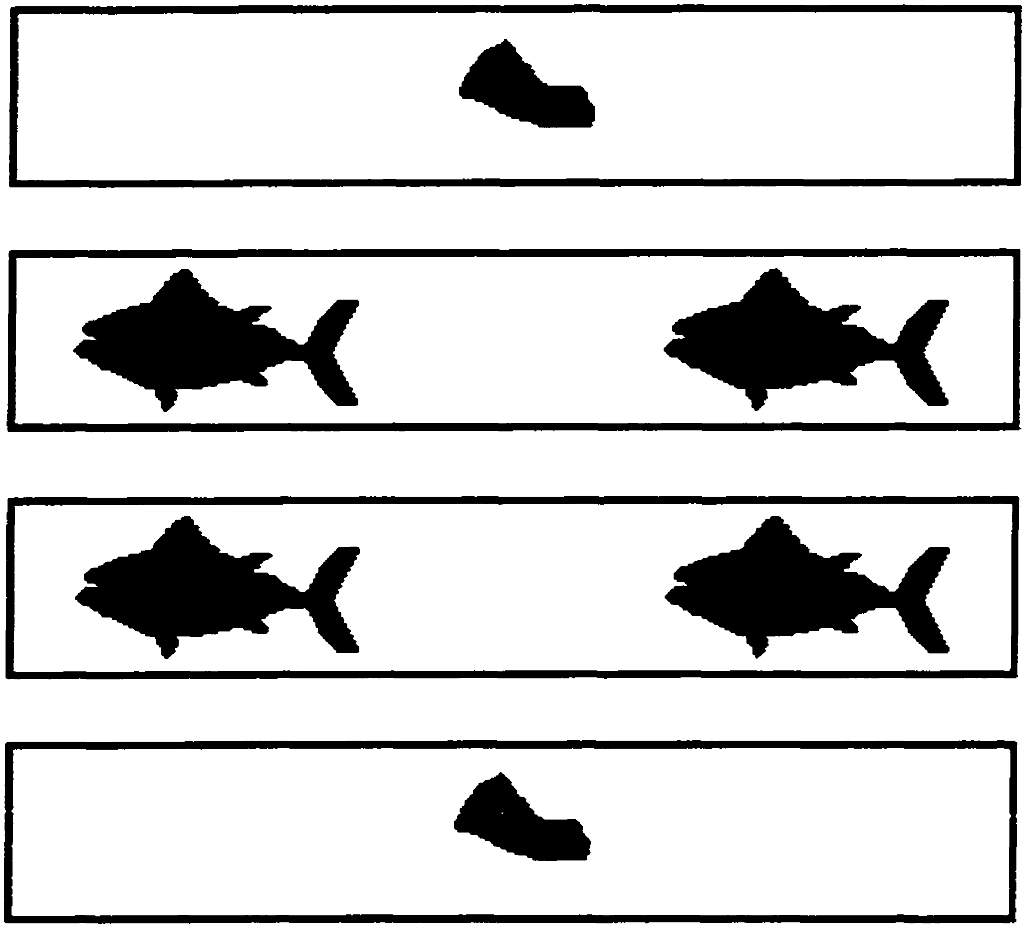

Figure 2. Example of cards presented in the demonstration and criterial trials in Experiment 1.

presented in the same order across all children. Furthermore, the placement of matching pairs was counterbalanced across the three trials. That is, on the demonstration trial, the first card matched the fourth card (and therefore, the middle two cards matched each other), whereas on Criterial Trial 1, the first card matched the second card, and on Criterial Trial 2, the first card matched the third card. Additionally, each cue of each dimension (e.g., orange for colour or small for size) was used twice across these three preliminary trials.

2.2.2.b. Flexible Item Selection Task. Trial sets presented in the FIST each consisted of sets of three cards that were identical on two (irrelevant) dimensions (e.g., size and number) but differed on two relevant dimensions (e.g., colour and shape). Two of the three cards matched on one of the relevant dimensions (e.g., colour) and a different pair of cards matched on the other relevant dimension (e.g., shape). For example, in a given trial set, one card might depict one medium purple phone, the second card might depict one medium pink phone, and the third might depict one medium pink fish (see Figure 3). Thus, in all trial sets, a pivot card (e.g., the pink phone) matched one of the other cards on one relevant dimension (e.g., colour; the other pink one) and matched the remaining card on the other relevant dimension (e.g., shape; the other phone).

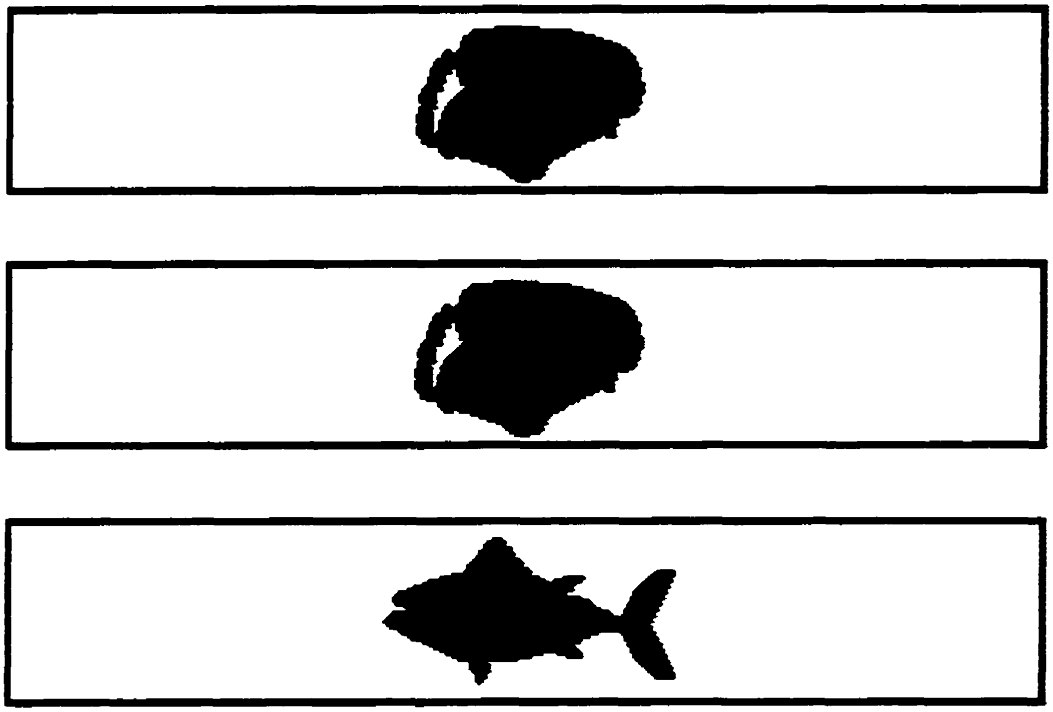

Figure 3. Example of cards presented in the Flexible Item Selection Task in Experiment 1.

All children received the same 12 trial sets (presented in the same order), which consisted of two blocks of six trial sets. Within each trial block, each of the six possible **relevant-dimension pairs** (i.e., colour and shape, colour and size, colour and number, shape and size, shape and number, and size and number) was presented once in a random order (see Table 1 in Appendix B for more detailed counterbalancing information). Each dimension on its own was thus relevant on 6 of the 12 trial sets (2 trial sets with each of the other possible dimensions), but the same dimension was never presented on more than two consecutive trials. The cue of the relevant dimension by which cards matched can be referred to as the **dominant cue** (e.g., pink for colour and phone for shape in the example in Figure 3), whereas the cue of the nonmatching third card can be referred to as the **nondominant cue** of the relevant dimension (e.g., purple for colour and fish for shape in that example). Within each trial block, each cue within each dimension appeared once as the dominant cue of the relevant dimension, once as the nondominant cue of the relevant dimension, and once as the **irrelevant cue** of an irrelevant dimension (i.e., one of the dimensions that did not vary across the three cards; e.g., medium for size and one for number in that example).

Pivot-card placement was also counterbalanced. The pivot card could appear as Card 1, Card 2, or Card 3. For example, in Figure 3, the placement of the pivot card is Card 2 (the middle card in the figure). Each of the three possible pivot-card placements was used on four trials (two trials within each trial block) but the same placement was not used on more than two consecutive trials.

2.2.3. Procedure

Children were tested individually in one testing session at their respective daycare centres or at the university. They were given 1 demonstration trial, 2 criterial trials, and 12 trials on the FIST itself, and were introduced to the task by being told that they would see some cards with pictures on them. On the demonstration trial, the experimenter then placed four cards side by side and told children,

> Look! Here's a card, here's another card, here's another card, and here's another card. I'm going to pick two cards that are the same in one way. So I'll pick these two cards [simultaneously pointing to two identical cards; i.e., Card 1 and Card 4]. These two cards are the same because they both have one little pink pair of socks. So they're the same. Now I'm going to pick two cards that are the same but in a different way. So I'll pick these two cards [simultaneously pointing to the other pair of identical cards; i.e., Card 2 and Card 3]. These two cards are the same because they both have two medium orange fish on each card. That's why they're the same. So these two cards here are the same [pointing to the first pair] and these two cards here are the same [pointing to the second pair], but see, these two cards here are different from those two cards. You know what? Now it's your turn to show me some cards.

Children were subsequently given two criterial trials. On each criterial trial, they were instructed to, "Show me (put your fingers on) two cards that are the same in one way." (Selection 1). Once children responded, they were then asked to, "Show me two cards that are the same but in a different way." (Selection 2). Children were not given feedback on their responses because these criterial trials served to ensure that children understood the basic task instructions, including the requirement of selecting two cards, the concept of same, and the concept of same but in a different way. If children erred on even one selection on either of the two criterial trials, they did not receive the task proper.

Children who succeeded on both criterial trials were then given 12 trials of the FIST, on each of which they were also required to make two selections. The instructions on the FIST were identical to those given on the criterial trials, except that FIST trials involved only three cards instead of four. If children failed to select two cards on either selection within a particular trial, then the experimenter prompted them on the first occasion in which this occurred (i.e., "Is there another one that's the same as that one?"). Similarly, if children seemed to hesitate after selecting two cards, then the experimenter prompted them on the first instance in which this occurred by asking, "Are you done?"

2.3. Results

2.3.1. Performance on Criterial Trials

The percentages of 2-, 3-, 4-, and 5-year-olds who failed at least one of the two criterial trials were 85%, 22%, 12%, and 3%, respectively. A chi-square test confirmed that successful performance on the criterial trials varied as a function of age group, $\chi^2(3, \underline{N} = 189) = 92.20$, $\underline{p} < .001$ (assuming an alpha level of .05; see Table 1). Separate Fisher's Exact tests were conducted on each pair of age groups to determine which groups differed from each other. Two-year-olds differed from all other age groups ($\underline{p}$s < .001) and 3-year-olds differed only from 5-year-olds, ($\underline{p} < .05$). Three-year-olds did not differ from 4-year-olds and 4-year-olds did not differ from 5-year-olds ($\underline{p}$s > .10).

Table 1

Number of Children Who Failed or Passed the Criterial Trials in Experiment 1 as a Function of Age Group

	Performance	
Age Group	Failed	Passed
2-year-olds	46	8
3-year-olds	11	40
4-year-olds	6	43
5-year-olds	1	34

2.3.2. Performance on the Flexible Item Selection Task

2.3.2.a. Main analyses. Given that only 8 two-year-olds succeeded on the criterial trials–and as a result, completed the task proper–they were excluded from further analyses because they constituted a nonrepresentative sample of 2-year-olds. Thus, only 3-, 4-, and 5-year-old children who succeeded on the criterial trials were included in the analyses on the FIST. Although the basic design of the experiment was a 3 x 2 x 2 (Age x Sex x Selection) design with repeated measures on selection, using selection as a repeated measure was problematic because Selection 2 necessarily followed from and, in fact, depended upon Selection 1. Consequently, this violated the assumption underlying repeated-measures designs of no carry-over effects between treatments (Neter, Wasserman, & Kutner, 1990). Therefore, separate analyses were conducted for each selection.

Inspection of the raw data indicated that the distributions of scores on Selection 1 were negatively skewed, presumably because of ceiling effects on performance. However, as Kirk (1982) noted, "[s]kewed populations have very little effect on either the level of significance or the power of the F test for the fixed-effect model" (p. 75). Thus, a 3 x 2 (Age Group x Sex) analysis of variance (ANOVA) on Selection 1 performance was conducted and a significant main effect of age group was detected, F (2, 111) = 6.81, MSE = 2.58, p < .01 (see Figure 4). Post hoc pairwise comparisons using Tukey's HSD tests revealed that 3-year-olds (M = 9.65, SD = 1.53) did significantly worse than both 4- and 5-year-olds (M = 10.86, SD = 1.51, p < .01, and M = 10.74, SD = 1.75, p < .05, respectively), and that 4- and 5-year-olds did not differ from each other (p > .10). Pairwise effect-size comparisons (Cohen, 1988) further revealed that the difference between means was large for 3- and 4-year-olds (d = 0.80) and moderate for 3- and 5-year-olds (d = 0.66). In contrast, as predicted from the ANOVA,

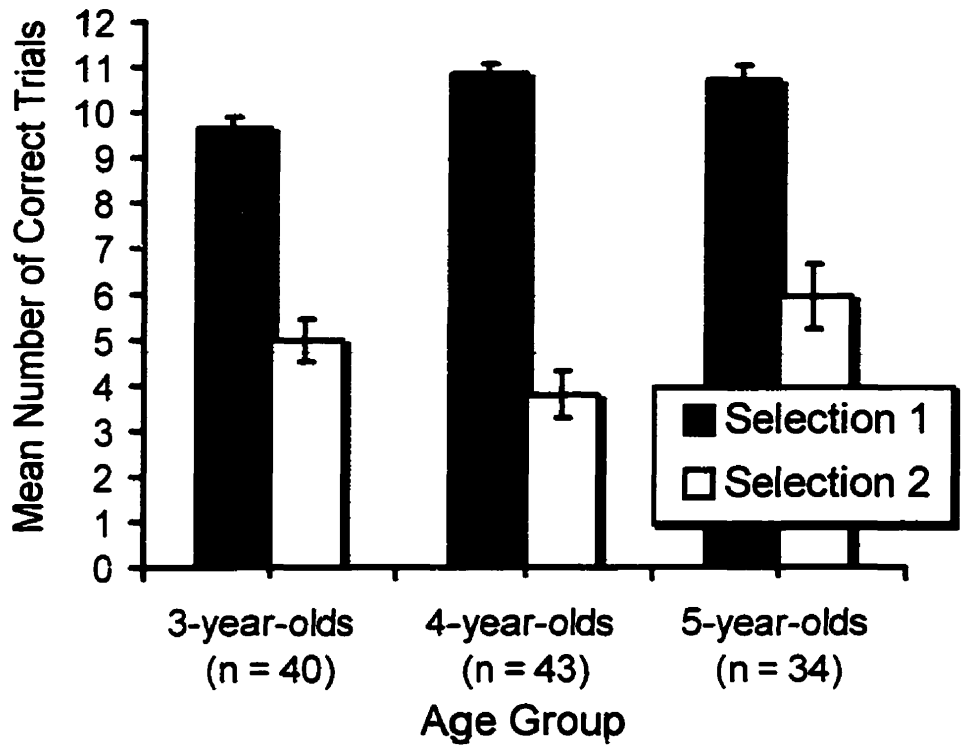

Figure 4. Mean numbers (and standard errors) of correct trials in Experiment 1 as a function of age group and selection (N = 117). (Note that the mean and standard error of 3-year-olds' Selection 2 responses are also presented despite the fact that 3-year-olds were not included in the analysis on Selection 2 performance because of their relatively poor Selection 1 performance.)

the difference between means for 4- and 5-year-olds was negligible (d = 0.08).

Given that Selection 2 performance was meaningful as a measure of flexibility only in the context of relatively good performance on Selection 1, data from 3-year-olds were omitted from the analyses of Selection 2 performance because these children did significantly worse on Selection 1 than both 4- and 5-year-olds. A 2 x 2 (Age Group x Sex) ANOVA on 4- and 5-year-olds' Selection 2 performance revealed a significant difference between age groups, F (1, 73) = 6.08, MSE = 14.61, p < .05 (refer to Figure 4 again). Four-year-olds (M = 3.79, SD = 3.49) performed significantly worse than 5-year-olds (M = 5.94, SD = 4.20) on Selection 2 despite equivalent performance on Selection 1. This difference was moderate in terms of effect size, d = 0.56.

2.3.2.b. Response-pattern analyses. Selection 1 responses were classified into five mutually exclusive (and exhaustive) categories. Responses were classified as **correct pair** if children selected a correct pair of cards (note that two correct pairs were possible on Selection 1). Incorrect responses were classified as **wrong pair** if children selected the (only) wrong pair of cards (e.g., the purple phone and the pink fish in the example in Figure 3). The other three response categories occurred when children selected either more or fewer than two cards, including selecting (a) **all cards**, (b) **one card**, or (c) **no cards**.

Table 2 illustrates the percentages of Selection 1 (and Selection 2) responses that were of each type summed across all trial sets and across all children within each age group. Although no inferential statistics could be conducted on these data because observations were not independent of each other, it is worth noting that 3-year-olds selected the wrong pair on Selection 1 almost three times more often than older children. This finding is consistent with the results of the ANOVA, which revealed that 3-year-olds had more difficulty on Selection 1 than the older children. Also of interest is the finding that on Selection 1, the percentages of other errors (i.e., all cards, one card, and no cards) were relatively low (less than 5% combined for each age group) and were similar across all age groups.

Selection 2 responses were also classified into separate categories. In addition to the five categories used for Selection 1 responses, two other response categories were added. On Selection 2, it was also possible for children to err by selecting the **same pair** of cards that

Table 2

Overall Percentages for Each Possible Response Category Across Children in Experiment 1 as a Function of Selection and Age Group

	Response Categories						
Age Group	Correct Pair	Wrong Pair	Same Pair	All Cards	One Card	Rem. Card	No Cards
Selection 1							
3-year-olds	80.4	16.9	–	0.2	2.5	–	0
4-year-olds	90.5	6.0	–	0.2	3.1	–	0.2
5-year-olds	89.5	5.9	–	0.2	2.7	–	1.7
Selection 2							
3-year-olds	41.5	20.8	12.9	0.2	1.7	22.7	0.2
4-year-olds	31.6	17.4	6.8	0.2	1.9	40.5	1.6
5-year-olds	49.5	15.4	10.5	0	2.9	16.9	4.7

Note. The "Same Pair" and "Rem. [Remaining] Card" response categories applied only to Selection 2 responses, and "One Card" responses for Selection 2 consisted of "other one-card" responses (see Section 2.3.2.b. for definitions on each response category). Each child within each age group contributed 12 responses for each selection.

they selected on Selection 1. Furthermore, one-card responses were classified into two separate categories: These included **remaining-card** responses, which occurred when children selected only the card that they had not selected on Selection 1, and **other** one-card responses, which included all other instances of one-card responses. Hence, response categories for Selection 2 included the following seven mutually exclusive and exhaustive categories: (a) correct pair, (b) wrong pair, (c) same pair, (d) all cards, (e) remaining card, (f) other one card, and (g) no cards.

The Selection 2 data were different from Selection 1 data in several respects. As

~~shown in Table 2, unlike Selection 1 in which children tended to err by selecting the wrong~~ pair, the most common type of error on Selection 2 was remaining-card responses, particularly for the 4-year-olds who erred in this manner approximately twice as often as either the 3- or 5-year-olds. The next most frequent types of incorrect responses for all age groups were wrong-pair and same-pair responses, whereas the other types of incorrect responses occurred infrequently. Morever, unlike Selection 1, it was the 4-year-olds who erred on a greater proportion of trials than the other age groups, not the 3-year-olds.

2.3.2.c. Task analyses. A series of analyses were also conducted to determine whether the age-related differences noted above were influenced by certain task variables. For the purpose of these analyses (except for the analysis on dimension preferences: see below), performance on each trial was summed across both selections so that an accurate estimate of the true difficulty of any given trial set could be obtained.

First, to determine whether or not children's performance improved (or worsened) over trials, performance across both selections on the first trial block was compared to performance on the second trial block.[4] A 3 x 2 x 2 (Age Group x Sex x Trial Block) ANOVA with repeated measures on trial block revealed only a main effect of age group, $\underline{F}$ (2, 111) = 3.52, $\underline{MSE}$ = 7.10, $\underline{p} < .05$. Pairwise comparisons revealed that both 3- ($\underline{M}$ = 14.63, $\underline{SD}$ = 2.71) and 4-year-olds ($\underline{M}$ = 14.65, $\underline{SD}$ = 3.52) differed from 5-year-olds ($\underline{M}$ = 16.68, $\underline{SD}$ = 4.93; $\underline{p}$s < .01), but that they did not differ from each other ($\underline{p} > .10$).[5]

[4] Recall that trial sets were presented in two trial blocks, each of which contained six trials that included one trial set with each of the six possible relevant-dimension pairs (refer to Table 1 in Appendix B) and two trial sets with each of the three possible pivot-card placements.

[5] Note that this pattern of age-group differences is different from the pattern observed on Selection 1 performance. This results from the fact that performance on both selections was combined in this analysis.

In addition, to determine if age-group differences in performance were attributable in part to age-related differences in children's tendencies to select cards according to specific dimensions (i.e., dimensional preferences), Selection 1 responses were examined more closely. As mentioned previously, cards matched on one of four dimensions (i.e., colour, shape, size, and number) and a particular dimension was relevant on 6 of the 12 trial sets (2 trial sets with each of the other three possible dimensions). If children exhibited no bias for selecting cards according to specific dimensions, then on Selection 1, they ought to have selected cards according to each dimension on approximately half of the trials in which each was relevant. Therefore, a 3 x 2 x 4 (Age Group x Sex x Selection 1 Dimension) ANOVA with repeated measures on Selection 1 dimension was conducted. The analysis revealed a significant main effect of age group, F (2, 111) = 6.81, MSE = 0.65, p < .01, and a significant main effect of Selection 1 dimension, F (3, 333) = 89.60, MSE = 1.97, p < .0001. Pairwise comparisons revealed that all dimensions were differentially selected on Selection 1 (all ps < .01): shape (M = 4.15, SD = 1.40) was selected most often, then colour (M = 3.03, SD = 1.54), then number (M = 1.94, SD = 1.12), and then, size (M = 1.28, SD = 1.05).

Similarly, to assess whether it was more difficult to detect certain dimensions within specific pairings, a 3 x 2 x 6 (Age Group x Sex x Relevant-Dimension Pair) ANOVA was conducted on both selections with repeated measures on relevant-dimension pair. In addition to the expected main effect of age group, F (2, 111) = 3.52, MSE = 2.37, p < .05, the analysis revealed a main effect of relevant-dimension pair, F (5, 555) = 11.71, MSE = 0.37, p < .0001. Table 3 presents the means and standard deviations for each relevant-dimension pair and the results of Tukey's HSD post hoc tests that were conducted to determine which relevant-dimension pairs differed from each other.

~~Table 3~~

Means, Standard Deviations, and Results of Pairwise Comparisons Between Relevant-Dimension Pairs in Experiment 1

Relevant-Dimension Pairs	M	SD
colour / size	2.77^{a}	0.91
colour / shape	2.68^{ab}	0.98
size / number	2.60^{abc}	0.88
shape / number	2.54^{bc}	0.86
colour / number	2.43^{cd}	0.79
shape / size	2.21^{d}	0.63

Note. Relevant-dimension pairs are ordered in terms of relative ease (from easiest to most difficult).

^{a,b,c}Means with the same letters did not differ from each other (the minimum difference needed between means for an alpha of .05 was 0.23).

Finally, to determine whether the placement of the pivot card (i.e., Card 1, Card 2, or Card 3) influenced children's tendency to detect both pairings, a 3 x 2 x 3 (Age Group x Sex x Pivot-Card Placement) ANOVA on both selections with repeated measures on pivot-card placement was conducted. Main effects of age group, F (2, 111) = 3.52, MSE = 4.73, p < .05, and pivot-card placement, F (2, 222) = 51.89, MSE = 0.93, p < .0001, in addition to a two-way interaction between age group and pivot-card placement, F (4, 222) = 4.92, p < .001, and a difficult-to-interpret three-way interaction between age group, sex, and pivot-card placement, F (4, 222) = 3.81, p < .01, were all significant. Pairwise comparisons between each pivot-card placement revealed that children did better on trials in which the pivot card appeared in the centre (Card 2; M = 5.88, SD = 1.74) than when it appeared on the left (Card 1; M = 4.69, SD = 1.32; p < .01) or on the right (Card 3; M = 4.66, SD = 1.47; p < .01), but

~~that their performance did not differ on trials in which the pivot card was on the left or on the~~ right (p > .10). Figure 5 presents the means and standard errors for each age group as a function of pivot-card placement. Simple main effects test for the interaction revealed that for each of the age groups, the pivot-card placements main effect was significant, F (2, 228) = 39.90, MSE = 0.97, p < .01, F (2, 228) = 16.23, p < .01, F (2, 228) = 10.16, p < .01, for 3-, 4-, and 5-year-olds, respectively, but that for each of the pivot-card placements, the age-group main effect was significant only when the pivot card was placed on the right-hand side (i.e., Card 3), F (2, 209) = 7.82, MSE = 2.22, p < .01.

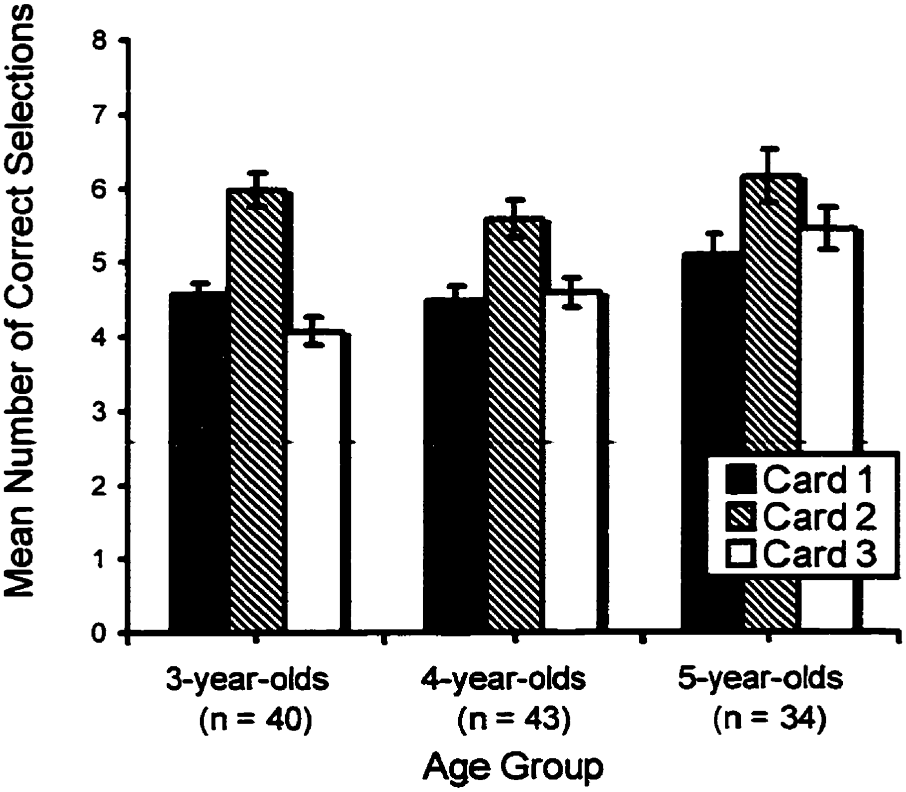

Figure 5. Mean numbers (and standard errors) of correct selections in Experiment 1 as a function of age group and pivot-card placement (N = 117).

2.4. Discussion

Several aspects of problem solving have been well-studied in preschoolers (e.g., rule use, inhibitory response control, working memory, planning, and error evaluation; see Dempster, 1992, 1993; Harnishfeger & Bjorklund, 1993; Zelazo, Carter, Reznick, & Frye, 1997; Zelazo & Jacques, 1996, for reviews), but tests assessing abstraction and cognitive flexibility have been lacking. In the current experiment, 2-, 3-, 4-, and 5-year-olds were tested on a new task, the FIST, which provides a relatively straightforward measure of these aspects of problem solving in preschoolers. Important and meaningful age-related differences emerged in this study across the entire age range tested. First, the majority of 2-year-olds showed no evidence of understanding basic task instructions, as revealed by their poor performance on criterial trials. On the criterial trials, children were given instructions that were identical to those given on the FIST; however, children did not have to detect a specific dimension by which to match a pair of cards (i.e., no abstraction component) because matching cards were identical on all dimensions. Furthermore, they did not have to select a particular card in multiple ways (i.e., no flexibility component) because there were four cards instead of three, and therefore, each card needed to be selected only once. Hence, 2-year-olds' difficulties on these criterial trials cannot be attributed to limitations in abstraction or cognitive flexibility per se.

In contrast, findings from the FIST itself converge to provide a clear picture of changes in abstraction and flexible thinking over the 3 to 5 year period. Most 3-year-olds succeeded on criterial trials, but performed worse than the older age groups on Selection 1. To perform well on Selection 1 on the FIST, children needed to detect or abstract dimensional information because matching cards were nonidentical. In contrast, the matching cards in the criterial trials were identical and therefore, it was not necessary for children to abstract dimensional information or match cards on the basis of a particular dimension. More specifically, in the criterial trials, it was possible for children to rely on a simpler perceptual-matching strategy in that they could match cards according to overall similarity without necessarily detecting a specific dimension. Ample evidence exists in the literature to support the notion that young children do indeed use this kind of simpler strategy (see Gentner & Ratterman, 1991; Smith, 1989a; Zelazo & Jacques, 1996, for reviews). The relatively poor performance of 3-year-olds on Selection 1 compared to their own good performance on the criterial trials (and to the performance of the older age groups on Selection 1) suggests that they had difficulty correctly detecting a dimension that was common to two cards when these two cards were not identical to each other. Thus, their difficulty on the FIST appears to be due primarily to difficulties in abstracting out a relevant dimension.[6]

In contrast, 4-year-olds performed as well as 5-year-olds on Selection 1 of the FIST, suggesting that they correctly detected how two nonidentical cards were identical along one dimension. In other words, they appeared to do well on the abstraction component of the task. However, their relatively poor performance on Selection 2 of the FIST compared to 5-year-olds is consistent with the suggestion that it was the requirement that they flexibly select one of the cards (e.g. the pink phone in Figure 3) according to two dimensions that rendered the task difficult for them. Indeed, when they erred on Selection 2, they tended to do so by

[6] Or alternatively, it may be due to a difficulty in abstracting out a relevant dimension but only when confronted with conflicting information. That is, it may not have been the requirement that they detect a common dimension from nonidentical items per se that was difficult. Rather, it may have been a difficulty in doing so when there were two conflicting matches possible (e.g., selecting the two phones on shape, or selecting the two pink ones on colour). I will return to this point again in Section 5.4.

~~selecting the remaining card alone (see below for further discussion of remaining-card~~ responses). On Selection 1 most errors occurred because children selected the wrong pair demonstrating that they at least understood to some degree that they were expected to select two cards. However, despite understanding that they needed to select two cards–as evidenced by their good performance on the criterial trials and on Selection 1 of the FIST–4-year-olds tended to select only one card on Selection 2. Selecting the remaining card alone raises the possibility that children failed to see how the pivot card (i.e., the card that needed to be selected twice; e.g., the pink phone in Figure 3) matched the remaining card on a different dimension. They may have perseverated in thinking of the pivot card in only one way (i.e., according to the first dimension by which they selected it), and consequently refused to select it with the only card that was left after they had selected cards on Selection 1.

It should be noted, however, that the failure to select the pivot card twice could also be due to a misunderstanding of the task requirements. The criterial trials required children to understand the basic task instructions but did not require them to abstract or represent flexibly dimensional information. Although inclusion of these trials was necessary to determine at which age children understood task instructions, the design of the criterial trials may have inadvertently led children to believe that each card ought to have been selected only once, given that correct responses in the criterial trials never required selecting any given card more than once. Of course, a combination of both of these explanations of remaining-card responses is also possible. For example, in the FIST, children may have failed to see how the pivot card matched the remaining card on a second dimension (i.e., cognitive inflexibility) and because of the lack of a fourth card, they may have opted to use the strategy of selecting the remaining card alone. On this account, their difficulty stemmed from

~~cognitive inflexibility, but how the difficulty manifested itself depended on task variables~~ (e.g., the inclusion of the criterial trials). To differentiate between these accounts, it was important to conduct another experiment with the FIST in which no criterial trials were included, thereby avoiding these potential ambiguities in interpretation. Moreover, given the results of Experiment 1, it seemed no longer necessary to include such trials, because it is clear that the large majority of 3-, 4-, and 5-year-olds understand basic task instructions.

In addition, the analyses on task variables suggest that the counterbalancing procedures and the stimuli that were used in the paper version of the FIST could be improved. For example, relevant-dimension pairs had a significant effect on performance as did pivot-card placements. Despite the fact that both relevant-dimension pairs and pivot-card placements were counterbalanced across trial sets, these two variables were not crossed and counterbalanced with each other. As a result, it is difficult to interpret the findings with respect to each of these variables, given that they may have been confounded with each other. Hence, in subsequent experiments, these two variables were crossed with each other so that their separate–as well as their joint–contributions to performance could be established.[7]

[7] Despite the fact that the results of the analyses on task-related variables are reported in this and in each of the subsequent experiments, the actual discussion of most of these results will be postponed until Section 5.1.3. in the General Discussion. This is done in part because no specific predictions were made with respect to these variables and they are somewhat tangential to the main hypotheses. Moreover, effects of task-related variables are interesting insofar as they are replicable across all experiments. As a result, the effects of specific task-related variables will be compared and contrasted across all experiments only after all pertinent results have been reported.

CHAPTER III

EXPERIMENT 2

AGE AND LABELLING EFFECTS ON PRESCHOOLERS' PERFORMANCE ON THE FLEXIBLE ITEM SELECTION TASK

3.1. Introduction

Experiment 2 was conducted in an attempt to replicate the findings of Experiment 1 with 3-, 4-, and 5-year-olds using a slightly different, and in many ways, a more refined, computerized version of the FIST, and also to begin to explore experimentally whether language plays an instrumental role in successful performance on this task. To do so, children were tested in one of three conditions that differed in terms of whether children were asked to label the stimuli, and if so, which aspects of the stimuli they were asked to label. Thus, aside from assessing children in a standard, **no-label condition**, two other labelling conditions were included: a **relevant-label condition**, in which children were asked to label the dimensions by which they selected items, and an **irrelevant-label condition**, in which they were asked to label the irrelevant dimension on each selection. Performance in the relevant-label condition was of particular interest, although the irrelevant-label condition was also important in that it served as a control condition for possible nonspecific effects of verbalization because it required that children talk about the items (as in the relevant-label condition), but it did not require that children talk about their selection per se. In past studies, improvements in task performance attributable to labelling effects appear to be limited to labelling relevant aspects of stimuli; labelling irrelevant aspects of stimuli has failed to help performance (e.g., Cantor & Spiker, 1976; Dickerson, 1970; Kendler & Kendler, 1961; Kobayaski & Cantor, 1974).

On the basis of the findings of Experiment 1, it was predicted that 3-year-olds would make more incorrect selections on Selection 1 than either 4- or 5-year-olds, who were not expected to differ from each other on that selection. In turn, 4-year-olds were expected to perform worse than 5-year-olds on Selection 2–at least those in the no-label and irrelevant-label conditions. In contrast, 4-year-olds in the relevant-label condition were expected, not only to outperform the 4-year-olds in the other conditions on Selection 2, but their Selection 2 performance was also expected to be indistinguishable from that of 5-year-olds in all conditions. In fact, only 4-year-olds in the relevant-label condition were expected to benefit from the labelling manipulations. In both kinds of labelling conditions, children were asked to label after each of their selections (see Section 3.2.2. for procedural details), and as a result, because labels followed each selection, it was difficult to assess performance on Selection 1 with respect to the labelling manipulations. Consequently, given the design of the experiment, it was only possible to assess performance on Selection 2 in relation to labels

~~provided on Selection 1.~~[8] ~~Hence, given that 3-year-olds were expected to fail to select items~~ correctly even on Selection 1, they were not expected to benefit on Selection 2 from labelling on Selection 1. On the other hand, explicit instructions to label the relevant dimension on Selection 1 was also not expected to help the performance of 5-year-olds, whose performance on Selection 2 should be already high. Indeed, it was hypothesized that the good performance of 5-year-olds on Selection 2 relative to that of 4-year-olds in Experiment 1 was due to their spontaneous (overt or covert) use of relevant labels. Therefore, 5-year-olds in the relevant-label condition were not expected to benefit further from explicit instructions to label compared to 5-year-olds in the other conditions (cf. Kendler & Kendler, 1961).

Four-year-olds, then, were predicted to be "transitional" children in a sense, in that it was hypothesized that their poor performance on Selection 2 in Experiment 1 was due to their failure to identify the relevant dimension spontaneously when selecting items on each selection. However, with explicit instructions to do so, not only were they expected to be able to identify the relevant dimension correctly on Selection 1, but they were also expected to reap the benefits of doing so on their Selection 2 performance. This idea of transitional phases in development in which children are able to produce specific strategies, but fail to do so spontaneously, is similar to Vygotsky's (1929; 1934/1986) notion of the zone of proximal development and to Flavell's (1970) notion of production deficiencies.

[8] This apparent limitation in the design did have its advantages: By examining performance on one selection in terms of labelling on another entirely separate selection, it was possible to prevent labels from acting only as attention-getting devices, and it was therefore possible to exclude simple attention-getting explanations as possible explanations for the labelling-related effects obtained on this task (see Section 5.2. for further discussion on this topic).

3.2. Method

3.2.1. Participants

A total of 40 three-year-olds (M = 41.0 months, SD = 3.3 months, range = 36.1 to 46.5 months), 38 four-year-olds (M = 53.1 months, SD = 3.4 months, range = 47.9 to 59.2 months), and 36 five-year-olds (M = 65.0 months, SD = 3.2 months, range = 59.8 to 69.7 months) participated in Experiment 2 (57 girls and 57 boys). However, 6 children were dropped from the final sample: 4 three-year-olds and 1 four-year-old were dropped because they refused to complete the experiment, and 1 four-year-old girl was excluded because she had been previously diagnosed with an expressive speech disorder (unbeknownst to the experimenter until after the child had been tested). Thus, the final sample consisted of 36 children at each of the three ages; half of the children at each age (and in each condition) were girls and half were boys. Children were recruited in the same manner as in Experiment 1. None of the children who participated in Experiment 2 had participated in Experiment 1 or in any other pilot experiments with the FIST. Informed consent was obtained from all parents of children who participated in the experiment.

3.2.2. Task Design

The computerized version of the FIST included 18 trials. On each trial, participants were shown three items, each of which appeared in a different window. For example, participants might be presented with a trial set that included a small yellow teapot, a small blue teapot, and a medium blue teapot (see Figure 6 for an example). The items were devised from the combination of three dimensions: colour, shape, and size. Each dimension was in turn represented by three cues. Colour was represented by blue, red, and yellow; shape was represented by boat, shoe, and teapot; and size was represented by small, medium (which was

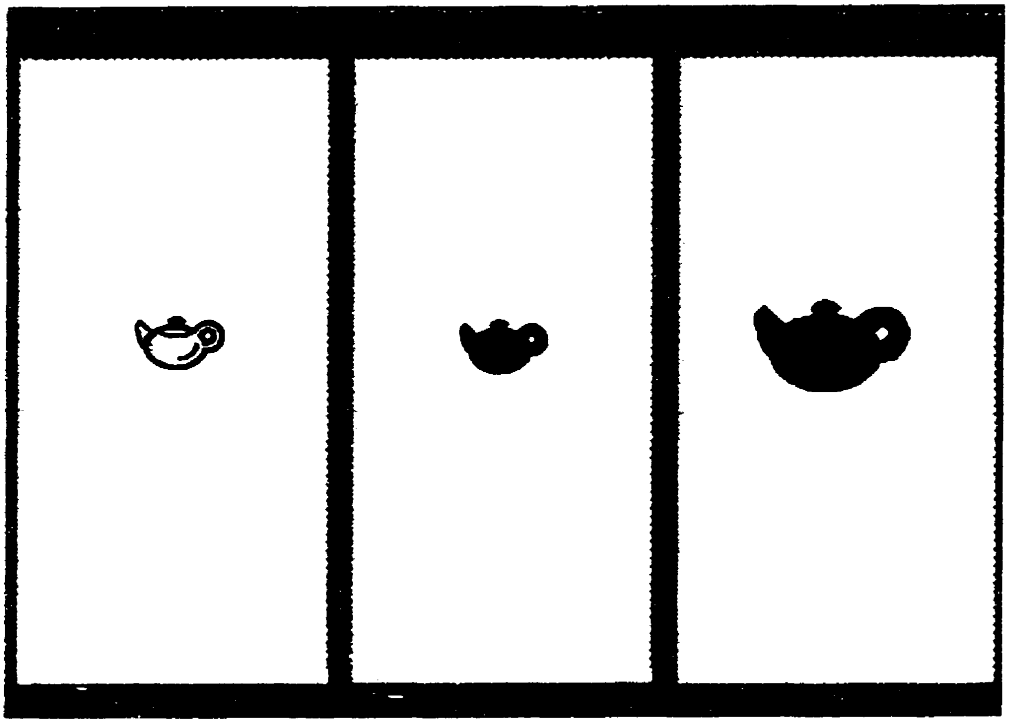

Figure 6. Computer-screen layout and example of items presented in the computerized Flexible Item Selection Task.

three times the area of the small items), and large (which was three times the area of the medium items). For each dimension, cues were selected so that they would be easily distinguishable from each other visually and phonetically. Moreover, the shapes were selected so that they would have similar height:width ratios.[1]

In each trial set, two dimensions were relevant (e.g., size and colour in the example in

[1]In Experiment 1, although the differently shaped stimuli had similar areas for any given size (see Section 2.2.2.), they did not have similar height:width ratios (e.g., the fish had longer widths relative to their height, whereas the socks had longer heights relative to their widths). As a result, it may have been difficult for children to select cards on the basis of size when shape was the other relevant dimension (i.e., when equal-sized stimuli were different shapes). Indeed, the results of Experiment 1 suggest that trial sets in which shape and size were relevant-dimension pairs were the most difficult for children (see Table 3).

Figure 6) and one dimension was irrelevant (e.g., shape). Each dimension was relevant on 12 of the 18 trials and irrelevant on the remaining 6 trials. Additionally, relevant-dimensions pairs (i.e., size and colour, colour and shape, or shape and size) each occurred six times.

The placement of items that matched each other on a relevant dimension was also fully counterbalanced. That is, a given pair of items that matched each other could be located in Windows 1 and 2, 2 and 3, or 1 and 3 (from left to right). Each possible placement of matching items (e.g., Windows 1 and 2) occurred on 12 of the 18 trials. Moreover, each possible placement of the two pairs of matching items (i.e., Windows 1 and 2, and 2 and 3; Windows 1 and 3, and 1 and 2; or Windows 2 and 3, and 1 and 3), or equivalently, of the pivot item, occurred six times. In addition, relevant-dimension pairs were fully crossed and counterbalanced with window placements. Specifically, there were 18 possible combinations of relevant-dimension pairs and window placements; each combination was presented only once.

As in Experiment 1, the number of times each cue (e.g., small, yellow, or teapot) was used as a dominant cue (i.e., the cue of a relevant dimension according to which two items matched; e.g., small for size or blue for colour in the example presented in Figure 6), a nondominant cue (i.e., the cue of a relevant dimension according to which the remaining item differed from the matching pair; e.g., medium for size or yellow for colour in that example), or an irrelevant cue (i.e., the cue of the irrelevant dimension that remained constant across all three items; e.g., teapot for shape in that example) was also controlled. More precisely, each cue (within each dimension) appeared four times as a dominant cue, four times as a nondominant cue, and twice as an irrelevant cue. Further, each possible pairing of dominant and nondominant cues within a given dimension (e.g., combining small as a dominant cue

~~and medium as a nondominant cue when size was a relevant dimension) occurred twice, once~~ with one of the other dimensions as relevant (e.g., colour) and once with the remaining dimension as relevant (e.g., shape).

Also, because three items were presented on each trial, a total of 54 items were presented across the 18 trials. Therefore, the final choice of the 54 items that were used across the 18 trials, given all of the above-mentioned constraints, was determined such that each of the 27 unique items occurred at least once and no more than three times across the 18 trials. As a result, 6 items appeared only once; 15 appeared twice; and 6 appeared three times.

In addition, unlike Experiment 1, more than one trial-set presentation order was used. Six quasi-random orders of the 18 trial sets were devised with several restrictions pertaining to relevant-dimension pairs and pivot-item placement. First, in terms of relevant-dimension pairs, the quasi-random orders were devised with the restriction that (a) the first three trials include one (and only one) trial set from each of the three possible relevant-dimension pairs (i.e., size and colour, colour and shape, and shape and size) because these three trials served as demonstration and practice trials; (b) identical relevant-dimension pairs not appear on two or more consecutive test trials; and (c) a particular dimension not appear as a relevant dimension (as opposed to a relevant-dimension pair) on more than three consecutive test trials. In addition, for the demonstration and practice trials, each possible order of relevant-dimension pairs occurred only once across the six quasi-random orders (e.g., in Order 1, the order of the relevant-dimension pairs was size and colour, shape and size, and colour and shape; whereas for Order 2, the order was size and colour, colour and shape, and shape and size). Likewise, each trial set (e.g., Trial-set C) was used only once across the six quasi-random orders for the demonstration and practice trials (e.g., Trial-set C was used for the colour and size relevant-dimension pair in Order 1, Trial-set F was used for that relevant-dimension pair in Order 2). Finally, in terms of the pivot-item placement, the restrictions were that for each quasi-random order, the pivot item appeared only once in each window (i.e., Window 1, Window 2, or Window 3) across the demonstration and practice trials and that it appeared in the same window on no more than two consecutive trials on the test trials (see Tables 2 to 5 in Appendix B for more detailed counterbalancing information).

3.2.3. Experimental Design and Procedure

All children were given four tasks that were always presented in the same order: an Item Identification Task, a Favourite Items Task, the FIST, and the Peabody Picture Vocabulary Test-Revised (PPVT-R, Dunn & Dunn, 1981). All children received identical versions of the Item Identification Task, the Favourite Items Task, and the PPVT-R, but they received one of three versions of the FIST (see below). Before administering the tasks, the experimenter put a small sticker on one of the children's fingernails (i.e., the index finger of their dominant hand). This device served as an external reminder for children to use only that finger to select items, and it was designed to reduce the risk that they would inadvertently touch an unintended window. More precisely, the experimenter said,

> You and I are going to play some pick-some-pictures games together. But before we start, I'm going to put a little sticker on my magic pointing finger, so I can remember which finger is my magic pointing finger. And you know what? When I point to pictures in my games, I can only use my magic pointing finger. Now, I'm going to put a sticker on your magic pointing finger, and that's so you remember which finger is your magic pointing finger. And when you point to pictures in my games, you can only use your magic pointing finger. Do you think you can do that? I think you can too!

3.2.3.a. Item Identification Task. The Item Identification Task was included to

~~determine whether children could correctly identify all cues of each dimension used in the~~ FIST (i.e., blue, red, and yellow for colour; boat, shoe, and teapot for shape; and small, medium, and large for size) and to provide them with appropriate labels for each cue and for each dimension in the event that they could not identify any one of these correctly. More specifically, the task consisted of three trials. On each trial, children were presented with a 21 x 28 cm sheet depicting all cues of one dimension (i.e., colour, shape, or size). For example, for colour identification, children were shown a blue, a red, and a yellow rectangle (see Figure 7). Encouraging children to use their magic pointing finger, the experimenter then asked children to identify each cue (e.g., "Show me red.") and provided positive (or negative) feedback to children depending on whether or not their choice was correct (e.g., "Good job! That's right, that's the red one.", or, "Good try but I think that this one here is the red one. What do you think?"). In the few cases in which children misidentified a cue, the experimenter asked them to re-identify the cue after they had identified the other remaining ones. After all cues of a given dimension had been identified correctly, the experimenter mentioned the actual dimension term (e.g., "You really know your colours!"). The experimenter then presented a second and a third sheet displaying cues of the other dimensions (see Figures 8 and 9, which depict the cues of the shape and size dimensions, respectively) and repeated the procedure for each.

Six possible orders of dimension presentation existed (i.e., colour-shape-size; colour-size-shape; shape-colour-size; shape-size-colour; size-colour-shape; size-shape-colour) and children were randomly assigned to one of these orders with the restriction that each order occurred equally often within each age group. In addition, independently for each dimension, children were randomly assigned to one of the six possible orders in which the experimenter

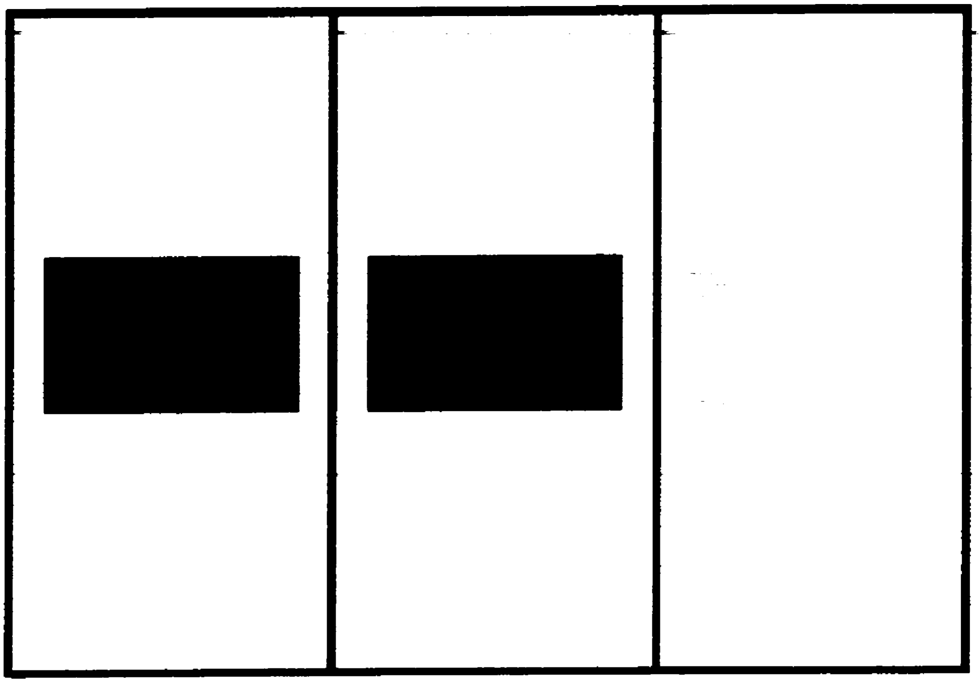

Figure 7. Items presented in the Item Identification Task to assess knowledge of colour terms.

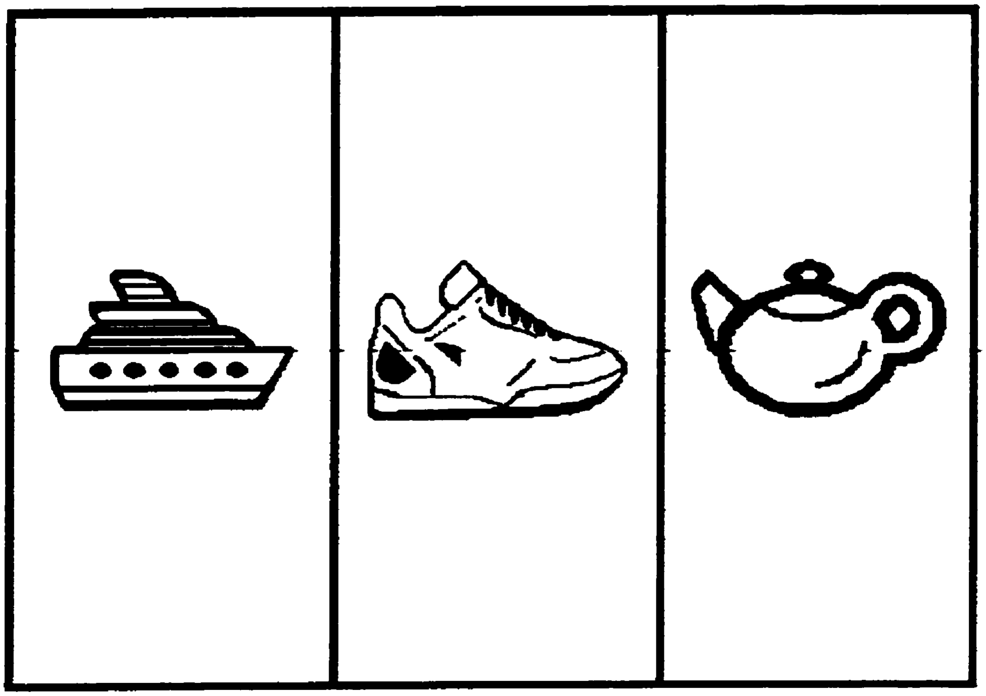

Figure 8. Items presented in the Item Identification Task to assess knowledge of shape terms.

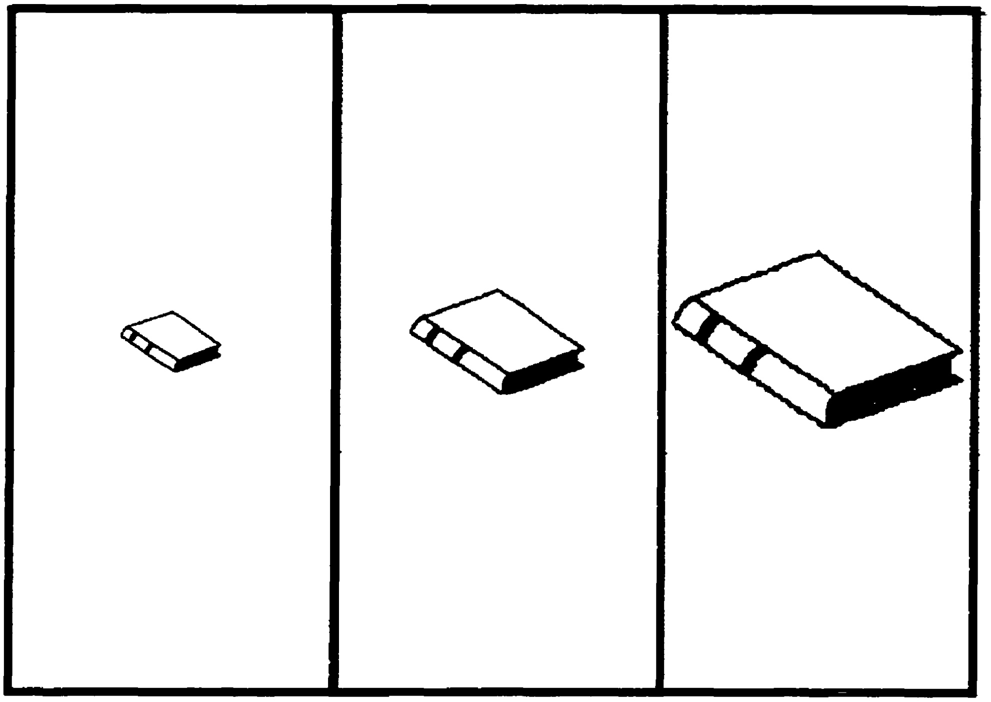

Figure 9. Items presented in the Item Identification Task to assess knowledge of size terms.

could ask them to identify the cues themselves (again with the restriction that each order occurred equally often within each dimension for each age group).

3.2.3.b. Favourite Items Task. The Favourite Items Task, like the FIST, was presented on a Pentium® 166 laptop computer with a 31 cm computer screen equipped with a removable EZTouch® 30.5 cm custom-made touch screen. The Favourite Items Task consisted of one demonstration trial and three practice trials, and on each trial, children were shown three unrelated items (e.g., a heart, a train, and blocks; see Figure 10. for an example). The purpose of this task was to teach children how to select items on the touch screen and to teach them how to select <u>two</u> items.

The experimenter told children,

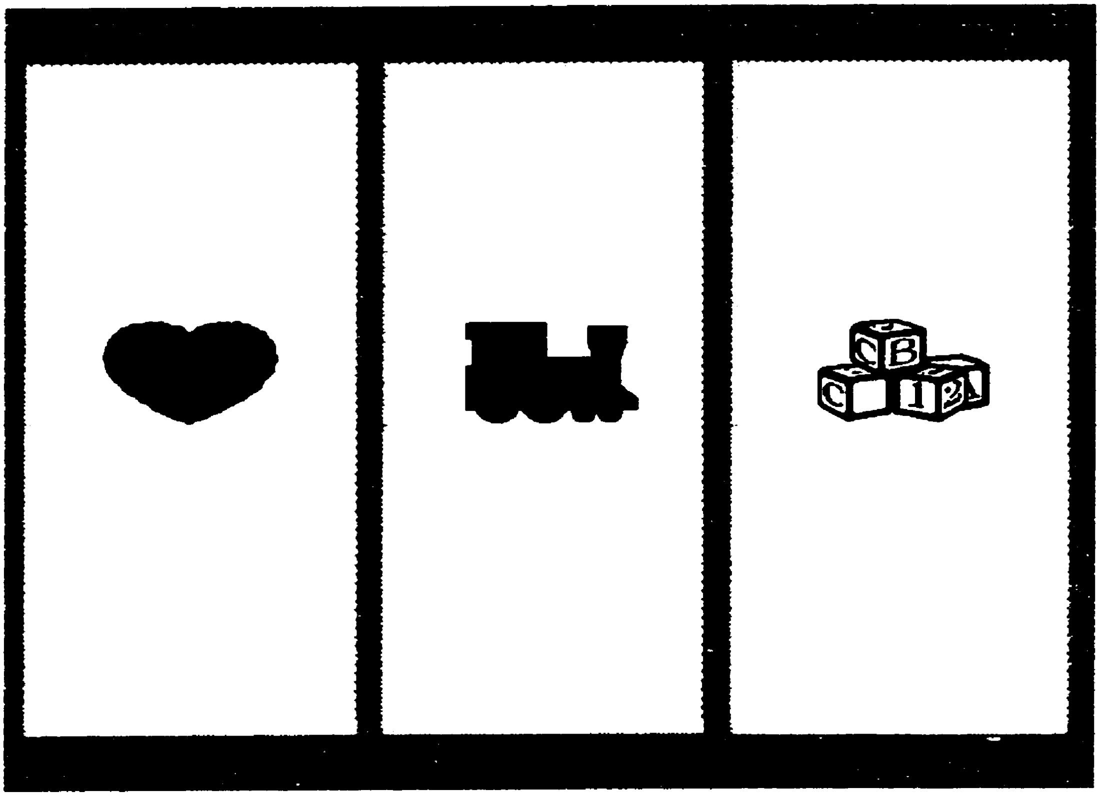

Figure 10. Example of items presented in the Favourite Items Task.

> You and I are going to pick some of our favourite pictures together. I'm going to pick my favourite pictures first, just to show you how we pick our favourite pictures, and then it will be your turn. OK? See, here's a picture, here's another picture, and here's another picture. I'm going to pick my two favourite pictures. So I'm going to put my magic pointing finger on this picture here because that's one of my favourite pictures, and I'm going to put my magic pointing finger on this picture here because that's my other favourite picture. Picture one and picture two. So these two pictures here are my two favourite pictures. I'm not going to touch that picture over there because that's not one of my favourite pictures. I'm only going to touch these two pictures here because these two pictures are my two favourite pictures.

The two items that the experimenter selected on the demonstration trial (and the order in which they were selected) were randomly determined for each child with the restriction that each combination (six possible combinations) occurred equally often within each age group. On each of the three practice trials that followed, three new items were presented on the computer screen. Children were asked to touch their two favourite items with their magic pointing finger. The experimenter intervened (a) if children selected one item by prompting them to select another picture, or (b) if they selected all three by requesting and providing more explicit prompts that they select only two items. Children were given positive feedback when they finally made a correct response (e.g., "So these two pictures are your two favourite pictures? Good job! You didn't touch that picture over there because that's not one of your favourite pictures, is it? No! So you only touched your two favourite pictures, and you only used your magic pointing finger. Good for you!").

3.2.3.c. Flexible Item Selection Task. Children in all three conditions were then given the FIST itself. The experimenter introduced the FIST as a new computer game and began by presenting a demonstration trial to show children how to play the game. That is, the experimenter said,

> Now, you and I are going to play a different computer game. We're going to pick some more pictures together with our magic pointing finger. But we are going to play a different pick-some-pictures game. I'm going to pick some pictures first, just to show you how we pick pictures in this new game, and then it will be your turn. OK? I'm going to pick <u>two</u> pictures that <u>go together in one way</u>.[10] So I'm going to put my magic pointing finger on this picture here and on this picture here because these two pictures here go together in one way. That picture over there doesn't go with these two pictures here, does it? No! So these two pictures here go together in one way.

[10] Note that in Experiment 1, children were asked to select items that "were the <u>same</u> in one way". The word "same" used in Experiment 1, may have inadvertently led children to believe that they had to select identical items. Coupled with the inclusion of criterial trials in Experiment 1, this specific wording may have also contributed to children's difficulties in selecting the pivot item twice. Therefore, even though criterial trials were not included in Experiment 2, the wording was changed as well to the less ambiguous wording of "go together".

> ~~Now, you know what I'm going to do? I'm going to pick~~ two ~~pictures that~~ go together, but in another way. So I'm going to put my magic pointing finger on this picture here and on this picture here, because these two pictures here go together, but in another way. That picture over there doesn't go with these two pictures, does it? No! So these two pictures here go together, but in another way.

The experimenter concluded the demonstration trial by summarizing both selections (i.e., "So see, these two pictures here go together in one way and these two pictures here go together, but in another way."). For half of the children within each age group, condition, and trial order, the experimenter selected items from one of the relevant dimensions on Selection 1 and for the other half, the experimenter selected items from the other relevant dimension.

Children then received two practice trials. These trials were presented in a similar fashion as the demonstration trial except that children were asked to make each selection themselves (e.g., "Now, it's your turn to pick some pictures. I want you to put your magic pointing finger on two pictures that go together in one way".). If children selected a correct pair, the experimenter stated,

> You know what? I think you're right! That's right, these two pictures here go together in one way. That picture over there doesn't go with these two pictures here, does it? No! Good job! So these two pictures here go together in one way. So you're right!

If they made an incorrect selection the experimenter simply said, "That's a good try. But you know what? I think that these two pictures here go together in one way [pointing to the correct pictures]. What do you think? That's right! These two pictures here go together in one way . . .", and continued with the same feedback as when children responded correctly.

Note that the experimenter did not label any items, but if children spontaneously said something about an item or about their selection (or simply spoke), the experimenter acknowledged their utterance by simply repeating it, but without making any statements about the utterance's accuracy or relevance to the task. As in the demonstration trial, the experimenter summarized both selections before proceeding to the next trial.

The three conditions included a no-label condition, a relevant-label condition, and an irrelevant-label condition. Although the demonstration and practice trials instructions were identical for the three conditions, the conditions differed in the instructions presented in the test trials (see Table 4 for a summary of these instructions). At the beginning of each trial, children in all conditions were asked to make a first selection (i.e., "Show me two pictures that go together in one way."). The conditions differed in terms of whether or not the experimenter requested verbal responses from children after they made their selection. In the no-label condition, the experimenter simply proceeded to the next selection (or to the next trial). In the relevant-label condition, children were asked to justify each of their selections. More specifically, the experimenter asked, "Why do these pictures go together?" Children in the irrelevant-label condition were asked about the irrelevant dimension. For example, if shape was the irrelevant dimension, they were asked, "What thing are these pictures?"[11] Analogous instructions were devised for Selection 2.

Thus, the design included three age groups (3 years, 4 years, and 5 years) and three conditions (no label, relevant label, and irrelevant label), and children at each age were assigned randomly to one of the conditions with the restriction that there be equal numbers of girls and boys in each condition (6 girls and 6 boys in each condition, at each age).

[11] The word "thing" was used for the shape dimension instead of the word "shape" because the shapes were in fact real-world objects, and as a result, the interpretation for the term "shape" was ambiguous (i.e., shape could refer to the outline of the object, as well as to the object itself).

Table 4

Instructions Used in the Test Trials of the Flexible Item Selection Task in Experiment 2 for Each Selection as a Function of Condition

Condition	Selection	Instructions
No Label		
	Selection 1:	"Show me two pictures that go together in one way."
	Selection 2:	"Now show me two pictures that go together but in another way."
Relevant Label		
	Selection 1:	"Show me two pictures that go together in one way." [after selection] "Why do these pictures go together?"
	Selection 2:	"Now show me two pictures that go together but in another way." [after selection] "Why do these pictures go together?"
Irrelevant Label[a]		
	Selection 1:	"Show me two pictures that go together in one way." [after selection] "What colour / thing / size are these pictures?"
	Selection 2:	"Now show me two pictures that go together but in another way." [after selection] "What colour / thing / size are these pictures?"

[a] Although the instructions for this condition varied from trial to trial depending on the dimension that was irrelevant on each trial, they were identical for each selection within a particular trial.

Furthermore, as described in the Task Design section (Section 3.2.2.), six quasi-random presentation orders of trial-sets in the FIST were devised. One sixth or 2 children (1 girl and 1 boy) in each age and condition were randomly assigned to each of the quasi-random orders.

3.2.3.d. Peabody Picture Vocabulary Test-Revised. Finally, the experimenter gave children the Peabody Picture Vocabulary Test-Revised (PPVT-R), a standardized test that provides a rough estimate of receptive language skills, and administered it in the standardized manner (see Dunn & Dunn, 1981). Briefly, children were shown a picture book. Four items appeared on each page (e.g., a rope, a zipper, a rake, and a wheel), and children were asked to identify a predetermined item on each page (e.g., "Show me rope."). The task was administered until children failed six out of eight consecutive items, and a raw score was then calculated on the basis of the number of items that they identified correctly (see Dunn & Dunn, for more information on how raw scores are calculated).

3.3. Results

3.3.1. Performance on the Item Identification Task

The Item Identification Task was administered to determine whether children could correctly identify each cue of each dimension on the basis of its respective label, and to provide them with labels in cases in which they failed to identify them correctly. The majority of children at all ages (22 of 36 three-year-olds, 33 of 36 four-year-olds, and 35 of 36 five-year-olds; or 61%, 92%, and 97%, respectively) selected all cues correctly for all dimensions, whereas 14 children made one error, 3 made two errors, and 1 made four errors. In other words, 24 cues were misidentified out of a total of 972 possible identifications (9 identifications per child). The most common error was to misidentify the "medium" cue for size: 11 out of the 24 errors were of this type. Of the remaining 13 errors, 6 were errors in identifying "little", 2 in identifying "big", 1 in identifying "boat", 1 in identifying "teapot", 2 in identifying "blue", and 1 in identifying "yellow". It is clear from these findings that children–particularly the 3-year-olds–were somewhat prone to misidentify the cues of the size dimension, but had virtually no difficulty in identifying the cues of the colour or shape dimensions.

3.3.2. Performance on the Favourite Items Task

The Favourite Items Task was administered to provide children with practice at

~~selecting two (of three) items on a laptop computer monitor equipped with a touch screen~~. The Favourite Items Task included one demonstration trial and three practice trials on which children were asked to select their two favourite pictures and were given feedback on their selections. To be correct on any trial, then, children needed to select any two (but only two) items. Table 5 depicts the number of children at each age who obtained 0, 1, 2, or 3 trials correct. A Fisher's exact test revealed that there was a significant association between age group and the number of correct trials that children obtained (p < .0001): 3-year-olds made significantly more incorrect responses than both 4- or 5-year-olds (ps < .0001), who did not differ from each other (p > .10). Of the 9 children who failed to select two items on every trial, 4 children always selected all three items, 3 children always selected only one item, and the remaining 2 children switched between selecting all items to selecting only one item. Furthermore, of the 11 children who selected two items on two of the three trials, 9 erred on the first practice trials (5 selected all items and 4 selected only one), suggesting that they benefited on subsequent trials from the feedback that they received on this first practice trial.

Table 5

Number of Children Who Obtained 0, 1, 2, or 3 Correct Trials on the Favourite Items Task in Experiment 2 as a Function of Age Group

	Number of Correct Trials			
Age Group	0	1	2	3
3-year-olds	8	1	10	17
4-year-olds	1	1	1	33
5-year-olds	0	0	0	36

Interestingly, the performance of those children who performed correctly on all trials (i.e., those who selected two items on all trials) suggested that there were also differences between the age groups in the strategies that children used to select the items that they considered to be their favourite items. Table 6 shows that the youngest children were more likely than older children to select their favourite items from the same pair of windows on each trial. In fact, a Fisher's exact test revealed that the number of different window pairs from which children selected their favourite items was dependent on age ($\underline{p} < .001$). Again, 3-year-olds differed from both the 4- and 5-year-olds ($\underline{p} < .05$ and $\underline{p} < .001$, respectively), who in turn, did not differ from each other ($\underline{p} > .10$).

Table 6

Number of Children Who Selected Two Items on All Trials of the Favourite Items Task in Experiment 2 as a Function of Age Group and of the Number of Different Window Pairs That They Selected

	Number of Window Pairs Selected		
Age Group	One	Two	Three
3-year-olds	10	7	0
4-year-olds	7	23	3
5-year-olds	4	23	9

Assuming that each item pair had an equal probability of being selected on any given trial,[12] then on the basis of chance alone, the probability of randomly selecting item pairs

[12] This is somewhat of a tenuous assumption because there may have been certain items on any given trial that children strongly favoured over others. However, let us adopt this assumption for the sole purpose of describing differences between the age groups in selection patterns and not for the purpose of making claims

located in the same two windows (e.g., Window 1 and Window 2) across all trials ought to occur only 11% of the time (3 possibilities out of 27 possible combinations of window pairs). Moreover, the probability of selecting item pairs from the same two windows on two of the three trials ought to occur 67% of the time, whereas selecting item pairs from three different window pairs ought to occur 22% of the time. Goodness-of-fit chi-square statistics were then calculated for each age group to determine whether children's selection tendencies differed from these expected values. The number of different window pairs from which children selected their favourite items differed significantly from these expected values for 3-year-olds, χ^2 (2, N = 17) = 40.26, p < .001, it was almost statistically significant for 4-year-olds, χ^2 (2, N = 33) = 5.64, p < .10, but not for 5-year-olds, χ^2 (2, N = 36) = 0.17, p > .10.

3.3.3. Performance on the Flexible Item Selection Task

3.3.3.a. Preliminary analyses. Children in each age group were randomly assigned to each condition, but to ensure that any existing differences between conditions were not due to differences in the age of the participants (within each age group) or to differences in language comprehension, two separate 3 x 3 x 2 (Age Group x Condition x Sex) ANOVAs were conducted using age and PPVT-R raw scores as response measures. As expected, age-group differences were detected for both age, F (2, 90) = 487.30, MSE = 10.45, p < .0001, and PPVT-R raw scores, F (2, 89) = 34.42, MSE = 246.51, p < .0001,[13] but more important, no main effects of condition or sex, nor any interactions were found.

3.3.3.b. Main analyses. The assumption of homogeneity of variance was violated for

about preferences for specific items.

[13] The mean PPVT-R raw score for 3-year-olds in the no-label condition is based on 11 children, because 1 girl refused to finish this task.

Selection 1 responses, as indicated by the Box-Scheffé method. That is, analyses on variances for Selection 1 scores revealed a significant main effect of age group, F (2, 26) = 13.92, MSE = 0.94, p < .01, but no main effect of condition, F (2, 26) = 0.23, p > .10, or interaction between age group and condition, F (4, 26) = 0.95, p > .10. Pairwise comparisons using Tukey's HSD tests revealed that the variances for Selection 1 scores of 3-year-olds were significantly greater than those of 4- and 5-year-olds, but that the variances for Selection 1 scores of 4- and 5-year-olds did not differ from each other. However, this violation is of little concern when the number of observations in each cell are equal because in this situation, the F test is fairly robust to violations of the homogeneity of variance assumption (see Kirk, 1982). Thus, a 3 x 3 x 2 (Age x Condition x Sex) ANOVA on Selection 1 was conducted and it detected a significant main effect of age group, F (2, 90) = 17.70, MSE = 9.99, p < .0001, and an unexpected condition by sex interaction, F (2, 90) = 3.13, p < .05, but no main effect of condition or sex, or other interactions were detected (ps > .10). Pairwise comparisons between age groups using Tukey's HSD tests revealed that 3-year-olds (M = 9.89, SD = 4.56) differed significantly from 4- and 5-year-olds, (p < .05) who did not differ from each other (M = 12.86, SD = 2.64, and M = 14.22, SD = 1.51, respectively; p > .10; see Figure 11). Similarly, the differences between means for 3- and 4-year-olds and 3- and 5-year-olds were large (d = 0.80 and d = 1.28, respectively), whereas the difference between means for 4- and 5-year-olds was medium (d = 0.63).[14] Analyses of simple main effects were also conducted to determine the exact nature of the interaction between condition and sex. A significant main effect of sex was detected within the no-label condition, F (1, 90) = 4.67, p < .05, indicating that boys performed more poorly than girls in that condition (see Figure 12).

[14] This difference between the means of 4- and 5-year-olds is somewhat larger than anticipated on the basis of the findings in Experiment 1 in which Cohen's d was 0.08 for the analogous comparison.

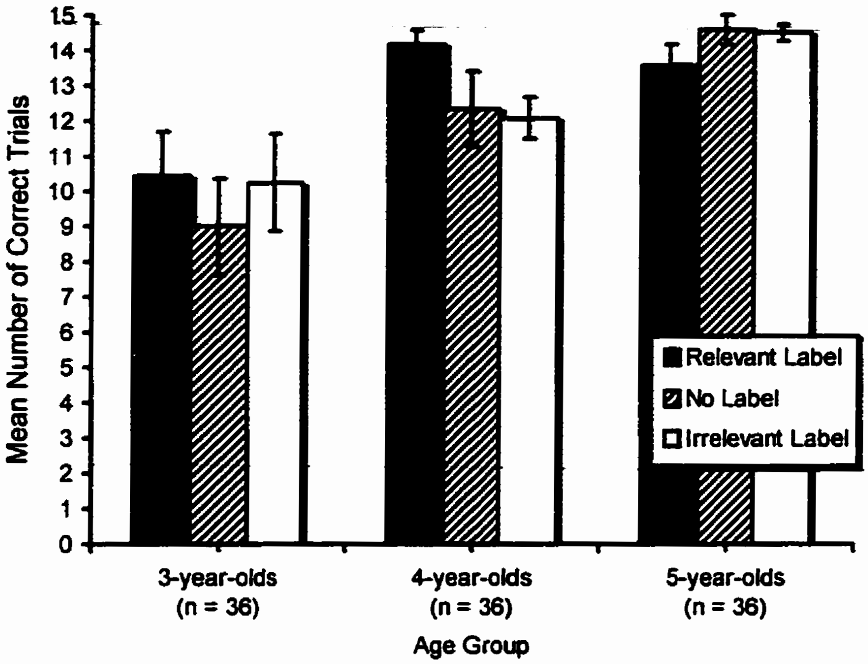

Figure 11. Mean numbers (and standard errors) of correct Selection 1 responses in Experiment 2 as a function of age group and condition (N = 108).

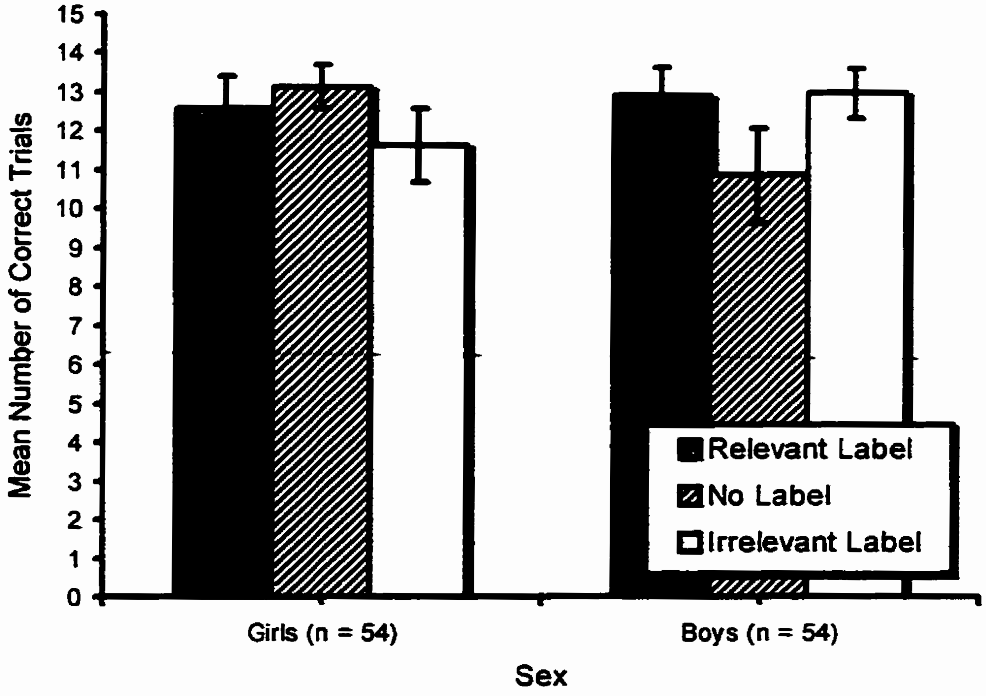

Figure 12. Mean numbers (and standard errors) of correct Selection 1 responses in Experiment 2 as a function of condition and sex (N = 108).

As in Experiment 1, the data of 3-year-olds were omitted from the analyses on Selection 2 performance because of their poor performance on Selection 1 relative to 4- and 5-year-olds. A 2 x 3 x 2 (Age Group x Condition x Sex) ANOVA on the number of correct Selection 2 responses revealed a significant main effect of age _F_ (1, 60) = 10.44, _MSE_ = 12.26, _p_ < .01, a significant age by condition interaction, _F_ (2, 60) = 3.94, _p_ < .05, and a significant age by condition by sex interaction, _F_ (2, 60) = 3.50, _p_ < .05. Overall, 4-year-olds did worse than 5-year-olds (_M_ = 9.97, _SD_ = 4.27 and _M_ = 12.64, _SD_ = 3.36, respectively) and the difference between the two age groups was moderate, _d_ = 0.69 (see Figure 13).

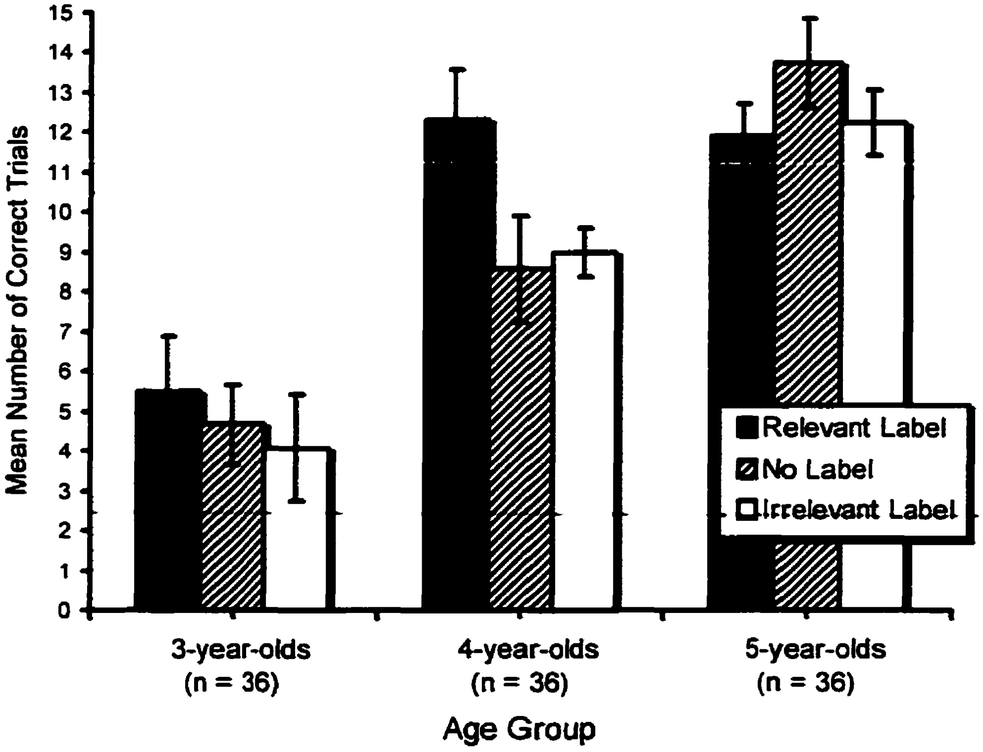

Figure 13. Mean numbers (and standard errors) of correct Selection 2 responses in Experiment 2 as a function of age group and condition (_N_ = 108). (Note that the means and standard errors of 3-year-olds are also presented despite the fact that 3-year-olds were not included in the analyses on Selection 2 performance because of their relatively poor Selection 1 performance.)

Simple main effects were used to examine further the predicted age by condition interaction to determine if, as predicted, 4-year-olds in the relevant-label condition performed better than 4-year-olds in the other conditions and performed as well as 5-year-olds. As predicted, there was a significant main effect of condition for 4-year-olds, F (2, 60) = 4.14, p < .05, but not for 5-year-olds, F (2, 60) = 0.93, p > .10. Conversely, as predicted, there was a significant main effect of age for children in the no-label condition, F (1, 60) = 13.07, p < .01, and for children in the irrelevant-label condition, F (1, 60) = 5.17, p < .05, but no main effect of age for children in the relevant-label condition, F (1, 60) = 0.09, p > .10 (refer to Figure 13 again). In addition, pairwise effect-size estimates were also calculated to determine which differences between pairs of means were substantial: These estimates are presented in Table 7. The results of the simple main effects tests and of the pairwise effect-size estimates together support the prediction that children in the relevant-label condition performed as well as 5-year-olds, and that they performed better overall than the 4-year-olds in the other conditions.

Simple main effects were also used to analyse the unexpected age by condition by sex interaction (see Figure 14). In short, the interaction between age and condition was significant for both girls, F (2, 60) = 4.20, p < .05, and boys, F (2, 60) = 3.23, p < .05. However, the significant main effect of age held only for girls in the irrelevant-label condition, F (1, 60) = 8.81, p < .01, whereas it held only for boys in the no-label condition, F (1, 60) = 12.58, p < .01. Thus, although performance of 4-year-olds in the no-label and irrelevant-label conditions were equivalent overall (and worse than that of 4-year-olds in the relevant-label condition), boys did more poorly within the no-label condition, whereas girls did more poorly within the irrelevant-label condition. Note, however that there were only 6

Table 7

Results of Pairwise Effect-Size Analyses on Selection 2 Performance in Experiment 2 as a Function of Age Group and Condition

	4-year-olds			5-year-olds		
Group	Rel. Lab.	No Lab.	Irr. Lab.	Rel. Lab.	No Lab.	Irr. Lab.
4-year-olds						
Relevant Label	–	d = 0.91**	d = 0.90**	d = 0.10	d = 0.49	d = 0.03
(M = 12.33, SD = 3.47)						
No Label		–	d = 0.10	d = 0.72*	d = 1.44**	d = 0.96**
(M = 8.58, SD = 4.64)						
Irrelevant Label			–	d = 0.68*	d = 1.52**	d = 0.96**
(M = 9.00, SD = 3.91)						
5-year-olds						
Relevant Label				–	d = 0.51*	d = 0.09
(M = 11.92, SD 4.64)						
No Label					–	d = 0.61*
(M = 13.75 SD = 2.09)						
Irrelevant Label						–
(M = 12.25, SD 2.80)						

* medium effect sizes, ** large effect sizes

children of each sex within each of these conditions, so the interpretation of this three-way interaction must be tempered by the small cell sizes on which it is based.

3.3.3.c Response-pattern analyses. As in Experiment 1, Selection 1 and Selection 2 responses were classified according to how children responded (five categories for Selection 1 and seven for Selection 2; see Section 2.3.2.b. for definitions on each category). Overall percentages for each response category were calculated separately for children in each age

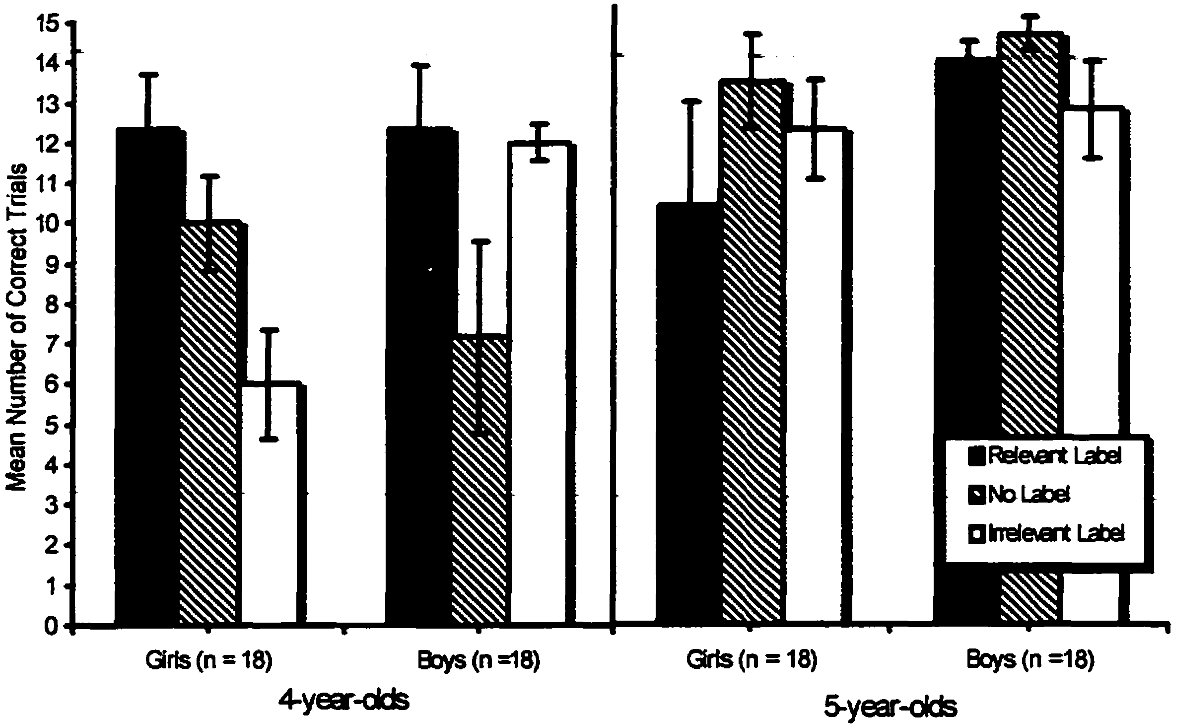

Figure 14. Mean numbers (and standard errors) of correct Selection 2 responses of 4- and 5-year-olds in Experiment 2 as a function of condition and sex (N = 72).

group and in each condition. Table 8 presents the percentages of each type of responses for 3-year-olds as a function of selection and condition. As can be seen from this table, 3-year-olds in all conditions responded correctly on the majority of their first selections, although they made a substantial number of incorrect responses (i.e., 31 to 40% incorrect responses across all conditions). Moreover, in all conditions, their most common type of incorrect responses were all-items responses, followed by one-item responses, and then by wrong-pair responses. In contrast, 4- and 5-year-olds did well on a greater percentage of their first selections (i.e., between 81 to 97% correct Selection 1 responses), and when they erred, they tended to do so by selecting the wrong pair (also see Table 8 for the percentages of 4- and 5-year-old children in each condition).

Three-year-olds performed correctly on approximately one third of their second selections (refer to Table 8 again). The two most frequent types of incorrect responses for

Table 8

Overall Percentages for Each Possible Response Category Across 3-, 4-, and 5-year-olds in Experiment 2 as a Function of Age Group, Selection, and Condition

	Response Categories						
Group and Selection	Correct Pair	Wrong Pair	Same Pair	All Items	One Item	Rem. Item	No Items
3-year-olds							
Selection 1							
Relevant Label	69.4	5.0	–	18.9	6.7	–	0
No Label	60.0	10.6	–	16.7	12.8	–	0
Irrelevant Label	68.3	2.2	–	23.3	6.1	–	0
Selection 2							
Relevant Label	36.7	13.9	16.7	22.2	8.9	1.7	0
No Label	31.1	12.8	30.0	14.4	11.1	0.6	0
Irrelevant Label	27.2	12.2	32.2	21.1	6.7	0.6	0
4-year-olds							
Selection 1							
Relevant Label	94.4	5.0	–	0.6	0	–	0
No Label	82.2	7.8	–	4.4	5.6	–	0
Irrelevant Label	80.6	15.0	–	4.4	0	–	0
Selection 2							
Relevant Label	82.2	13.9	2.7	1.1	0	0	0
No Label	57.2	21.1	6.7	1.7	7.2	6.1	0
Irrelevant Label	60.0	26.7	11.0	0.6	1.1	0	0.6

(table continues)

Group and Selection	Response Categories Correct Pair	Wrong Pair	Same Pair	All Items	One Item	Rem. Item	No Items
5–year-olds							
Selection 1							
Relevant Label	90.6	6.1	–	3.3	0	–	0
No Label	97.2	2.7	–	0	0	–	0
Irrelevant Label	96.7	3.3	–	0	0	–	0
Selection 2							
Relevant Label	79.4	8.3	5.0	3.3	0	3.9	0
No Label	91.7	7.8	0.6	0	0	0	0
Irrelevant Label	81.7	13.9	3.9	0	0	0.6	0

Note. The "Same Pair" and "Rem. [Remaining] Item" response categories applied only to Selection 2 responses, and "One Item" responses for Selection 2 include "other one item" responses (see Section 2.3.2.b. for definitions on each response category). Each child within each age group and condition contributed 15 responses (corresponding to the 15 test trials) for each selection.

children in this age group were all-items responses and same-pair responses. Although 4-year-olds erred more frequently than 5-year-olds on Selection 2–except for 4-year-olds in the relevant-label condition–children in these age groups erred in a similar way (see Table 8 again). That is, as was the case with Selection 1 responses, wrong-pair responses were by far the most common type of incorrect responses exhibited by children in both age groups.

3.3.3.d. Task analyses. As in Experiment 1, analyses were also conducted on specific task-related variables to ensure that age- and condition-related differences noted above were not mitigated by the influence of specific task variables. First, to assess whether specific trial orders affected performance differentially across age groups and conditions, a 3 x 3 x 6 (Age Group x Condition x Trial Order) ANOVA was conducted on both selections summed together.[15] Only a significant main effect of age was found, F (2, 54) = 35.86, MSE = 38.94, p < .0001; no main effect or interactions involving trial order were found (ps > .10). Pairwise comparisons using Tukey's HSD tests revealed that all age groups differed from each other (M = 14.64, SD =7.01, M = 22.83, SD = 6.10, and M = 26.86, SD = 4.25, for 3-, 4-, and 5-year-olds, respectively; all ps < .01).

Second, to determine if performance between groups improved or worsened differentially over trials, the 15 test trials were divided into three equal trial blocks (i.e., the first 5, the middle 5, and the last 5 trials). A 3 x 3 x 2 x 3 (Age Group x Condition x Sex x Trial Block) ANOVA with repeated measures on trial block was conducted. A main effect of age group, F (2, 90) = 42.62, MSE = 10.92, p < .0001, a main effect of trial block, F (2, 180) = 10.55, MSE = 1.48, p < .0001, and an interaction between age group and trial block, F (4, 180) = 3.23, p < .05, were found. Pairwise comparisons between the three trial blocks revealed that children performed better during the first trial block (M = 7.57, SD = 2.54) than they did during the second (M = 7.03, SD = 2.80; p < .01) or third (M = 6.84, SD = 2.94; p < .01) blocks, but that their performance during these latter two trial blocks did not differ (p > .10). Figure 15 depicts performance on each of the three trial blocks as a function of each age group. As one might predict, the analysis of simple main effects revealed that the age-group

[15] Note that sex was not included in this analysis because there were not sufficient degrees of freedom available to include it.

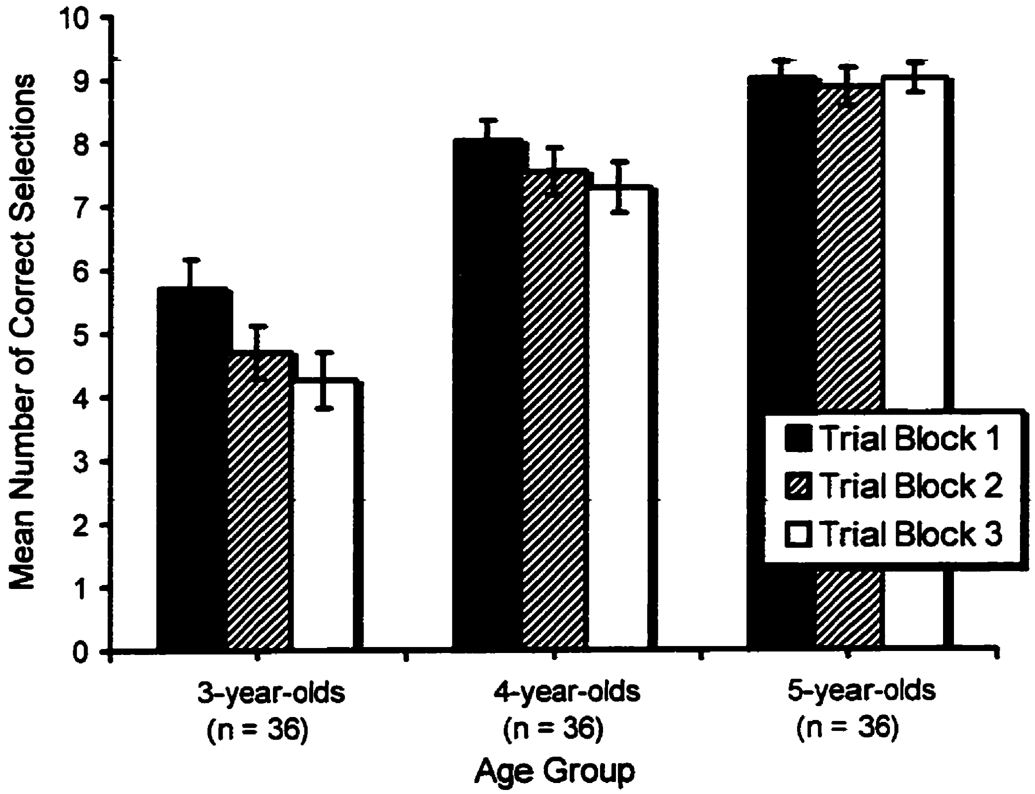

Figure 15. Mean numbers (and standard errors) of correct selections in Experiment 2 as a function of age group and trial block (N = 108).

main effect was significant for each block, F (2, 155) = 21.74, MSE = 4.78, p < .01, F (2, 155) = 34.10, p < .01, F (2, 155) = 43.56, p < .01, for the first, second, and third blocks, respectively. However, the block main effect was significant only for the 3-, and 4-year-olds, F (2, 210) = 14.42, MSE = 1.37, p < .01, F (2, 210) = 3.84, p < .05, respectively, suggesting that their performance deteriorated somewhat over trials but not that of the 5-year-olds.

Third, as in Experiment 1, to determine whether or not children had a preference for selecting items according to certain dimensions more than according to other dimensions on Selection 1, a 3 x 3 x 2 x 3 (Age Group x Condition x Sex x Selection 1 Dimensions) ANOVA with repeated measures on Selection 1 dimensions was conducted. In addition to the expected main effect of age group, F (2, 90) = 17.70, MSE = 3.33, p < .0001, and the condition by sex interaction, F (2, 90) = 3.13, p < .05, a main effect of Selection 1 dimension, F (2, 180) = 26.39, MSE = 7.62, p < .0001, a Selection 1 dimension by sex interaction, F (2, 180) = 3.54, p < .05, and an uninterpretable four-way interaction between age group, condition, sex, and Selection 1 dimension, F (8, 180) = 2.15, p < .05, were found. Pairwise comparisons were conducted to determine the overall order of dimensional preferences. Children were less likely to select items by size (M = 2.54, SD = 2.14) than they were to select items by colour (M = 5.00, SD = 2.82; p < .01) or shape (M = 4.79, SD = 2.73; p < .01) on Selection 1, but they were equally likely to select items by colour or shape first (p > .10). The interaction between Selection 1 dimension and sex was analysed with the use of simple main effects. Both girls, F (2, 212) = 23.99, MSE = 7.62, p < .01, and boys, F (2, 212) = 5.96, p < .01, selected items according to the three dimensions differentially on Selection 1 (see Figure 16). However, girls were less likely than boys to select items by size on Selection 1, F (1, 301) = 4.43, MSE = 6.55, p < .05, despite not differing significantly from boys in their tendency to select items by colour or shape.

Fourth, in contrast to Experiment 1, because the three relevant-dimension pairs and the three pivot-item placement were crossed with each other, it was possible not only to assess the independent contribution of each of these variables to performance, but also to assess their joint contributions. Thus, a 3 x 3 x 2 x 3 x 3 (Age Group x Condition x Sex x Relevant-Dimension Pair x Pivot-Item Placement) ANOVA with repeated measures on both relevant-dimension pair and pivot-item placement was performed. Significant main effects of age group, F (2, 90) = 42.62, MSE = 3.64, p < .0001, relevant-dimension pair F (2, 180) = 5.45, MSE = 0.33, p < .01, and pivot-item placement, F (2, 180) = 12.33, MSE = 0.51, p < .0001, were found in addition to an interaction between sex and pivot-item placement, F (2,

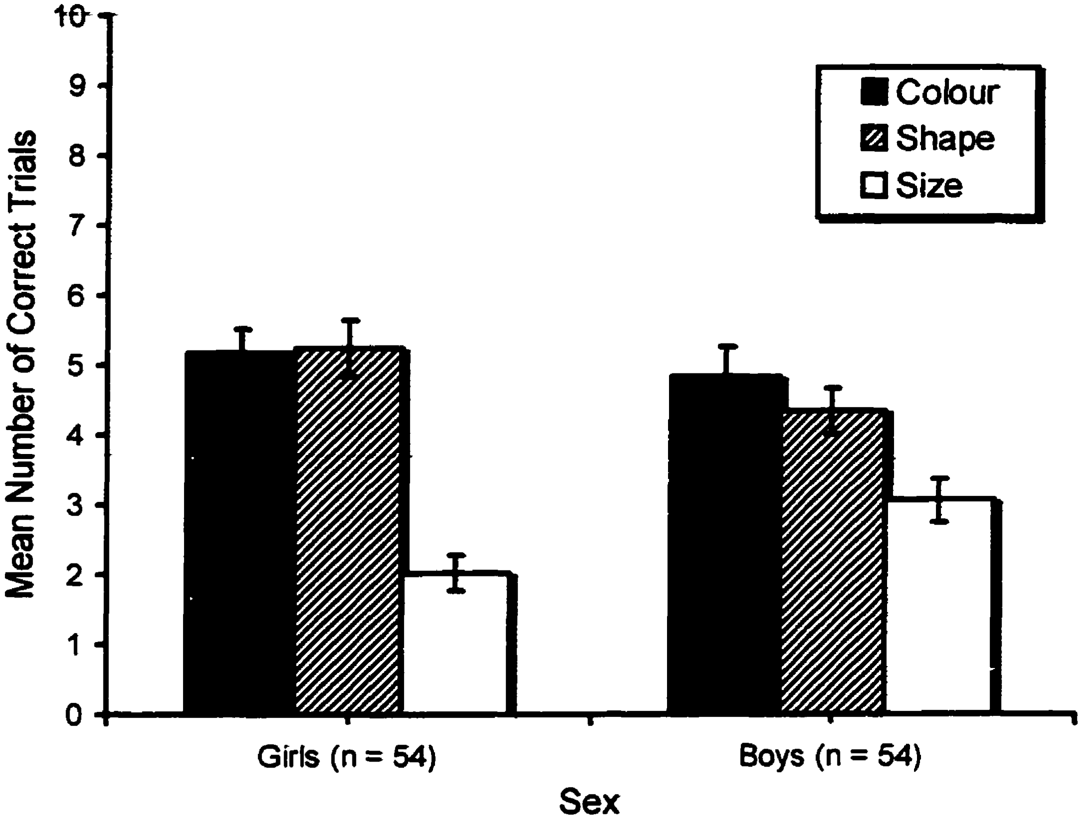

Figure 16. Mean numbers (and standard errors) of Selection 1 responses in Experiment 2 as a function of sex and the relevant-dimension pair according to which children selected items (N = 108).

180) = 3.83, $p < .05$. Pairwise comparisons using Tukey's HSD tests were conducted for the relevant-dimension pair and pivot-item placement main effects to determine which means differed from each other. The significant main effect of relevant-dimension pair was due to the fact that trials in which size and shape (M = 6.89, SD = 2.89) were relevant were more difficult than trials in which either colour and size were relevant (M = 7.30, SD = 2.67; $p < .01$) or colour and shape were relevant (M = 7.26, SD = 2.56, $p < .01$). These latter two pairings did not differ from each other ($p > .10$). Furthermore, as anticipated from the results of Experiment 1, the main effect of pivot-item placement resulted from the fact that children

performed better on trials in which the pivot item was located in the centre window (Window 2; M = 7.61, SD = 2.74) than on trials in which it was located in the left window (Window 1; M = 6.81, SD = 2.78; p < .01) and on trials in which it was located in the right window (Window 3: M = 7.03, SD = 2.85: p < .01). Despite the fact that the main effect of pivot-item placement was significant for both girls, F (2, 212) = 9.18, MSE = 1.59, p < .01, and boys, F (2, 212) = 6.26, p < .01, the sex by pivot-item placement interaction resulted from the fact that girls were less likely than boys to select items from the left window (i.e., Window 1), although the difference was not very reliable, F (1, 140) = 2.86, MSE = 7.76, p < .10 (see Figure 17).

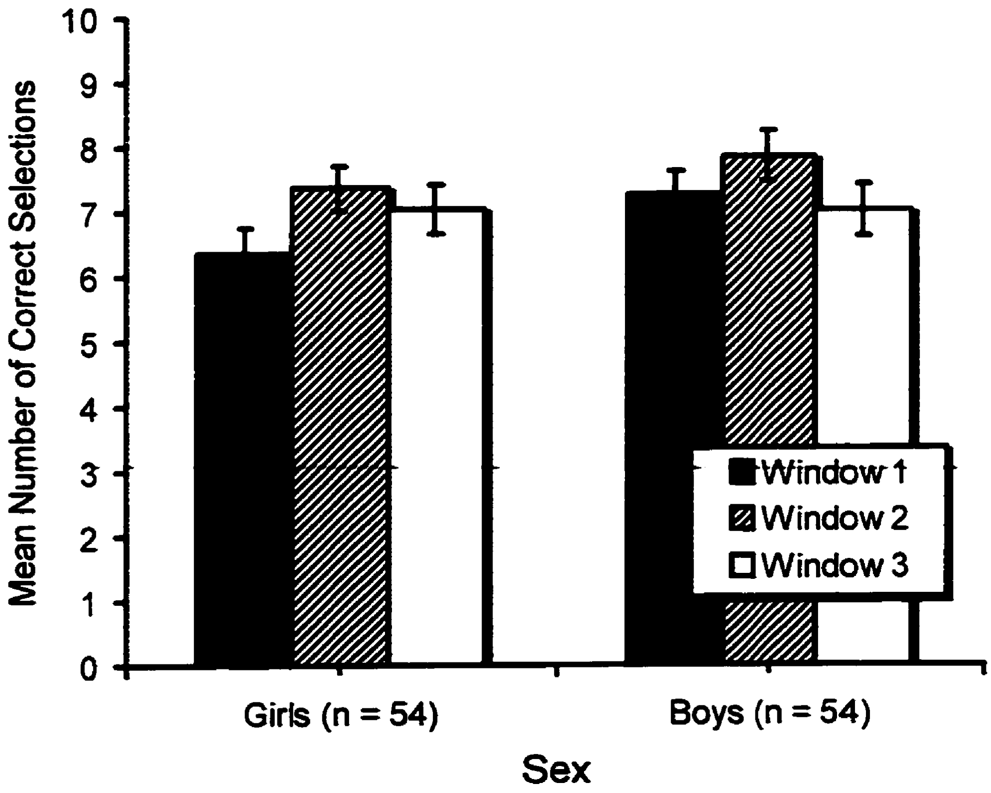

Figure 17. Mean numbers (and standard errors) of correct selections in Experiment 2 as a function of sex and pivot-item placement (N = 108).

3.3.3.e. Labelling analyses. In an attempt to better understand how labels might come to influence performance on the FIST, specific labels that children provided were coded and analysed. Two broad label categories could be differentiated: **spontaneous** and **induced labels**. Spontaneous labels consisted of unprompted utterances that children made about the items (e.g., "It's a red boat!") or about their selection (e.g., "They match because they're both red."). Exploratory analyses on these spontaneous labels are presented in the Spontaneous Label section. In contrast, induced labels were utterances that children in the relevant- and irrelevant-label conditions made in responses to the experimenter's queries in the test trials. Exploratory analyses on these induced labels are presented in the Induced Labels section separately for children in the relevant- and irrelevant-label conditions.

Spontaneous and induced labels were coded in the same way for the purposes of this exploratory investigation. That is, each label was coded into one of four mutually exclusive and exhaustive categories: **relevant labels**, **irrelevant labels**, **wrong labels**, and **other labels**. Relevant labels included any and all labels that referred to the relevant dimension of selected items (e.g., selecting Windows 1 and 2 in Figure 6 and referring to size in some way). Thus, by definition, relevant labels only occurred when the selection itself was correct. In contrast, irrelevant labels included any label that referred to the irrelevant dimension (e.g., selecting Windows 1 and 2 or Windows 1, 2, and 3, and referring to shape in some way), irrespective of whether the selection was correct or not, and irrespective of whether zero, one, two, or three windows were selected. Any label that referred to a dimension other than a relevant or irrelevant dimension was categorized as a wrong label (e.g., selecting Windows 1 and 2 and referring to colour; selecting a wrong pair such as Windows 1 and 3 and referring to colour or size; or selecting all three windows and referring to colour or size), again irrespective of whether the selection was correct or not, and irrespective of whether zero, one, two, or three windows were selected. Finally, other labels encompassed all other task-relevant utterances that did not make specific reference to any of the three possible dimensions (e.g., "they're the same"; "I don't know"; "one, two").

Three points should be noted about these categories. First, labels themselves were not scored for accuracy. For example, children who selected Windows 1 and 2 in Figure 6 were categorized as having provided a relevant label if they referred to size in some way, regardless as to whether they stated that, "they're the same size", "they're both small / little", "they're not big like the other one", "they're the same height", and even, "they're both medium". Children often mislabeled cues, especially cues of the size dimension because size can be construed not only as an absolute dimension in which cues are assigned specific labels, but also as a relational dimension in which cues are defined in relation to other present cues. And indeed, because only one or two cues of the size dimension were present on any given trial, some children used only dichotomized labels ("big" and "little") to refer to size. For this reason, all mislabeled cues were coded as though they had been correct. Second, idiosyncratic labels were categorized in accordance with what children intended to mean rather than what they actually said if it was clear that children used the labels consistently to refer to a specific dimension or cue. For example, 2 children referred consistently to size using the label "age" as in, "they're both the same age". Some children referred to yellow (and only yellow) with terms such as "green", "grey", or "pale". Third, on any given selection, some children provided more than one codable label. In these cases, only the first label was coded and analysed. The decision to analyse only the first label–as opposed to (say) considering all labels and ordering them in order of importance–was based on the fact that, in the case of induced labels, some children who provided multiple labels appeared to not understand the questions that the experimenter asked. Instead, they seemed to be simply describing or listing the characteristics of each items. For example, upon being asked, "Why do these pictures go together?" after selecting Windows 1 and 2 in Figure 6, a child might say, "this one is yellow and little and it's a teapot and this one is a blue teapot and it's little". By only coding the first label (in this case, colour), children who were truly answering the experimenter's question (and therefore consistently identifying the relevant dimension first or alone) would have most of their labels categorized as relevant, whereas children who were simply denoting specific aspects of each item without necessarily answering the experimenter's question would have only a few labels categorized as relevant (i.e., only those trials in which they happened to mention the relevant dimension first).

Spontaneous labels. As mentioned previously, spontaneous labels were unsolicited, task-relevant utterances that children made, irrespective of the condition in which they took part. Further, children in all conditions received identical versions of the demonstration and practice trials, the conditions themselves differed only in the instructions presented in the actual test trials. As a result, it was possible to assess whether spontaneous labelling during these preliminary trials might relate to Selection 1 and Selection 2 performance on the test trials, above and beyond the condition in which children took part. To assess whether children's propensity to label spontaneously the relevant dimension in these preliminary trials was related to their subsequent performance on the actual test trials, children who mentioned a relevant label at least once on any selection during any of these preliminary trials were classified as spontaneous labellers ($\underline{n}$ = 57), whereas those who did not mention a relevant

~~label even once were classified as nonlabellers (n = 51).~~[16] ~~A chi-square test revealed that~~ number of children classified as labellers or nonlabellers was related to age group, χ^2 (2, N = 108) = 6.24, p < .05 (see Table 9). The 3- and 4-year-olds did not differ from each other in terms of their propensity to label spontaneously, χ^2 (1, N = 72) = 0.23, p > .10 , but 3-year-olds did differ from 5-year-olds, χ^2 (1, N = 72) = 5.63, p < .05, and the difference between 4- and 5-year-olds was close to statistical significance, χ^2 (1, N = 72) = 3.66, p < .06.

Table 9

Number of Children in All Conditions in Experiment 2 Who Did (Labellers) or Did Not (Nonlabellers) Spontaneously Produce a Relevant Label in the Preliminary Trials of the Flexible Item Selection Task as a Function of Age Group

	Status	
Age Group	Nonlabellers	Labellers
3-year-olds	21	15
4-year-olds	19	17
5-year-olds	11	25

A 3 x 2 x 2 (Condition x Labeller Status x Sex) ANOVA was then conducted on Selection 1 responses. A significant main effect of labeller status, F (1, 96) =11.05, MSE = 11.28, p < .01, and significant interactions between condition and sex, F (2, 96) = 5.72, p < .01, and between labeller status, condition, and sex, F (2, 96) = 5.37, p < .01, were detected.

[16] Note that nonlabellers also included children who labelled a cue of an irrelevant dimension, a cue of a wrong dimension, or made other task-relevant comments but produced no relevant labels. These children were not included with children who produced a relevant label for the same reason discussed previously for considering only children's first labels.

Overall, labellers performed better on Selection 1 (M = 13.07, SD = 2.93) than nonlabellers (M = 11.49, SD = 4.14) and the difference between these groups was moderate (d = 0.44). As in the analyses on Selection 1 performance presented in the Main Analyses section, simple main effects revealed that the condition by sex interaction was due to the fact that boys in the no-label condition did more poorly overall than girls in the same condition, F (1, 96) = 4.14, p < .05 (refer again to Main Analyses section 3.3.3.b., and to Figure 12).

Analyses of simple main effects showed that the three-way interaction between labeller status, condition, and sex resulted in part from the fact that there was a significant main effect of labeller status for girls in the irrelevant-label condition, F (1, 96) = 4.14, p < .05, and a comparable effect for boys in the no-label condition, F (1, 96) = 14.22, p < .01 (see Figure 18). In addition, boys who did not label in the irrelevant-label condition did better than girls who did not label in the same condition, F (1, 96) = 5.37, p < .05, whereas the opposite was true for boys and girls who did not label in the no-label condition, F (1, 96) = 12.10, p < .01. Finally, there was a significant effect of condition for boys who did not spontaneously label, F (2, 96) = 9.01, p < .01.

Similar results were obtained for Selection 2. That is, the main effect of labeller status was also significant, F (1, 96) = 8.29, MSE = 22.86, p < .01, as was the interaction between labeller status, condition, and sex, F (2, 96) = 5.31, p < .01, although the interaction between condition and sex did not quite reach statistical significance, F (2, 96) = 2.68, p < .08. As was the case with Selection 1, labellers (M = 10.16, SD = 4.87) did better on Selection 2 than nonlabellers (M = 7.96, SD = 5.07) and this difference was also moderate (d = 0.44). Moreover, analyses of simple main effects also revealed that the three-way interaction between labeller status, condition, and sex was due in part to the fact that there was a

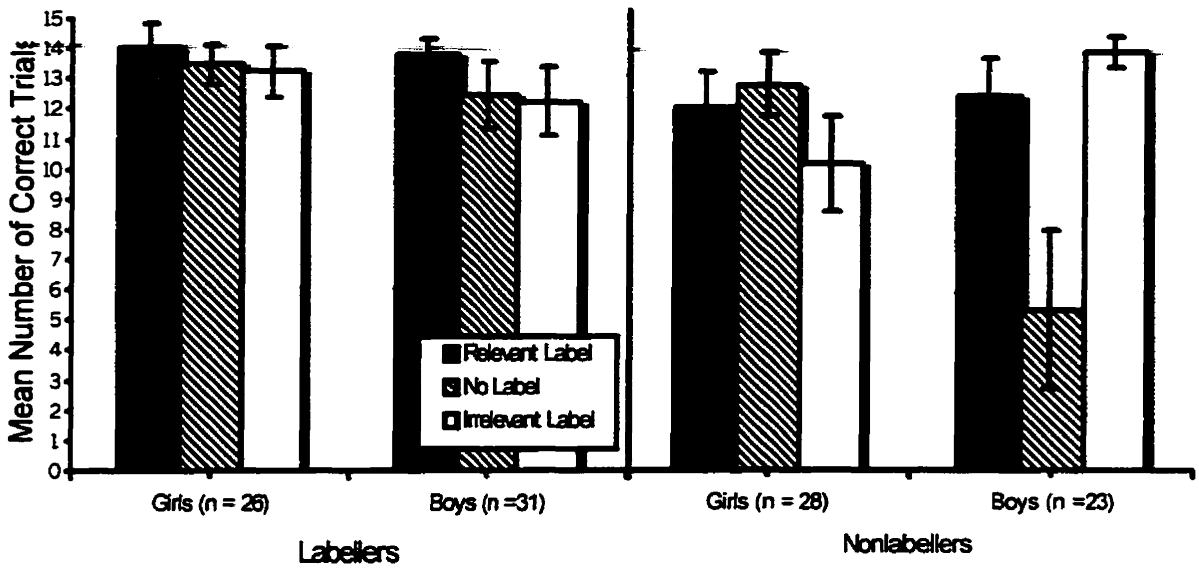

Figure 18. Mean numbers (and standard errors) of correct Selection 1 responses of spontaneous labellers and nonlabellers in Experiment 2 as a function of condition and sex (N = 108).

significant main effect of labeller status for boys in the no-label condition, F (1, 96) = 10.85, $p < .01$, but unlike Selection 1, the main effect of labeller status for girls in the irrelevant-label condition was not significant (refer to Figure 19). As was the case for Selection 1, boys who did not label in the irrelevant-label condition did better than girls who did not label in the same condition, F (1, 96) = 4.29, $p < .05$, and the opposite pattern was again found for boys and girls who did not label in the no-label condition, F (1, 96) = 6.87, $p < .05$. Moreover, as with Selection 1, a significant effect of condition for boys who did not spontaneously label was also found, F (2, 96) = 4.82, $p < .05$.

Induced labels. Results of these analyses are presented separately for children in the relevant-label condition and those in the irrelevant-label condition. Moreover, only Selection 1 labels were examined: Selection 2 labels were not analysed because they followed, and in fact, depended upon Selection 2 performance, and therefore, they provided little in terms of

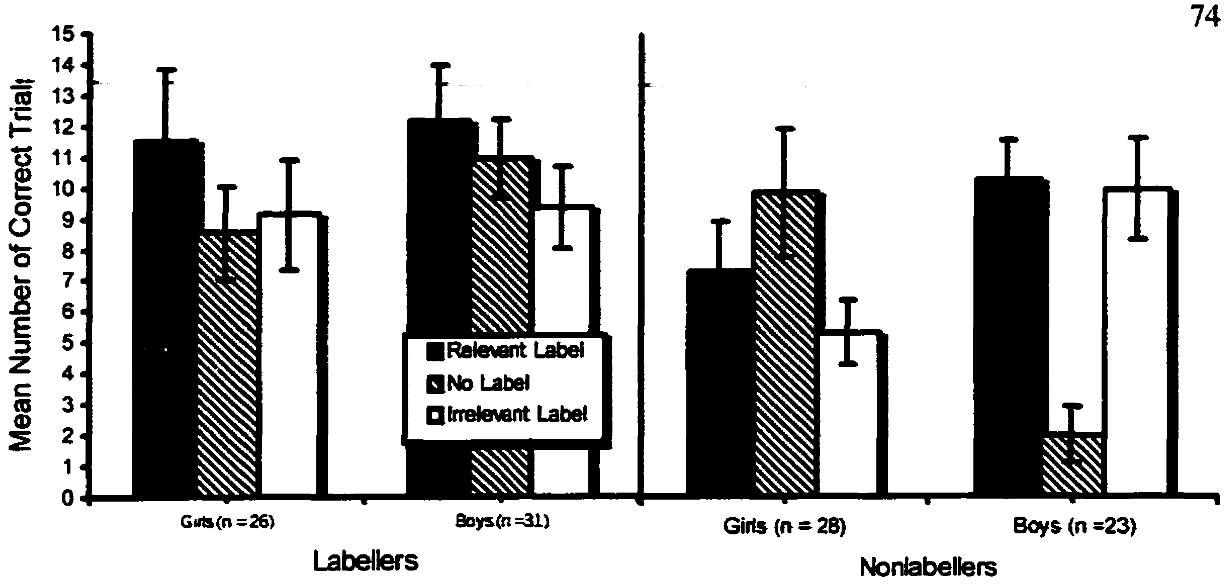

Figure 19. Mean numbers (and standard errors) of correct Selection 2 responses of spontaneous labellers and nonlabellers in Experiment 2 as a function of condition and sex (N = 108).

predictive utility for either selection.

As mentioned previously, relevant labels were defined as those labels that referred to the dimension by which a pair of items were selected. Consequently, for a label to be defined as relevant, then the selection itself had to first be correct. For this reason, assessing the influence of Selection 1 relevant labels on Selection 1 performance was not done because the results would have been uninterpretable. However, it was possible to classify children on the basis of the number of relevant labels that they provided on Selection 1, and then to compare them on their Selection 2 performance. Accordingly, a median split was done to classify children as providing few or many relevant labels on Selection 1. Children with 7 or fewer relevant labels on Selection 1 were classified as "low" in relevant labels ($\underline{n}$ = 19), whereas those with 8 or more were classified as "high" in relevant labels ($\underline{n}$ = 17). A chi-square test indicated that the likelihood of being classified as high or low in relevant labels depended on children's age group, χ^2 (2, $\underline{N}$ = 36) = 8.25, $\underline{p} < .05$ (see Table 10). More precisely, pairwise comparisons between each age group revealed that 3- and 5-year-olds differed from each other in their tendency to be classified as high or low in relevant labels, χ^2 (1, N = 24) = 8.22, p < .01; and that the difference approached significance for 3- and 4-year-olds, Fischer's exact test, p < .10, but not for 4- and 5-year-olds.

Table 10

Number of Children in the Relevant-Label Condition in Experiment 2 Who Were Classified as High or Low in Relevant Labels as a Function of Age Group

	Category	
Age Group	Low	High
3-year-olds	10	2
4-year-olds	6	6
5-year-olds	3	9

A 2 x 2 (Label Category x Sex) ANOVA was conducted to determine whether children who were classified as low in relevant labels on Selection 1 differed from those classified as high on their Selection 2 performance. As predicted, children who were classified as low in relevant labels on Selection 1 did more poorly on Selection 2 (M = 7.74, SD = 5.77) than those who were classified as high (M = 12.35, SD = 3.35), F (1, 32) = 6.58, MSE = 23.52, p < .05 and effect-size analyses indicated that the difference between these groups was large (d = 0.98).

Hence, children who were classified as high in relevant labels did better overall on Selection 2 than those who were classified as low. Likewise, as mentioned in the previous section on spontaneous labels, children who spontaneously labelled the relevant dimension at

~~least once in the preliminary trials also did better on Selection 2 than those who did~~ not, irrespective of the condition in which they participated. Perhaps children who provided many relevant labels when prompted by the experimenter in the relevant-label condition were also more likely to label the relevant dimension spontaneously on the preliminary trials than children who provided few induced relevant labels. Thus, to determine whether children who were classified as high in relevant labels were the same children within the relevant-label condition who spontaneously labelled the relevant dimension in the preliminary trials. children in the relevant-label condition were simultaneously classified on whether they were classified as spontaneous labellers or not, and on whether they were classified as high or low in relevant labels. As shown in Table 11, children who were classified as high in relevant labels seemed equally likely to be classified as labellers and nonlabellers, but those who were classified as low in relevant label appeared to be twice as likely to be classified as nonlabellers than as labellers. Despite this apparent difference. however. no significant

Table 11

Number of Children in the Relevant-Label Condition in Experiment 2 Who Were Classified as High or Low in Relevant Labels as a Function of Whether They Did (Labellers) or Did Not (Nonlabellers) Spontaneously Produce Relevant Labels in the Preliminary Trials of the Flexible Item Selection Task

	Category	
Status	Low	High
Nonlabellers	13	9
Labellers	6	8

relation was detected between these two types of classification schemes, χ^2 (1, N = 36) = 0.91, p > .10.

The spontaneous and induced measures of labels, then, appear to provide somewhat independent measures of labelling effects. Perhaps children who were classified as both high in relevant labels and as spontaneous labellers on the preliminary trials did better overall than children classified favourably into only one of these categories. In turn, perhaps those classified favourably into one of these categories did better on Selection 2 than those classified neither as high in relevant labels nor as spontaneous labellers on the preliminary trials. To determine whether or not this was in fact the case, children were classified into one of four groups (labellers high in relevant labels, nonlabellers high in relevant labels, labellers low in relevant labels, and nonlabellers low in relevant labels), and a one-way ANOVA was conducted on the Selection 2 responses of these four groups. The analysis revealed a significant main effect of group, F (3, 32) = 3.68, MSE = 22.55, p < .05. Pairwise comparisons using Tukey's HSD tests revealed that only the children in the two extreme groups differed significantly from each other: that is, only those children who were classified as labellers and as high in relevant labels differed from children who were classified as nonlabellers and as low in relevant labels (see Table 12 for the means and standard deviations of each group). However, given the apparent differences in the standard deviations between groups and given the low numbers of children within each group, such an analysis may not be entirely appropriate. It is worth noting, however, that being classified as a spontaneous labeller on the spontaneous-labelling measure and as high in relevant labels on the induced-labelling measure almost guaranteed near-perfect performance: no child in this group obtained less than 13 trials correct.

Table 12

Means and Standard Deviations of Correct Selection 2 Responses of Children in the Relevant-Label Condition in Experiment 2 as a Function of Their Simultaneous Classification on Relevant Labels and on Whether They Spontaneously Produce Relevant Labels in the Preliminary Trials of the Flexible Item Selection Task

Groups	*M*	*SD*
High on Relevant Labels		
Labellers (*n* = 8)	14.13^{a}	0.99
Nonlabellers (*n* = 9)	10.78^{ab}	3.96
Low on Relevant Labels		
Labellers (*n* = 6)	8.83^{ab}	6.97
Nonlabellers (*n* = 13)	7.23^{b}	5.37

^{ab}Means with the same letters did not differ from each other (the minimum difference between means for an alpha of .05 was 3.68).

A general problem exists in interpreting the data on relevant labels, however. It is possible that the existing difference on Selection 2 performance between children classified as high and low in relevant labels were not be due to the relevant labels themselves. Instead performance on Selection 1 itself may have caused these apparent relevant-labels effects on Selection 2 performance. That is, by definition, a label could be classified as relevant only if the selection that preceded was correct. Therefore, children who did better on Selection 1 probably also provided a greater number of relevant labels than those who did worse on Selection 1 simply because they had more opportunities to do so then those who did not do as well on Selection 1. As a result, classifying children only in terms of Selection 1 labels (and disregarding Selection 1 performance itself) is problematic because any differences on Selection 2 performance could be attributed not only to group differences in relevant labels.

but also to possible group differences in the number of correct Selection 1 responses children in each group obtained.

However, it is not immediately obvious how best to analyse relevant label while taking into consideration Selection 1 performance. For example, one could examine the performance of only those children who did well on Selection 1. However, it is unclear how best to determine good Selection 1 performance. Differences on Selection 1 performance continue to exist if the criterion for good performance is too lenient, but if it is too strict, too many children might be excluded. Alternatively, one could analyse the data of all children, but only using data from trials on which children responded correctly on Selection 1. However, this kind of analysis would require the use of proportions of trials instead of numbers of trials because children differ on the number of Selection 1 trials on which they respond correctly. The use of proportions instead of numbers is problematic because the overall weight given to a particular relevant label for any particular child will vary greatly depending on the number of trials on which the child responded correctly on Selection 1 initially. For example, a child who selected items correctly on Selection 1 on only two trials, labelled the relevant dimension correctly on only one of these trials, and then selected items correctly on Selection 2 of the one trial in which he or she labelled the items correctly would be scored as correct on 100% of trials in which he or she labelled the relevant dimension, and on 0% of those trials in which he or she did not label the relevant dimension. In contrast, another child who selected items correctly on all 15 trials on Selection 1, labelled the relevant dimension on six trials correctly, and succeeded on two of these correctly labelled trials would be scored as correct on only 33% of trials in which he or she labelled the relevant dimension. With the use of proportion, then, the first child would obtain a proportion of 1.00

for having labelled the relevant dimension correctly and having selected items correctly on Selection 2 on only one trial whereas the other child would obtain a proportion of .33 for having done the same on two trials. Clearly, the conclusions that could be drawn from analyses based on proportion would be tenuous at best.

A means of considering Selection 1 performance without having to use proportion data is to examine the effects of relevant labels on Selection 2 performance by pooling the data of all children in the relevant-label condition together and considering only those trials on which children selected items correctly on Selection 1. However, the disadvantage of such an approach is that no formal analyses could be conducted on these data because trials are not all independent of each other. Nonetheless, the results are interesting, and more important, they corroborate well the analyses presented above in which Selection 1 performance was not considered. Of 540 trials administered to children in the relevant-label condition, children selected items correctly on Selection 1 on 458 of these trials. Furthermore, children identified the relevant dimension correctly on about half of these trials (i.e., on 234 of the 458 trials), and misidentified it on the other half (i.e., 224 trials). As predicted, on trials in which the relevant dimension was identified correctly on Selection 1 (i.e., on 234 trials), children selected items correctly on Selection 2 on 81% of those trials. In contrast, children selected items correctly on Selection 2 on only 58% of those trials in which they had failed to identify the relevant dimension correctly.

Children in the irrelevant-label condition were not asked about their selection per se. Instead, they were asked to identify the irrelevant dimension on each selection (e.g., "What colour are these pictures?", "What thing are these pictures?", or "What size are these pictures?"). For them, then, a correct answer to the experimenter's question consisted of an irrelevant label, not a relevant one. Indeed, to furnish a relevant label in this condition would be incorrect. Consequently, as was done for children in the relevant-label condition, a median split was done between children who provide few or many correct irrelevant labels on Selection 1. Children who identified the irrelevant dimension correctly on 10 or fewer test trials were considered "low" in irrelevant labels ($n = 18$) whereas those who identified it correctly on 11 or more test trials were considered high ($n = 18$). As was the case for children in the relevant-label condition, a chi-square test revealed that the likelihood of being classified as high or low in irrelevant labels was associated with age group, χ^2 (2, $N = 36$) = 8.00, $p < .05$ (see Table 13). The 3-year-olds differed from both the 4- and 5-year-olds (χ^2 [1, $N = 24$] = 6.17, $p < .05$, for each comparison) in their likelihood of being classified as low or high, but 4- and 5-year-olds did not differ from each other ($p > .10$).

Table 13

Number of Children in the Irrelevant-Label Condition in Experiment 2 Who Were Classified as High or Low in Irrelevant Labels as a Function of Age Group

	Category	
Age Group	Low	High
3-year-olds	10	2
4-year-olds	4	8
5-year-olds	4	8

Two 2 x 2 (Label Category x Sex) ANOVAs were then conducted, one using

Selection 1 as the response measure and the other using Selection 2.[17] The analysis revealed that there were no differences on Selection 1 between children who were classified as low in irrelevant labels (M = 11.72, SD = 3.92) and those who were classified as high (M = 12.83, SD = 2.90), F (1, 32) = 0.54, MSE = 11.74, $p > .10$. There was also no significant main effect or interaction involving sex, $p > .10$. In contrast, children who were classified as low in irrelevant labels on Selection 1 did significantly worse on Selection 2 (M = 6.06, SD = 4.19) than those who were classified as high (M = 10.83, SD = 3.79), F (1, 32) = 10.45, MSE = 16.25, $p < .01$. Effect-size analyses corroborate the results of both ANOVAs. That is, the effect size was small on Selection 1 (d = 0.32) but large on Selection 2 (d = 1.19).

Furthermore, to determine if spontaneous and induced measures of labels were independent of each other for children in the irrelevant-label condition–as was found for children in the relevant-label condition–children in the irrelevant-label condition were also concurrently classified on whether they were high or low in irrelevant labels and on whether they were spontaneous labellers or nonlabellers on the preliminary trials. Table 14 clearly shows that there is no correspondence between each type of labelling status and a chi-square test confirmed it, χ^2 (1, N = 36) = 0.44, $p > .10$. To determine whether being classified favourably on both labelling measures resulted in better performance on each selection than being classified favourably on only one measure or on neither, separate one-way ANOVAs were conducted on the Selection 1 and Selection 2 performance of the resulting four groups (labellers high in irrelevant labels, nonlabellers high in irrelevant labels, labellers low in

[17] Unlike relevant labels, irrelevant labels were coded independently of the selections that children made. Thus, it was possible to categorize children in the irrelevant-label condition according to the number of correct irrelevant labels that they provided on Selection 1, compare them on that selection and obtain interpretable results.

Table 14

Number of Children in the Irrelevant-Label Condition in Experiment 2 Who Were Classified as High or Low in Irrelevant Labels as a Function of Whether They Did (Labellers) or Did Not (Nonlabellers) Spontaneously Produce Relevant Labels in the Preliminary Trials of the Flexible Item Selection Task

	Category	
Status	Low	High
Nonlabellers	10	8
Labellers	8	10

irrelevant labels, and nonlabellers low in irrelevant labels). There was no effect of group status on Selection 1 performance, F (3, 32) = 0.87, MSE = 11.99, p > .10, but there was a significant main effect of group status on Selection 2 performance, F (3, 32) = 4.76, MSE = 16.19, p < .01 (see Table 15). Tukey's HSD tests for pairwise comparisons on the means revealed that those children who were classified favourably in at least one way (i.e., either on the spontaneous or induced measures of labelling, or on both) did not differ from each other, and that the groups who were classified favourably on the induced measure of labelling differed from those classified unfavourably on both types of labelling measures (i.e., nonlabellers low in irrelevant labels).

3.3.4. Performance on the Peabody Picture Vocabulary Test-Revised

Several correlations were calculated to determine whether or not performance on the FIST relates to overall receptive language development. First, overall correlations with PPVT-R raw scores were calculated separately for Selection 1 and Selection 2, and they were calculated with and without age partialled out: Table 16 presents the actual values of these

Table 15

Means and Standard Deviations of Correct Selection 2 Responses of Children in the Irrelevant-Label Condition in Experiment 2 as a Function of Their Simultaneous Classification on Irrelevant Labels and on Whether They Spontaneously Produce Relevant Labels in the Preliminary Trials of the Flexible Item Selection Task

Groups	M	SD
High on Irrelevant Labels		
Labellers (n = 10)	10.70^{a}	4.37
Nonlabellers (n = 8)	11.00^{a}	3.21
Low on Irrelevant Labels		
Labellers (n = 8)	7.38^{ab}	4.34
Nonlabellers (n = 10)	5.00^{b}	3.97

^{a,b}Means with the same letters did not differ from each other (the minimum difference between means for an alpha of .05 was 5.17).

correlations, all of which were statistically significant. In addition, Table 16 also presents the correlations between PPVT-R raw scores and Selection 1 performance and between PPVT-R raw scores and Selection 2 performance for each condition, separately–again with and without age partialled out. For all conditions, PPVT-R raw scores correlated significantly with both Selection 1 and Selection 2. However, when age was partialled out, overall receptive language development related only to Selection 2 performance for children in the no-label condition and to both selections for children in the relevant-label condition.

3.4. Discussion

The purpose of Experiment 2 was to begin to explore experimentally the role of language on the development of cognitive flexibility in preschoolers. To do so, labelling

Table 16

Correlations (With and Without Age Partialled Out) Between PPVT-R Raw Scores and Performance on Selection 1 and Selection 2 Across All Children and as a Function of Condition in Experiment 2

Condition	Type of Relation	Selection 1	Selection 2
All (N = 107)			
	Correlation	.51**	.62**
	Age Partialled	.23*	.32**
Relevant Label (n = 36)			
	Correlation	.61**	.63**
	Age Partialled	.41*	.46**
No Label (n = 35)			
	Correlation	.53**	.72**
	Age Partialled	.12	.35*
Irrelevant Label (n = 36)			
	Correlation	.42*	.56**
	Age Partialled	.17	.17

Note. The values for children in the no-label condition are based on 35 children (as opposed to 36) because one 3-year-old in that condition refused to finish the PPVT-R. For this reason, the total N is 107 instead of 108.

* $p < .05$. ** $p < .01$

manipulations were used on a modified and improved computerized version of the FIST. In terms of the basic age-related differences on the task itself, the results of Experiment 2 replicated those of Experiment 1, with one notable exception. That is, as in Experiment 1, 3-year-olds did worse than both 4- and 5-year-olds on Selection 1 and 4-year-olds did worse than 5-year-olds on Selection 2. Therefore, the overall pattern of age-related differences was identical to the one found in Experiment 1. However, the manner in which children tended to err on the task differed across the two experiments. Children in Experiment 2 tended to err by selecting either the wrong pair or the same pair twice on Selection 2, whereas children in Experiment 1 tended to err by selecting the remaining item alone on Selection 2.

This difference in response tendencies may be due to the fact that criterial trials were not given in Experiment 2. Instead, children were given a demonstration trial and two practice trials with trial sets that were identical in form to the trial sets presented in the test trials. As a result, children in Experiment 2 were explicitly shown in the demonstration trial–and given feedback in the practice trials on–how to select two items on each selection (and therefore, how to select one of the items twice). Thus, unlike Experiment 1, because of this explicit training in selecting two items on each selection, children who did not know how a particular specific pair of items matched on a given selection, may nonetheless have known to restrict their responses on each selection to selecting only pairs of items (see also Section 5.1.1., Age Effects, for further discussion on this topic). This difference in procedure may account for why children in Experiment 2 erred on Selection 2 by selecting either wrong or same pairs, whereas those in Experiment 1 (who did not receive such explicit training) erred by selecting the remaining item alone. More important, however, these results taken together suggest that despite the fact that such procedural differences (e.g., including criterial vs. practice trials) may have affected <u>how</u> children manifested their underlying difficulty with switching flexibly between dimensions, these procedural differences did not affect children's overall <u>tendency</u> to experience difficulty. In other words, the findings pertaining to the overall age-related changes in performance on this task appear to be relatively robust.

Interesting results also emerged with respect to the labelling manipulations. As predicted, 3- and 5-year-olds in the relevant-label condition did not benefit from the labelling manipulation when they were compared to children of the same age in the no-label and irrelevant-label conditions. Presumably, this lack of an effect occurred for different reasons across these age groups: On the one hand, it appears as though 3-year-olds did not benefit because they were doing too poorly on the task in the first place. Given that 3-year-olds had difficulty with the abstraction component of the task (as assessed by their Selection 1 performance) and that they did not benefit from the labelling manipulation, perhaps abstraction may be a necessary prerequisite for the emergence of linguistically mediated symbolic thought, and that there is such a thing as abstraction without language (Werner, 1948). These ideas will explored in depth in Section 5.3. in Chapter V.

On the other hand, it seems as though 5-year-olds in the relevant-label condition did not benefit from the labelling manipulation because they were already doing too well on the task. Perhaps the near-ceiling Selection 2 performance of 5-year-olds in all conditions was due to the fact that they were already spontaneously labelling the relevant dimension to solve the task. Indeed, older children–namely, 5-year-olds–were more likely to identify a relevant dimension spontaneously in the preliminary trials than younger children, and in turn, children who made such identifications spontaneously did significantly better on Selection 2 than those who did not. Admittedly, however, it is also possible that developments in other areas of cognitive development may actually underlie the emergence of both spontaneous labelling and cognitive flexibility in 5-year-olds; labelling and flexibility may simply co-exist together rather than exist in a causal relation with each other.

On the contrary, 4-year-olds in the relevant-label condition did benefit substantially from the labelling manipulation in that they did significantly better on Selection 2 than 4-

year-olds in the no-label and irrelevant-label conditions. In fact, their performance was indistinguishable from that of the 5-year-olds. The finding that 4-year-olds in the relevant-label condition benefited from the labelling manipulation, but not those in the irrelevant-labelling condition, suggests that the act of labelling itself or the act of labelling the stimuli themselves were not sufficient in affecting performance. It appears that it was the requirement that children talk about the relevant dimension, the dimension by which they selected items, that was important. Analyses on the induced labels that children provided in the relevant-label condition further suggests that neither was it sufficient for children to be in the relevant-label condition nor was it sufficient for them to be asked about the relevant dimension for their performance to improve. Actually, children needed to identify the relevant dimension correctly most of the time on Selection 1 in order for them to do better on Selection 2. This might also explain in part why 3-year-olds in the relevant-condition did not do better than 3-year-olds in the other conditions. That is, 3-year-olds were more likely than the other age groups to be classified as low in relevant labels. Perhaps because they failed to identify the relevant dimension on most trials on Selection 1, they consequently failed to reap the benefits of participating in such a condition.

The labelling manipulation presented in the irrelevant-label condition was different. In that condition, the experimenter asked children to label the dimension that did not vary across the three items (e.g., "What size are these picture?") and children were only expected to label the cue of that dimension. Therefore, if children were familiar with each dimensional term (i.e., "colour", "thing", and "size"), and with how each cue related to each of these dimensions (e.g., red is a colour, boat is a thing, little is a size), then they ought to have been able to provide accurate labels on all trials, irrespective of how they did on each of their individual selections. Interestingly, those children in the irrelevant-label condition who were classified as high in the number of correct irrelevant labels that they provided did significantly better on Selection 2 than those classified as low, despite equivalent performance on Selection 1. This finding together with the finding that participating in the irrelevant-label condition per se did not help improve children's overall performance at any age suggests that those children in the irrelevant-label condition who did well on Selection 2 already had dimensional information relatively well organized. Perhaps a well organized representational structure relating dimensions and cues is a prerequisite for allowing children to switch flexibly between such dimensions (cf. Kendler, 1963; Smith, 1989a; Werner, 1948; Zelazo & Frye, 1997). I will return to this issue again in the next chapter and in Chapter V.

One unexpected labelling effect was also found. That is, children who spontaneously labelled a relevant dimension in the preliminary trials not only did better on Selection 2, as one might predict, but they also did better on Selection 1. No specific predictions were made with respect to the role of labelling on Selection 1 for several reasons. At the outset, the task itself was not designed to permit an adequate examination of the role of labelling on the abstraction component of the task. That is, Selection 1 labels were elicited by the experimenter <u>after</u> children made their Selection 1 responses. Therefore, in a time-related manner at least, Selection 1 <u>labels</u> were dependent upon children's selections, even though it is still conceivable that Selection 1 responses could be dependent on appropriate labels being covertly generated first. Furthermore, the underlying assumption driving the present series of studies was that it is possible to perform well on Selection 1 (i.e., the abstraction component) without generating relevant labels for the items, but that it is impossible to do well on Selection 2 (i.e., the cognitive flexibility component) without such labels. If relevant labels are necessary for good performance on both of these components, then children who do not produce such labels ought to perform poorly on both selections whereas those who do produce these labels ought to automatically succeed on both selections. Obviously, given the presence of a dissociation between the performance of 4-year-olds on Selection 1 and on Selection 2 in Experiment 1, it is clear that labelling cannot have the same effect on both abstraction and cognitive flexibility. For this reason, no specific predictions were made in regards to labelling effects on Selection 1 other than to assume that relevant labels were not necessary for successful performance on that selection. Note that this assumption, however, does not preclude the possibility that labels can still facilitate or be associated with Selection 1 performance, even if they are not an essential prerequisite for good performance on that part of the task. The finding that children who spontaneously provided relevant labels in the preliminary trials subsequently did better on Selection 1 is consistent with this view. In general, the finding that spontaneous labellers improved on both selections suggests that labelling relevant information does relate with performance on both selection, but this finding on its own does not permit one to determine the precise nature of the relation that exists between spontaneous relevant labels and abstraction, and between these labels and cognitive flexibility. The same problem applies to interpreting the finding that general receptive language development relates to performance on both Selection 1 and Selection 2. More will be said in Sections 5.2. and 5.3. about possible relations that may exist between language, abstraction, and cognitive flexibility.

Across several analyses an unanticipated condition by sex interaction was found: Overall, the results suggest that 4-year-old boys did somewhat more poorly in the no-label condition than girls in the same condition, and a similar pattern was found for 4-year-old girls

~~in the irrelevant-label condition, despite overall equivalent performance across both~~ conditions (and across both sexes). However, given that no differences were anticipated, that no other obvious sex-related differences were found in the other experiments, and that there were only 6 children of each sex at each age in each condition, it is quite possible that this significant interaction occurred as a result of some slight pre-existing differences between the boys who were randomly assigned to the no-label condition and those assigned to the irrelevant-label condition, and between girls randomly assigned to the same conditions (but in the opposite direction). Consistent with this interpretation, if one looks at the mean PPVT-R raw scores of 4-year-olds as a function of sex and condition, the mean PPVT-R raw scores for boys and girls in the no-label and irrelevant-label conditions followed a similar pattern of higher and lower scores as the Selection 2 responses (see Figure 20). Although the differences in PPVT-R raw scores between groups were not statistically significant, PPVT-R raw scores did relate to Selection 2 performance in all conditions. Perhaps scores on Selection 2 may simply have been more sensitive for detecting these pre-existing differences between the groups (i.e., this slight pre-existing difference may have become exaggerated on Selection 2).

The results of the labelling manipulations and of the actual labels that children provided are in line with the general theoretical framework underlying the present work. However, there are remaining difficulties in interpreting these findings. For example, within the relevant-label condition, children not only differed in terms of the actual number of correct selections that they obtained on Selection 1, but also in terms of the actual number of relevant labels that they provided. On any given trial then, the accuracy of children's Selection 2 responses may have been influenced by how they did on Selection 1, how they

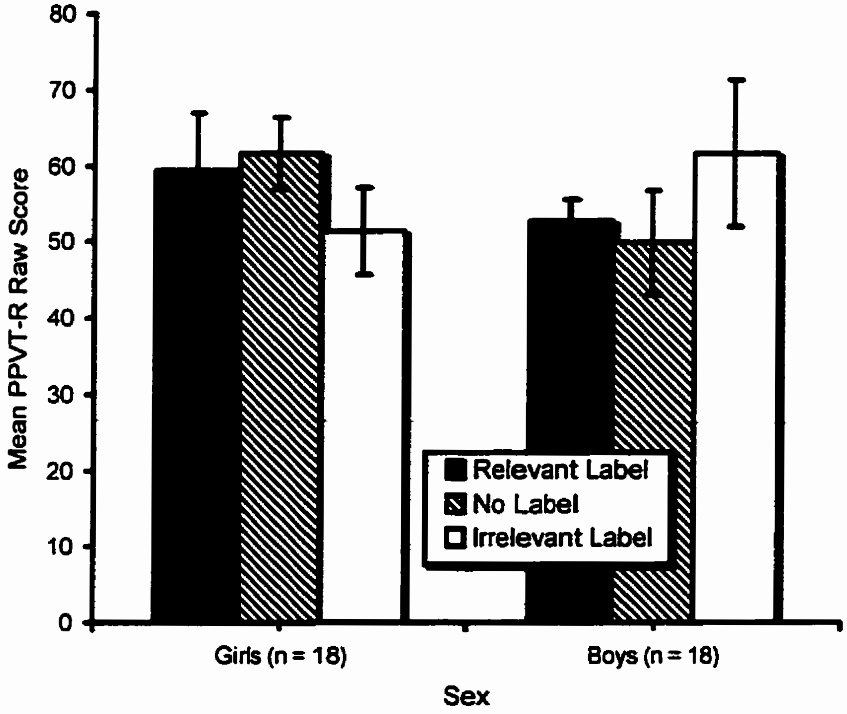

Figure 20. Mean PPVT-R raw scores (and standard errors) of 4-year-olds in Experiment 2 as a function of condition and sex (N = 36).

labelled that selection, or a combination of both of these variables. To properly assess the role of relevant labels on Selection 2 performance, it would be important to ensure that all children were exposed to the same number of correct Selection 1 responses and to the same number of correct relevant labels on Selection 1. One way of doing so is to have an experimenter always select items on Selection 1 (according to a predetermined order), and to have the experimenter provide specific labels depending on the labelling manipulation for each condition.

This kind of change in the procedure would also help address an additional limitation in interpreting Selection 2 responses in the preceding experiments. In both experiments, the interpretation of data rests on the assumption that Selection 1 responses measure abstraction.

whereas Selection 2 responses measure cognitive flexibility. Of course, as mentioned previously, abstraction is also necessary for selecting matching pairs on Selection 2. However, one might reasonably expect that if children were to have difficulty with the abstraction component of the task then they would have difficulty with abstraction on both selections. Therefore, if children succeed on Selection 1, but not on Selection 2, one can conclude that their difficulty with Selection 2 is due to a difficulty with flexibility, not abstraction. However, if children were to always select the dimension that they found easier to abstract on Selection 1, then it is possible that poor performance on Selection 2 might also arise from difficulty with abstraction, one pertaining to difficulty abstracting more difficult–or less preferred–dimensions instead of arising from difficulty with switching between dimensions per se (i.e., cognitive flexibility component). In other words, it is possible that both Selection 1 and 2 responses instead provide measures of the relative ease of (or preference for) abstracting certain dimensions. Further, given that in Experiments 1 and 2, children showed biases for selecting certain dimensions on Selection 1, this interpretation remains a plausible, and problematic, one.

Changing the procedure by having the experimenter instead of children make all Selection 1 responses provides a way of differentiating between these two possible interpretations of Selection 2 responses. That is, if the experimenter always selected items on Selection 1, then presumably, on half of trials, the experimenter would select items consistent with children's preferred dimension, whereas on the other half of the trials, the experimenter would likely select items from children's nonpreferred dimension. Consequently, if Selection 2 responses simply provide a second measure of abstraction abilities, instead of a measure of cognitive flexibility per se, then on trials in which the experimenter selected items consistent

~~with children's preferred dimension, children ought to perform on Selection 2 in a way~~ similar to how they might have responded had they selected the items themselves on Selection 1. However, on trials in which the experimenter selected items on the basis of children's nonpreferred dimension, children ought to perform better on Selection 2 because selecting items according to their preferred dimension would still be a viable option. Thus, if Selection 2 responses measure only the relative ease of abstracting dimensions, then children ought to perform better overall with this type of administration procedure than if the task were to be presented in the standard manner (i.e., children select items on both selections). In contrast, if Selection 2 responses provide a measure of cognitive flexibility (i.e., children's ability to switch between dimensions), then they ought to perform on Selection 2 like children who select items themselves on Selection 1. Even though children may have a bias for selecting certain dimensions first, it does not necessarily follow that they also have more difficulty abstracting less preferred dimensions. Indeed, the finding that children in this experiment performed equivalently on trial sets in which colour and shape were relevant (i.e., two preferred dimensions) as on trials in which colour and size were relevant (i.e., one preferred and one nonpreferred dimension) supports this claim.

CHAPTER IV

EXPERIMENT 3

DIFFERENTIAL EFFECTS OF EXPERIMENTER LABELS ON FOUR-YEAR-OLD'S PERFORMANCE ON THE FLEXIBLE ITEM SELECTION TASK

4.1. Introduction

Only 4-year-olds participated in Experiment 3. Unlike previous experiments, in this experiment, it was the experimenter who always selected items on Selection 1 and who always labelled these items in particular ways. The purpose of assessing children only on their Selection 2 performance was twofold: First, as discussed in Experiment 2, it provides a means of assessing whether Selection 2 performance is indicative of difficulties in switching flexibly between dimensions or of difficulties in abstracting less preferred dimensions. Second, and more important, because of the added control gained from exposing all children only to correct Selection 1 responses and only to correct relevant or irrelevant labels (or no labels at all), the role of labelling on Selection 2 performance could be better evaluated. That is, the finding in the previous experiment that showed that children in the relevant-label

~~condition who tended to label the relevant dimension correctly on Selection 1 did better on~~ Selection 2 then those who did not suggests that for children to benefit from participating in the relevant-label condition, they may actually have to identify the relevant dimension correctly on Selection 1. However, the causal nature of the relation between Selection 1 relevant labels and Selection 2 performance cannot be determined precisely on the basis of the data from Experiment 2 alone because it may be that more advanced children both labelled more accurately and did better on the task without there being a precise causal connection between the two. However, in Experiment 3, by having the experimenter select items on Selection 1 and then label these items (differently depending on the condition) in some systematic way, the causal nature of the relation between Selection 1 labels and Selection 2 performance–if one exists–could be firmly established.

Moreover, an additional aim of Experiment 3 was to determine whether or not presenting children with relevant or irrelevant labels at different dimensional levels would influence their performance. That is, would referring to the dimension (e.g., "colour") or to the cue (e.g., "red") on Selection 1 affect children's performance on Selection 2 differently? In the previous experiment, children in the irrelevant-label condition who provided more correct irrelevant labels performed better on Selection 2 than those who provided few such labels. On the basis of that finding, it was suggested that a good understanding of the relational structure between dimensions and cues was an important precursor for children to switch flexibly between dimensions. Moreover, as alluded to in the previous experiment, several researchers emphasize the importance of differentiating between representing information at the level of the dimension and representing it at the level of the cues themselves (cf. Kendler, 1963; Smith, 1989a; Werner, 1948; Zelazo & Frye, 1997), the former type of representation being identified as the more sophisticated. On these accounts, one might expect that children who are provide with the more abstract dimensional terms might benefit more on Selection 2 than children provided with the terms for the cues themselves.

Hence, children were tested in one of five conditions. In one condition, the experiment did not provide any labels after selecting items on Selection 1. In two of the other conditions, the experimenter referred to the relevant dimension, although in one condition, the experimenter referred to the actual dimensional term (e.g., "colour") and in the other, the experimenter referred to the cue (e.g., "red"). The two remaining conditions were analogous to the two conditions in which the experimenter referred to the relevant dimension, except that the experimenter referred to the irrelevant dimension instead. On the basis of the previous experiment, it was predicted that only relevant labels provided by the experimenter on Selection 1 would affect children's Selection 2 performance, and children provided with relevant dimensional terms were expected to outperform those provided with relevant cue terms.

4.2. Method

4.2.1. Participants

A total of 94 four-year-olds (M = 53.7 months, SD = 3.4 months, range 48.1 to 59.7 months) participated in the experiment (45 girls and 49 boys), although 4 boys were subsequently excluded from the analyses. One boy was excluded because he refused to complete the experiment, 2 boys were excluded because their parents provided information during the experiment, and 1 boy was excluded because of experimenter error (the

~~experimenter inadvertently selected a wrong pair on Selection 1). Children were recruited in~~ the same manner as those who participated in the previous experiments. None of the children who participated in Experiment 3 had participated in Experiments 1 or 2 or in any other pilot experiments with the FIST.

4.2.2. Experimental Design and Procedure

The design of the task was identical to that of Experiment 2 (see Section 3.2.2). In addition, children were tested on the same four tasks that were used in Experiment 2 (i.e., the Item Identification Task, the Favourite Items Task, the FIST, and the PPVT-R), which were presented in the same order as in Experiment 2. The procedures for administering the Item Identification Task, the Favourite Items Task, and the PPVT-R[18] were identical to those of Experiment 2, although the procedure for administering the FIST was somewhat different.

Children were randomly assigned to one of five conditions with the restriction that there be equal numbers of girls and boys in each condition (9 girls and 9 boys in each condition). The conditions differed in terms of the version of the FIST that was administered. As in Experiment 2, all children received 1 demonstration trial, 2 practice trials, and then 15 test trials on the FIST. However, unlike Experiment 2, it was the experimenter who always selected items on Selection 1, and therefore, children in all conditions selected items on Selection 2 only. The five conditions included a **no-cue condition**, a **relevant-dimension**

[18] Due to an experimenter error, PPVT-R raw scores were not obtained for 8 of the children in the experiment (3 in the irrelevant-cue condition, 2 in the irrelevant-dimension condition, and 1 in each of the other conditions). The experimenter had mistakenly assumed that if no basal score (i.e., obtaining 8 consecutive trials correct) could be obtained when administering the task backwards, then it was unnecessary to continue administering the task until a ceiling score had been reached, and therefore, she stopped testing these children without obtaining a ceiling score (see Dunn & Dunn, 1981, for more information on basal scores).

condition, a **relevant-cue condition**, an **irrelevant-dimension condition**, and an **irrelevant-cue condition**. In all conditions, the experimenter began by saying,[19] "I'm going to pick two pictures that go together in one way, so I'm going to pick this picture here and this picture here [the experimenter selects two items correctly on Selection 1]." However, the conditions differed in how the experimenter subsequently characterized each of her Selection 1 response. In the no-cue condition, the experimenter characterized her selection by providing children with nonspecific information. That is, in addition to saying, "I'm going to pick two pictures that go together in one way, so I'm going to pick this picture here and this picture here," the experimenter then added, "These two pictures here go together because they are both the <u>same in one way</u>." In contrast, in both the relevant-dimension and relevant-cue conditions, the experimenter characterized her selection by providing children with a relevant label. More specifically, for children in the relevant-dimension condition, the experimenter instead added, "These two pictures here go together because they are both the same <u>size</u>.", whereas for children in the relevant-cue condition, the experimenter added, "These two pictures here go together because they are both <u>little</u>." The relevant-dimension and relevant-cue conditions differed only in terms of the level at which the experimenter labelled the relevant dimension (i.e., in terms of the dimension itself or in terms of the cue). The irrelevant-dimension and irrelevant-cue conditions were analogous to the relevant-dimension and relevant-cue conditions, respectively, in how items were labelled. However, instead of providing a relevant label, the experimenter provided an irrelevant one. That is, in the irrelevant-dimension condition, the experimenter said, "These two pictures here go together.

[19] For illustration purposes, the examples presented for each type of instruction are based on the experimenter selecting Windows 1 and 2 in Figure 6.

Table 17

Instructions Used in the Test Trials of the Flexible Item Selection Task in Experiment 3 as a Function of Selection and Condition

Selection	Condition	Instructions
Selection 1		
	All Conditions	I'm going to pick two pictures that go together in one way, so I'm going to pick this picture here and this picture here [Experimenter makes Selection 1]. These two pictures here go together . . . **[insert appropriate instructions for each condition; see below]**
	No Cue	**because they are both the same in one way.**
	Relevant Dimension	**because they are both the same size [i.e., the relevant dimension].**
	Relevant Cue	**because they are both little [i.e., the relevant cue].**
	Irrelevant Dimension	**Oh look, they are both the same thing [i.e., the irrelevant dimension].**
	Irrelevant Cue	**Oh look, they are both teapots [i.e., the irrelevant cue].**
Selection 2		
	All Conditions	Now can you show me two pictures that go together, but in another way? [Child makes Selection 2]

Note. The examples are based on the experimenter selecting Windows 1 and 2 in the Example presented in Figure 6.

Oh look, they are both the same thing.", and in the irrelevant-cue condition, the experimenter said, "These two pictures here go together. Oh look, they are both teapots." (see Table 17 for a summary of the instructions presented in each condition). Following the labelling

~~manipulation, the experimenter then asked children in all conditions to select items on~~ Selection 2 (i.e., "Now can you show me two pictures that go together, but in another way?"). Irrespective of the condition, children were never asked to provide any kind of label on either selection.

For the sake of comparison, an attempt was made to maintain as much similarity between Experiment 2 and Experiment 3 in administering the task except for the fact that the experimenter made the first selection and labelled it. Recall that children in the relevant-label condition in Experiment 2 were asked, "Why do these pictures go together?", which required, at least implicitly, a "because" response. In contrast, children in the irrelevant-label condition in Experiment 2 were asked, "What colour / thing / size are these pictures?", which did not require a "because" response. As a result, the instructions presented in the relevant-dimension and relevant-cue conditions and those presented in the irrelevant-dimension and irrelevant-cue conditions were worded differently to parallel the instructions provided in Experiment 2.

In sum, the design of the experiment included five conditions (no cue, relevant dimension, relevant cue, irrelevant dimension, and irrelevant cue) and half of the children in each conditions were girls and half were boys (9 girls and 9 boys in each condition). Further, 3 children within each condition received one of the six possible quasi-random presentation orders. In addition, because the experimenter always made the first selection, the items that the experimenter selected needed to be predetermined in a counterbalanced manner to avoid possible selection biases. To achieve this, relevant-dimension pairs and pivot-item placements were both considered in determining which items the experimenter would select on each trial. That is, relevant-dimension pairs (i.e., colour and size, colour and shape, and shape and size) were each relevant on one third of the trials, and pivot-item placement (i.e.,

~~Window 1, Window 2, and Window 3) was crossed with relevant-dimension pair, resulting in~~ two trials of each combination of relevant-dimension pair and pivot-item placement. For example, there were two trial sets in which colour and size were relevant and in which the pivot item was located in Window 1. On the basis of this counterbalancing, then, the experimenter selected one dimension (e.g., colour) on Selection 1 for one of the two trials and selected the alternated dimension for the other trial (e.g., size). Further, two versions were created in such a way that the experimenter's choices on Selection 1 in one version were the opposite of her choices in the other version (i.e., if colour was selected on one trial and size was selected on the other in Version 1, size would be selected on the former trial and colour on the latter in Version 2). Half on the children received one version, whereas the other half received the other version.

4.3. Results

4.3.1. Performance on the Item Identification Task

As in Experiment 2, the majority of 4-year-olds (66 out of 90) made no errors in identifying all cues of the dimensions used in the experiment, although the percentage of 4-year-olds in Experiment 3 (27%) who did make errors was higher than in Experiment 2 (8%). Twenty children made one error, 3 made two errors, and 1 made three errors. As in Experiment 1, the most common error was to misidentify the "medium" cue for size with 10 of the 29 errors being of this kind. The remaining errors resulted from the fact that 6 children misidentified the "little" cue, 3 misidentified the "big" cue, 4 misidentified the "blue" cue, 1 misidentified the "red" cue, 2 misidentified the "yellow" cue, and 3 misidentified the "boat" cue.

4.3.2. Performance on the Favourite Items Task

The vast majority (80%) of children performed at ceiling on the Favourite Items Task, and there were no differences between conditions (Fischer's exact test, p > .10; see Table 18). Six of the 8 children who failed to select two items on all trials always selected only one item, and of the remaining 2 children, 1 child switched between selecting all items to selecting only one and the other child did the opposite. In addition, 6 of the 8 children who selected two items on two of the three trials erred on the first practice trials (5 selected only one item and 1 selected all items). Further, there were no differences between conditions in children's tendencies to select their favourite items from the same windows (Fischer's exact test, p > .10; see Table 19). However, 4-year-olds who performed correctly on all trials were more prone to select their items from the same windows across all trials than one might expect on the basis of chance alone,[20] χ^2 (2, N = 72) = 24.02, p < .01 (disregarding condition).

Table 18

Number of Children Who Obtained 0, 1, 2, or 3 Correct Trials on the Favourite Items Task in Experiment 3 as a Function of Condition

	Number of Correct Trials			
Condition	0	1	2	3
Relevant Dimension	0	0	0	18
Relevant Cue	1	0	3	14
No Cue	2	1	3	12
Irrelevant Dimension	2	1	1	14
Irrelevant Cue	3	0	1	14
Total	8	2	8	72

[20] Refer to Section 3.3.2. again for an explanation of how expected values were calculated for this test.

Table 19

Number of Children Who Selected Two Items on All Trials of the Favourite Items Task in Experiment 3 as a Function of Condition and of the Number of Different Window Pairs That They Selected

	Number of Window Pairs Selected		
Condition	One	Two	Three
Relevant Dimension	5	11	2
Relevant Cue	7	4	3
No Cue	4	7	1
Irrelevant Dimension	3	9	2
Irrelevant Cue	2	9	3
Total	21	40	11

4.3.3. Performance on the Flexible Item Selection Task

4.3.3.a. Preliminary analyses. As in Experiment 2, to ensure that differences between conditions were not due to possible pre-existing age- or language-related differences between conditions, separate 5 x 2 (Condition x Sex) ANOVAs were conducted using children's age and PPVT-R raw scores as response measures. No condition or sex main effects, or interaction between these were statistically significant for either age or PPVT-R raw scores.

4.3.3.b. Main analyses. To determine whether Selection 2 performance differed between conditions, a 5 x 2 (Condition x Sex) ANOVA on the mean number of correct Selection 2 responses was carried out. As predicted, a main effect of condition was detected, F (9, 80) = 5.17, MSE = 19.47, p < .001 (see Figure 21), but no main effect or interaction involving sex was detected. Pairwise comparisons using Tukey's HSD tests were then conducted: Children in the relevant-dimension condition did better than those in the no-cue

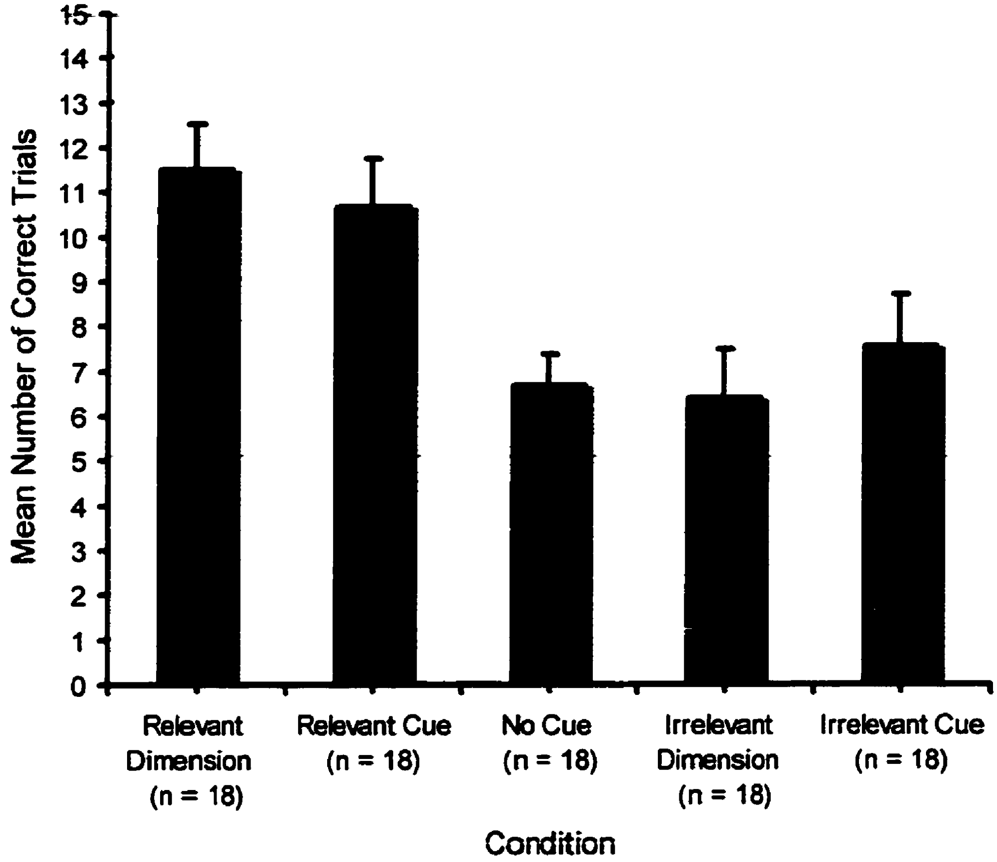

Figure 21. Mean numbers (and standard errors) of correct Selection 2 responses in Experiment 3 as a function of condition (N = 90).

and irrelevant-dimension conditions, (ps < .05), but they did not differ from children in the relevant-cue condition (p > .10). The difference between children in the relevant-dimension condition and children in the irrelevant-cue condition did not quite reach statistical significance (p < .07), despite a large effect size (d = 0.85). Similarly, the difference between children in the relevant-cue condition and those in the no-cue condition did not quite reach statistical significant (p < .06), despite a large effect size (d = 1.04). Children in the relevant-cue condition did do significantly better than those in the irrelevant-dimension condition (p < .06 and p < .05, respectively). However, they did not differ statistically from children in the

Table 20

Results of Pairwise Comparisons (Including Tukey's HSD Tests and Effect-Size Analyses) on Selection 2 Performance Between Each Condition in Experiment 3

	Rel. Dim.	Rel. Cue	No Cue	Irr. Dim.	Irr. Cue
Relevant Dimension					
(M = 11.50, SD = 4.37)					
HSD	–	p > .10	p < .05	p < .05	p < .07
Effect Size	–	d = 0.19	d = 1.29	d = 1.14	d = 0.85
Relevant Cue					
(M = 10.67, SD = 4.56)					
HSD		–	p < .06	p < .05	p > .10
Effect Size		–	d = 1.04	d = 0.93	d = 0.65
No Cue					
(M = 6.67, SD = 2.99)					
HSD			–	p > .10	p > .10
Effect Size			–	d = 0.07	d = 0.22
Irrelevant Dimension					
(M = 6.39, SD = 4.59)					
HSD				–	p > .10
Effect Size				–	d = 0.24
Irrelevant Cue					
(M = 7.56, SD = 4.94)					
HSD					–
Effect Size					–

irrelevant-cue condition (p > .10), as had been predicted, despite a moderate effect size (d = 0.65). Table 20 summarizes the findings of these HSD pairwise comparisons and also includes the corresponding effect size value (ds) of each comparison.

4.3.3.c. Covariate analyses. In an attempt to minimize the variability in the data in order to determine whether all predicted differences between conditions could be detected, an analysis of covariance (ANCOVA) was also performed on Selection 2 performance using PPVT-R raw scores as a concomitant variable. The use of PPVT-R raw scores as a concomitant variable seemed appropriate given that the overall correlation between Selection 2 performance and PPVT-R raw scores was high, r = .55, p < .0001.[21] The analysis revealed that the covariate did account for a substantial proportion of the variability in performance, F (1, 76) = 43.36, p < .0001. In addition, the analysis also revealed a significant main effect of condition, F (4, 76) = 6.94, MSE = 12.94, p < .0001.[22] Pairwise comparisons using Tukey HSD tests on the adjusted means revealed that children in the relevant-dimension and relevant-cue conditions, who did not differ from each other (p > .10), performed significantly better than children in the no-cue, irrelevant-dimension, and irrelevant-cue conditions (ps < .05), who in turn, did not differ from each other (ps > .10; see Figure 22).

4.3.3.d. Response-pattern analyses. As in the previous experiments, Selection 2 responses were categorized into one of seven possible response categories. As shown in Table 21 the way in which children erred differed from the way in which 4-year-olds erred in previous experiments. That is, the predominant pattern of incorrect response for children in all conditions except for children in the irrelevant-dimension condition were same-pair responses which were followed by wrong-pair responses. In contrast, children in the irrelevant-dimension condition erred most often by selecting all items and then by selecting the same pair and the wrong pair (in that order).

[21] PPVT-R raw scores could not be used as a covariate on Selection 2 analyses in Experiment 2 because, as one might predict, PPVT-R raw scores interacted with age group.

[22] Note that preliminary analyses were done to ensure that assumptions underlying the ANCOVA were met (e.g., no differences in the slopes of the regression lines of each condition; i.e., no interaction between condition and PPVT-R raw scores). No problems were detected.

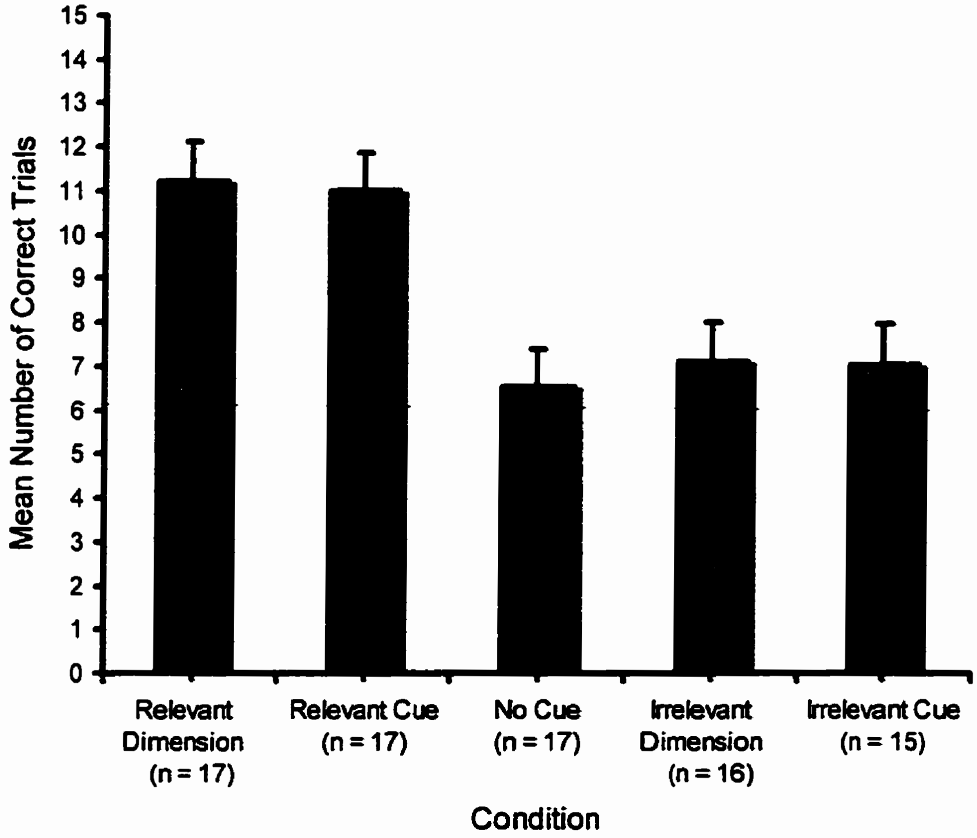

Figure 22. Adjusted means (and standard errors) of correct Selection 2 responses in Experiment 3 as a function of condition (N = 82; adjusted for PPVT-R raw scores).

4.3.3.e. Task analyses. As in Experiment 2, analyses were conducted to determine whether or not specific task variables (i.e., trial orders, trial blocks, dimensional preferences, relevant-dimension pairs, and pivot-item placements) affected performance, and more important, whether potential effects of these task variables differed for children in different conditions. A 5 x 2 x 6 (Condition x Sex x Trial Order) ANOVA detected only a condition main effect, F (4, 30) = 4.75, MSE = 20.23, p < .01, but no other main effect or interaction

~~Table 21~~

Overall Percentages for Each Possible Selection 2 Response Category Across Children in Experiment 3 as a Function of Condition

	Response Categories						
Condition	Correct Pair	Wrong Pair	Same Pair	All Items	One Item	Rem. Item	No Items
Relevant Dimension							
Selection 2	76.7	5.9	11.9	0	5.6	0	0
Relevant Cue							
Selection 2	71.1	10.0	18.9	0	0	0	0
No Cue							
Selection 2	44.4	19.6	28.5	5.9	0.4	1.1	0
Irrelevant Dimension							
Selection 2	42.6	10.7	14.1	22.6	5.6	4.4	0
Irrelevant Cue							
Selection 2	50.4	19.3	23.0	0	3.7	3.7	0

Note. "One Item" responses included "other one item" responses (see Section 2.3.2.b. for definitions of each response category). Each child within each condition contributed 15 Selection 2 responses.

were detected (ps > .10).[23] Likewise (but unlike Experiment 2), a 5 x 2 x 3 (Condition x Sex x Trial Block) ANOVA with repeated measures on trial block revealed a significant main effect of condition, F (4, 80) = 5.17, MSE = 6.49, p < .001, but no main effect of sex or trial block, nor any interactions were evident (ps > .10).

To determine whether on not children were more likely to succeed on Selection 2 when they had the opportunity to select items according to specific dimensions, a 5 x 2 x 3

[23] The result of the pairwise comparisons for the condition main effect are identical to those presented in Section 4.3.3.b. given that only Selection 2 responses were used for these analyses as well.

(Condition x Sex x Selection 2 Dimension)[24] ANOVA with repeated measures on Selection 2 Dimension (i.e., the number of correct selections on the basis of colour, shape, and size) was carried out. Aside from the expected condition main effect, F (4, 80) = 5.17, MSE = 6.49, $p < .001$, no other main effect or interactions were found, ps > .10.

Finally, a 5 x 2 x 3 x 3 (Condition x Sex x Relevant-Dimension Pair x Pivot-Item Placement) ANOVA with repeated measures on both relevant-dimension pair and pivot-item placement was conducted. The condition main effect was again detected, F (4, 80) = 5.17, MSE = 2.16, $p < .001$. Unlike previous experiments, there was no main effect of relevant-dimension pair, F (2, 160) = 1.01, MSE = 0.21, $p > .10$, and the main effect of pivot-item placement was only a trend toward statistical significance, F (2, 160) = 2.99, MSE = 0.31, $p < .06$. No interactions between any of these variables were detected (ps > .10). Pairwise comparisons for the main effect of pivot-item placement revealed that children performed better when the pivot-item appeared in the centre window (Window 2, M = 3.03, SD = 1.73), than when it appeared in the left window (Window 1, M = 2.83, SD = 1.78; $p < .05$) or in the right window (Window 3, M = 2.69, SD = 1.77, $p < .01$), and these latter two placements did not differ from each other ($p > .10$).

4.4. Discussion

In Experiment 3, it was the experimenter who always selected items on Selection 1 in a predetermined and counterbalanced manner, and 4-year-olds were assessed on the FIST on

[24] Note that Selection 2 dimension as opposed to Selection 1 dimension was used as a response measure in this experiment because the experimenter always (correctly) selected items on Selection 1 in a counterbalanced fashion. Therefore, children's preferences for dimensions had to be assessed using what was their first selection.

their Selection 2 performance in one of five conditions. In all conditions, however, the experimenter selected items on Selection 1 in a predetermined manner, and characterized each of these selections in a particular way depending on the condition in which children participated. Children who participated in the conditions in which the experimenter characterized her Selection 1 responses in terms of the relevant dimension (i.e., the relevant-dimension and relevant-cue conditions) performed better on Selection 2 than children in the conditions in which the experimenter either provided no labels (i.e., the no-cue condition) or characterized her Selection 1 responses in terms of the irrelevant dimension (i.e., the irrelevant-dimension and irrelevant-cue conditions). These findings strengthen and extend the conclusions drawn from the previous experiments in at least two ways.

First, as described previously, by having the experimenter select items on Selection 1 in a predetermined manner, it was possible to determine whether Selection 2 responses provide a measure of abstraction difficulties for less preferred dimensions or whether they provide a measure of cognitive flexibility. That is, because the experimenter selected items in a counterbalanced manner, the experimenter ought to have selected items on the basis of children's preferred dimension on about half of the trials and on the basis of children's less preferred dimension on the other half. On the one hand, if Selection 2 responses provide a measure of abstraction difficulties, then children ought to have been more likely to select items correctly on Selection 2 on those trials in which the experimenter selected items on the basis of their less preferred dimension on Selection 1 (because selecting items according to their preferred dimension was still possible). Thus, their overall performance on Selection 2 ought to have been better in Experiment 3 than that of the 4-year-olds in Experiment 2. On the other hand, if Selection 2 responses provide a measure of cognitive flexibility (i.e., the

~~ability to switch between dimensions), then the performance of children on Selection 2 in~~ Experiment 3 ought to have been similar to the performance of 4-year-old who selected items themselves on Selection 1. Indeed, as predicted, the performance of 4-year-olds in Experiment 3 was very much like that of the 4-year-olds in the previous experiments. That is, even when the relative abstraction difficulty of each selection was controlled, 4-year-olds in the no-cue, irrelevant-dimension, and irrelevant-cue conditions continued to exhibit difficulty on Selection 2, supporting the notion that Selection 2 responses do in fact provide an index of children's ability to switch between dimensions and not their ability to detect less preferred dimensions.

Second, as predicted on the basis of previous labelling studies and from those of Experiment 2, 4-year-olds who participated in the conditions in which they were told about relevant dimension performed better than those who were told nothing or told about the irrelevant dimension. The effect of relevant labels on Selection 2 performance is more convincing in Experiment 3 than in Experiment 2. In Experiment 3 the experimenter selected items on Selection 1 and labelled the items in a systematic way and therefore, the labelling effects cannot be attributed to differential Selection 1 performance among the groups, or to differential exposure to certain labels.

Another aim of Experiment 3 was to determine whether or not the level at which dimensional information was labelled (i.e., in terms of the dimensional term or in terms of the cue) significantly affected performance. It is clear from comparing the performance of children in the relevant-dimension and relevant-cue conditions that this variable did not have any observable impact on the performance of children in either of these conditions. Yet, many theories posit in one way or another that representing specific exemplars within the context of their respect conceptual or dimensional structures has implications for behaviour (Kendler, 1963; Smith, 1989a; Vygotsky, 1934/1986; Werner, 1948; Zelazo & Frye, 1997). It stands to reason, then, that proponents of these theories should also predict performance differences on cognitive tasks that relate to the specific level at which items are labelled. That is, labelling items by using their dimensional terms instead of the corresponding cue terms should lead to better overall performance. On these accounts, the finding in this experiment that the level at which the experimenter labelled items dimension did not affect performance is somewhat surprising, although Whitehill (1969) did obtain comparable results using labelling manipulations on the discrimination-shift paradigm (i.e., better performance in relevant-label conditions than in irrelevant-label conditions, but no differences with respect to whether general or specific labels were used within the two types of labelling conditions), and appears to go against these accounts.

However, the act of labelling may itself prompt the usage of relevant existing conceptual structure, irrespective of the specific terms that are used. That is, if labels act by providing children with a means representing information abstractly (i.e., representing information in relation to information contained within existing conceptual structures), then labelling a cue (or a dimension) should necessarily invoke the appropriate dimension itself. Within this theoretical approach, using the cue or the dimensional term should not matter at this age because both should be sufficiently linked that referring to one automatically conjures up the other (cf. Waxman & Gelman, 1986). Viewed in this way, then, it is less surprising that there were no were effects pertaining to the level at which items were labelled. Moreover, the finding in Experiment 2 that children in the irrelevant-label condition who produced the appropriate cue terms for each dimension on most of the trials also performed better on Selection 2 provides some indirect support for this interpretation of the results.

Yet, there was one finding with respect to how children erred that suggested that labelling at the level of the dimension may have had somewhat of an impact on performance. That is, despite the lack of differences in overall performance between children in the irrelevant-cue and irrelevant-dimension conditions, those in the latter condition tended to err by selecting all three items, whereas those in the irrelevant-cue condition tended to err by selecting the wrong or the same pair (as did children in the other three conditions). Perhaps labelling at the level of the dimension made it especially difficult to ignore the irrelevant dimension, leading some children to experience some difficulty in resisting the tendency to select all three items. Clearly, more work needs to be conduct to explore further the impact of dimensional and cue terms on performance on cognitive tasks.

CHAPTER V

GENERAL DISCUSSION

LANGUAGE, ABSTRACTION, AND COGNITIVE FLEXIBILITY

Several interesting findings emerged across the three experiments. In an attempt to make sense of them. the plan of the current chapter is first to provide an overview of the findings. and then to consider them in light of existing explanations that have been proposed to account for labelling effects. I then use an existing theoretical account proposed by Werner to explain the age- and labelling-related changes observed on the FIST. In a subsequent section. I discuss some of the limitations of the current study and propose some avenues for future research. Finally, I conclude the chapter by briefly discussing how the results relate to other existing data.

5.1. Findings

In this section, the findings of the three experiments are reviewed. Age-related

~~changes in performance found in Experiments 1 and 2 are reviewed first, followed by a~~ review of the results of the response-pattern analyses. I then discuss the findings of the task-related analyses to determine if any task-related variables mitigated any of the age- or labelling-related effects. Once these more mundane aspects of the data are discussed, the labelling effects are then reviewed in some detail immediately before I undertake the task of attempting to explain them.

5.1.1. Age-Related Changes

Experiments 1 and 2 assessed performance on slightly different versions of the FIST in different age groups. In Experiment 1, children between the ages of 2 and 5 years were tested and in Experiment 2, the sample was restricted to children between 3 to 5 years of age. The findings across the two experiments with respect to age-related change were similar in some respects, but differed in others.

More specifically, although 3-year-olds in both experiments performed consistently poorly on Selection 1 relative to their older counterparts, the 3-year-olds in Experiment 2 did somewhat more poorly on Selection 1 (correct responses ranged from 60% to 69% across conditions; see Table 8) than those in Experiment 1 (80% correct responses; see Table 2). This difference in overall success rate likely resulted from the fact that all 3-year-olds were included in the analyses in Experiment 2, whereas about one fifth of 3-year-olds in Experiment 1 did not even receive the FIST because of their poor performance on the criterial trials. Had these children also received the FIST in Experiment 1, then they presumably would have performed more poorly on Selection 1 than those who did well on the criterial trials. Indeed, the most common error on Selection 1 for 3-year-olds in Experiment 2 consisted of responses in which children selected more or fewer than two items, suggesting that some 3-year-olds in Experiment 2 did not understand the basic task instructions. In contrast, the most common error on Selection 1 for 3-year-olds in Experiment 1 was to select the wrong pair; few errors on Selection 1 in that experiment resulted from selecting more or fewer than two items. These findings, then, indicate that the inclusion of criterial trials in Experiment 1 was useful in differentiating between 3-year-olds who understood instructions and those who did not. Because criterial trials were not administered in Experiment 2, some 3-year-olds who did not understand the instructions were likely included in the analyses. This may have led to the overall decrease in 3-year-olds' Selection 1 performance in Experiment 2 compared to those in Experiment 1.

On the contrary, the exclusion of the criterial trials appears to have been instrumental in clarifying the interpretation of Selection 2 performance of 4- and 5-year-olds. Unlike 3-year-olds, the Selection 1 performance of 4- and 5-year-olds in all conditions in Experiment 2 was similar to that of 4- and 5-year-olds in Experiment 1. However, 4- and 5-year-olds in Experiment 2 did better overall on Selection 2 than their respective counterparts in Experiment 1, despite the fact that 5-year-olds in Experiment 3 still outperformed 4-year-olds in the no-label and irrelevant-label conditions. The improved performance of 4- and 5-year-olds in Experiment 2 on Selection 2 likely occurred because of differences in the strategies that children used in the two experiments to select items when they <u>did not</u> know how items ought to match. That is, if children who are unable to determine precisely which items go together on Selection 2 simply limited their responses to selecting only two items, then they should be more likely to select items correctly on the basis of chance alone than children who do not use this kind of blind strategy to restrict their manner of responding. As a result, children who use a two-item-only strategy for responding should perform better overall than those who do not use such a strategy, but without a corresponding increase in understanding. The evidence indicates that the use of such a strategy may explain why the 4- and 5-year-olds in Experiment 2 performed somewhat better overall on Selection 2 than the 4- and 5-year-olds in the previous experiment. That is, when 4- and 5-year-olds in Experiment 2 erred on Selection 2, they usually selected two items; whereas when 4- and 5-year-olds in Experiment 1 erred on Selection 2, they tended to do so by selecting the remaining item alone, although children occasionally erred by selecting the wrong or same pair.

Why would children in Experiment 2 have been more likely to limit their responses to selecting only two items than those in Experiment 1? One important difference between the two experiments concerns the fact that in Experiment 2, before the presentation of the test trials, a demonstration and practice trials (with feedback) were presented that included trial sets that were identical to the trial sets used in the task proper (i.e., three items consisting of two pairs of nonidentical matching items), and therefore, they were specifically trained to limit their responses to two items only. As a result, children in Experiment 2 who were unable to detect a second dimension along which two items matched on Selection 2 may have at least understood from these preliminary trials that they were always expected to select *pairs* of items on each selection. In contrast, in Experiment 1, the demonstration trial was identical in form to the criterial trials (i.e., four items consisting of two pairs of identical matching items). Moreover, children never received practice trials, and as a result, they never received feedback on any of their responses. Consequently, because children in Experiment 1 never received any direct training on the FIST in how to limit the possible ways in which they could respond, children who did not know how to respond correctly on Selection 2 had a wider range of possible response options at their disposal (i.e., selecting zero, one, two, or three items). Indeed, the fact that the majority of errors that children made in Experiment I were remaining-card responses instead of two-item responses supports this claim.

It is not surprising, then, that the overall number of correct Selection 2 responses was higher in Experiment 2 than it was in Experiment 1. However, this improvement does not appear to have been accompanied by an actual increase in 4-year-olds' tendency to be flexible. In fact, in Experiment 1, the mean number of correct Selection 2 responses that 4-year-olds obtained was approximately 64% of that of 5-year-olds (the means of 4- and 5-year-olds in Experiment 1 on Selection 2 were 3.79 and 5.94, respectively). Likewise, the mean number of correct Selection 2 responses that 4-year-olds in the no-label condition in Experiment 2–the condition that best matched the one presented in Experiment 1–obtained was only 62% (the means of 4- and 5-year-olds in the no-label condition on Selection 2 were 8.58 and 13.75, respectively).

5.1.2. Response Patterns

Although the overall age-related findings were similar across experiments, the way in which children–specifically, 4-year-olds–tended to err on Selection 2 varied widely across experiments. For example, in Experiment 1, 4-year-olds tended to err by selecting the item that they had not selected on Selection 1; in Experiment 2, they tended to err by selecting the wrong pair; in Experiment 3, they tended to err by selecting the same pair that the experimenter selected on Selection 1 (although wrong-pair errors were also frequent).[25]

As discussed in earlier sections, 4-year-olds' tendency to err by selecting the remaining item in Experiment 1 and to err by selecting the wrong pair in Experiment 2 may

[25] Children in the irrelevant-dimension condition are a notable exception: they behaved differently in that they tended to select all three items when they erred. See Section 4.4. for a discussion of this finding.

have been due to the inclusion of criterial trials in the former case and the inclusion of practice trials in the latter case. That, is, on the one hand, the inclusion of criterial trials at the beginning of the testing session in Experiment 1 may have inadvertently led children to assume that each item (card) ought to be selected only once. On the other hand, as discussed in the previous section, the inclusion of practice trials in Experiment 2 might also have increased 4-year-olds' tendency to limit their responses to selecting two (and only) items. In contrast, the tendency of 4-year-olds in Experiment 3 to err by selecting the same pair almost as often as the wrong pair may have been due to the tendency of some children to repeat the experimenter's choices in the face of uncertainty.

The bottom line, though, is that the modifications introduced in the administering of the FIST across the different experiments (e.g., inclusion of criterial trials vs. practice trials, child vs. experiment selected items on Selection 1) appear to have had some effect on how children manifested their tendency to experience difficulty on Selection 2 (e.g., selecting the remaining item, same pair, vs. wrong pair). More important, however, is that these modifications did not seem to affect whether children experienced difficulty per se. It appears, then, that the finding regarding 4-year-olds' difficulty with the cognitive flexibility component of the FIST is not only relatively robust, but also not simply the result of the specific way in which the task was administered.

5.1.3. Effects of Task-Related Variables

Several task-related variables were examined in each experiment to determine whether any of these contributed significantly to children's performance on the FIST. More important, the analyses were conducted to ensure that the age- and condition-related differences were not mitigated by differential effects of specific task-related variables across

~~ages and conditions~~. ~~Table 22 summarizes all significant effects detected in each experiment~~ that involved task-related variables. Several points are worth noting about these effects. First, as one might hope, no effects related to the order in which the trial sets were presented were detected. In addition, no effects of trial blocks were detected except for Experiment 2. The trial-block effect in Experiment 2 was due to the fact that 3- and 4-year-olds did worse on later trials than on earlier trials. However, even though fatigue appears to have played a small role in influencing the performance of the younger children, fatigue on its own cannot account for all age-related changes in performance seen on the task because age-related difference were detected on all blocks (i.e., if fatigue alone accounted for the age-related changes in performance, differences between age groups would have been detected only on later blocks; refer to Figure 15 again). Furthermore, the finding that performance deteriorated instead of improving over trials suggests that practice and learning did not contribute significantly to performance on this task or to the age- and condition-related differences that were found.

Second, relevant dimensions and the pairings of these relevant dimensions appear to have had an influence on children's performance in Experiments 1 and 2, but not in Experiment 3. For example, in both Experiments 1 and 2, children selected items more often on the basis of colour and shape first than on the basis of size (and number in Experiment 1) on Selection 1. Likewise, in both of these experiments, children did better on trials in which colour and size or colour and shape were relevant than on trials in which shape and size were relevant (pairings involving number in Experiment 1 were intermediate in difficulty; see Table 3).

Table 22

Comparison Across Experiments of Significant Main Effects and Interactions Involving Task-Related Variables on Performance on the Flexible Item Selection Task

	Task-Related Variables				
Experiment	Trial Orders[a]	Trial Blocks[b]	Dimensional Preferences[c]	Relevant Dimension Pairs	Pivot-Item Placements[d]
Experiment 1					
Main Effects	—	n.s.	shp > col > num > siz[e]	see Table 3 for ordering	Card 2 > (1 = 3)
Interactions[f]	—	n.s.	n.s.	n.s.	Age Group x Placement Age Group x Sex x Placement
Experiment 2					
Main Effects	n.s.	Blk 1 > (2 = 3)	(col = shp) > siz	(col-shp = col-siz) > shp-siz	Win. 2 > (1 = 3)
Interactions	n.s.	Age Group x Block	Sex x Dimension Age Group x Condition x Sex x Dimension	n.s.	Sex x Placement
Experiment 3					
Main Effects	n.s.	n.s.	n.s.	n.s.	Win. 2 > (1 = 3)
Interactions	n.s.	n.s.	n.s.	n.s.	n.s.

[a] The trials were presented in the same order for all participants in the version of the FIST used in Experiment 1, and therefore, no analyses were conducted on a trial-order variable in that experiment (in contrast, there were six possible presentation orders of trial sets in the computerized version used in subsequent experiments).

[b] In the version of the FIST used in Experiment 1, there were two trial blocks of six trials each, whereas in the

~~computerized version used in subsequent experiments, there were three trial blocks, each with five trials.~~

[c] Dimensional preferences were assessed on Selection 1 responses for children in Experiments 1 and 2, and on Selection 2 responses for children in Experiment 3 because the experimenter selected items for Selection 1 in this last experiment (and therefore, in Experiment 3, dimensional preference was synonymous with overall difficulty of each type of trial set).

[d] Cards (Cards 1, 2, or 3) instead of windows (Windows 1, 2, and 3) were used to denote the possible placements of the pivot item in Experiment 1.

[e] Four dimensions (colour, shape, size, and number) were varied in the version of the FIST used in Experiment 1, whereas only three of these (colour, shape, and size) were varied in the computerized version used in subsequent experiments.

[f] See text in relevant sections for a description of each of the two-way interactions; no attempt was made in the text to interpret the three and four-way interactions noted in the table.

Two notable differences exist between Experiments 1 and 2, and Experiment 3 that might explain the presence of these relevant-dimension effects in the first set of experiments but not in the latter ones. First, Experiments 1 and 2 differed from the other experiments in terms of who selected items on Selection 1: In Experiments 1 and 2, children selected items themselves on both selections; whereas in Experiment 3, the experimenter selected items on Selection 1 in a counterbalanced manner and children only selected items on Selection 2. Because it was the experimenter who selected items on Selection 1 in Experiment 3, it was only possible to detect children's preferences for specific dimensions in these experiments if these preferences translated into an increase in difficulty in selecting items according to less preferred dimensions. That is, because the experimenter selected items on Selection 1 in Experiment 3 in predetermined manner, on Selection 2 children had no other options but to select items on the basis of the remaining relevant dimension or to err. As discussed in Experiment 3, there is no evidence that children have more difficulty abstracting less preferred dimensions. Therefore, it is not surprising then that no dimensional preferences were detected for children in these experiments.

In Experiment 3, unlike the previous experiments, children did not have more difficulty on trials in which shape and size were relevant dimensions than on trials in which relevant dimensional pairings were used(e.g., colour and size). However, it is not immediately obvious how the fact that it was the experimenter, and not children, who selected items on Selection 1 in Experiment 3 might explain this failure to find significant differences between relevant dimension pairs in Experiment 3. Perhaps it is the other difference that exists between the first two experiments and the last experiment that explains this difference. In Experiments 1 and 2, 3-, 4-, and 5-year-olds were assessed on the FIST, but only 4-year-olds were tested in the third experiment. Perhaps it is the youngest children who contributed most to these dimension-related effects and because the means of older children were in the same directions, interactions were not evident by conventional statistical significance levels. Consistent with this account, the means for both Selection 1 dimensions and for relevant-dimension pairings of the 4-year-olds in Experiment 3 are in the same direction as that of children in previous experiments despite not reaching statistical significance.

Finally, in all experiments, children were significantly more likely to succeed on trials in which the pivot item was located in the centre position than when it was located in the left or right position, and this was especially true for younger children in Experiment 1 (but not for younger children in Experiment 2). Perhaps it was easier for children to notice similarities between two items when these matching items were adjacent to each other than when they were separated by another item. This explanation might account for children's improved performance on trials in which the pivot item appeared in the centre position, because only on such trials did the pivot item happen to be adjacent to both of the other items with which it needed to be matched. When the pivot item was located in either the left or right position, it necessarily was adjacent to only one of its matching items.

5.1.4. Labelling Effects

The primary motivation for the latter two experiments in this series was to attempt to determine the kinds of cognitive processes that might be critical for successful performance on this task, and how these processes might relate to each other. In particular, the experiments were designed to assess whether or not language contributes to the expression of flexible thought. As predicted, in all experiments, language development (as assessed with the PPVT-R, a standardized measure of receptive language development) related significantly and consistently with Selection 2 performance. Admittedly, however, despite the fact that this relation is at least consistent with the hypothesis that language development plays a role in the natural development of cognitive flexibility in preschoolers, the causal nature of this relation cannot be addressed with correlational techniques. Instead, to do so, labelling manipulations were also introduced in these experiments to examine whether language can be causally linked to the emergence of flexible thought in preschoolers. The findings with respect to these manipulations demonstrate clearly that language can contribute markedly to children's tendency to demonstrate flexible thinking, although the contexts in which these labelling manipulations actually assisted performance were somewhat constrained.

In other words, there appeared to be relatively well-defined limits in the extent to which labels were capable of inducing better Selection 2 performance on the FIST. First, as mentioned previously, labels only had a noticeable impact on the performance of 4-year-olds: 3- and 5-year-olds seemed unaffected on their Selection 2 performance by being asked to label items on Selection 1. The failure to find demonstrable effects of labelling on the

~~performance of 3- and 5-year-olds was attributed to the poor performance of the former group~~ at the outset, and to the already near-ceiling performance of the latter. Second, compared to the performance of children in the various no-label conditions, 4-year-olds exposed to irrelevant labels or asked about the irrelevant dimension did not show any observable improvement in their Selection 2 performance. It was not sufficient, then, for 4-year-olds simply to be exposed to labels for their performance to improve, they had to be exposed to particular kinds of labels (see Table 23).

Table 23

Comparison of Selection 2 Means and Standard Deviations of 4-year-olds in Experiments 2 and 3 as a Function of Experiment and Condition

Experiment and Condition	M	SD
Experiment 2		
Relevant Label (n = 12)	12.33	3.47
No Label (n = 12)	8.58	4.64
Irrelevant Label (n = 12)	9.00	3.91
Experiment 3		
Relevant Dimension (n = 18)	11.50	4.37
Relevant Cue (n = 18)	10.67	4.56
No Cue (n = 18)	6.67	2.99
Irrelevant Dimension (n = 18)	6.39	4.59
Irrelevant Cue (n = 18)	7.56	4.94

Note that the mean Selection 2 responses of 4-year-olds' in Experiment 1 are not included because the version of task presented in Experiment 1 was different from the version presented in Experiments 2 and 3.

Specifically, in Experiment 2, 4-year-olds who were asked to label the dimension by which they selected items on Selection 1 did significantly better on Selection 2 than children

~~who were not explicitly asked to label on Selection 1 (or asked to label the irrelevant~~ dimension). Moreover, 4-year-olds improved on Selection 2 whether they generated these relevant labels themselves (Experiment 2), or whether these relevant labels were provided by the experimenter (Experiment 3), and relevant labels on Selection 1 helped Selection 2 performance irrespective of whether the experimenter labelled the relevant dimension by its dimensional term (e.g., "colour"), or by its cue (Experiment 3). Furthermore, detailed analyses of the actual labels that children provided in Experiment 2 indicated that the children who were most accurate in identifying the relevant dimension on Selection 1 also did better on Selection 2 than those who were less accurate. Therefore, it did not seem to suffice for children to simply be in the relevant-label condition for them to do well. Rather, it appears that they actually needed to produce (or be exposed to) accurate relevant labels in order to benefit from participating in this type of condition.

Despite the fact that children in the various irrelevant-label conditions did not profit from participating in these conditions like children in the relevant-label conditions, children who tended to label the irrelevant dimension accurately on Selection 1 did better overall on Selection 2 (but not on Selection 1) than those who were incorrect more often. This finding suggests that children who have well-formed notions of how cues exist within specific dimensions are more flexible in their thinking than those who have poorly organized dimensional knowledge.

In a related vein, children who labelled spontaneously in the preliminary trials also outperformed those who did not on Selection 2 (and on Selection 1). Moreover, 5-year-olds were more likely to label spontaneously in these preliminary trials than either 3- and 4-year-olds who did not differ from each other (see Table 9 again). This finding is consistent with

~~the initial hypothesis that 5-year-olds do well on Selection 2 because they already~~ spontaneously label the relevant dimension when selecting items on Selection 1 (and presumably, on Selection 2). The additional finding that both 4- and 5-year-olds were somewhat more likely to label the relevant and irrelevant dimensions accurately than 3-year-olds also supports the hypothesis that despite not doing so spontaneously, 4-year-olds could be induced to produce accurate labels (refer to Tables 10 and 13 again).

5.2. Existing Explanations of Labelling-Related Effects

On the basis of the specific pattern of success and failure that 4-year-olds exhibited in the different labelling conditions in this series of experiments, some possible explanations of labelling effects can be easily rejected as explanations for 4-year-olds' Selection 2 improvements on the FIST. First, several authors have argued that labelling effects on different tasks occur because they serve only to help direct children's attention toward important information about stimuli, helping them both to notice relevant information and to disregard irrelevant information (cf. Gibson, 1969; House, 1989). Perhaps a simple selective-attention explanation of labelling effects might account for findings on other tasks (e.g., Roberts & Jacob, 1991), but this kind of explanation cannot adequately account for the labelling effects seen on the FIST, in part because labelling effects were not measured on same selections on which labels were provided or elicited. In other words, if only attention-directing properties of labels were operating on the FIST, then one might expect that they would function only in directing children's attention to dimensions that were relevant on given selections, and not by having their effects extend to performance on other selections, requiring attention to different dimensions. Improvements on Selection 2 were measured as a

~~function of labelling manipulations on Selection 1, not as a function of labelling~~ manipulations on Selection 2.[26]

In a related vein, Luria (1961; see also Tikhomirov, 1978) argued that in some instances, labels might only indirectly help performance on certain tasks such as Go-no-Go tasks by affecting children's response rate. Specifically, labels may exert their influence by simply forcing children to slow down their response rate, thereby reducing their overall tendency to err on tasks on which they may be apt to respond too quickly. However, on this account, any labels should be equally effective in slowing down responses. Obviously, the fact that the specific labels that children were asked to produce or that the experimenter provided on the FIST were differentially effective in helping 4-year-olds' performance argues strongly against this kind of nonspecific explanation of labelling effects.

A different group of explanations for labelling effects focus on the linguistic code that is actually generated as a result of labelling, rather than focussing on nonspecific or indirect effects of labelling. These kinds of verbal-mediation accounts share in common the assumptions (a) that to mediate verbally means to represent information abstractly using a linguistic (or verbal) representation, and (b) that the result of using this kind of linguistic representation to control behaviour (instead of, or in addition to, using a perceptually derived representation) can lead to qualitative changes in behaviour, in general, and in performance on select cognitive tasks, in particular (e.g., Cantor & Spiker, 1976; Dusek, 1978; Furth & Milgram, 1973; Kendler, 1963; Premack, 1984; Vygotsky, 1934/1986, 1978; Werner, 1948;

[26] In Experiment 2, children were asked to label on Selection 2 as well but they were asked to do so only after they had selected items on that selection, and furthermore, their Selection 2 performance was not assessed in terms of these particular labels.

Whorf, 1956). Generally, proponents of verbal-mediation accounts agree that experimentally invoked or spontaneously generated labels should in some way permit or facilitate the generation of linguistic representations. However, specific verbal-mediation accounts differ widely in how they explain the precise nature of the link between linguistically derived representations and behaviour. Yet, despite large differences in interpretations, specific verbal-mediation accounts can be differentiated into three broad types that can be identified as multiple-codes, linguistic-code, and linguistic-and-informative-code explanations, specifically.

First, labels may affect behaviour because they either permit or facilitate the generation of a second representation (which only happens to be linguistic) in addition to the perceptually derived representation, thereby allowing information to be held in multiple ways simultaneously. The net effect of holding multiple representations of information simultaneously in itself has consequences for behaviour. Proponents of this kind of explanation do not believe that behaviour is affected by labels because labels allow information to be coded into one particular kind of representation or another. Instead, labels affect behaviour simply because they allow information to be coded <u>twice</u> (e.g., Cantor & Spiker, 1976; Karmiloff-Smith, 1984; Kobayashi & Cantor, 1974; Kunen & Duncan, 1983; Paivio, 1969). That is, when children are provided with labels, they not only have this linguistic code at their disposal, but they also have the original perceptual code. Perhaps having two representations of specific information influences cognition and behaviour differently than having only one. There are at least two reasons why having two separate representations might give children the opportunity to process information more effectively: Perhaps having information coded into two separate representations increases the odds that

~~children will use that information simply because they have more chances of retaining it in at~~ least one of the codes, or perhaps having two codes per se permits children to process and use information more effectively (cf. Paivio, 1969). That is, on the one hand, the simple fact that the information is represented twice might make children (or adults) more likely to retain it, allowing them to process and use it to govern their behaviour. On the other hand, two representations considered in combination might make children better able to use the information more effectively (e.g., Cantor & Spiker, 1976; Kobayaski & Cantor, 1974). For example, Cantor and Spiker argued that as a result of using spontaneous or elicited relevant labels on the discrimination-shift learning task, children effectively have access to two identical, yet separately coded relevant dimensions, thereby strengthening their tendency to act on the basis of that dimension.

Within this framework, then, there is nothing particular about the linguistic representation itself that is useful for behaviour except for the simple fact that labels permit one kind of representation to exist alongside another kind representation. However, this kind of explanation is unlikely to account for why certain labels affected performance on the FIST because according to this account, improvements should be comparable across all label conditions for the same reasons as in the case of a slowing-down-of-response explanation. That is, all kinds of labels should be equally effective in generating a second linguistic code. However, the fact that only relevant labels affected performance on the FIST argues against a simple multiple-code account.

Alternatively, a second class of verbal-mediation accounts emphasize the specific features of the representational format that is generated by labels. That is, rather than stipulating a need for multiple codes, proponents of this kind of verbal-mediation explanation

~~agree that there is something inherent about the verbal code itself that leads to behavioural~~ changes. In fact, most well-known verbal-mediation accounts fall within this broad category (e.g., Kendler, 1963; Premack, 1984; Vygotsky, 1934/1986, 1978; Whorf, 1956). On these accounts, labels have an effect on performance because they facilitate or permit information to be coded in an abstract linguistic format, and conceptualizing information in this kind of representational format, in turn, permits the use of different cognitive processes, which can lead to different behavioural responses. However, as in the case of multiple-codes explanations, if all that labels accomplish is to induce children to represent information in an abstract linguistic code, then all labels should be equally effective in improving performance, but labels were not all equally effective in improving Selection 2 performance on the FIST. Hence, this kind of linguistic-code verbal-mediation account, on its own, does not adequately explain performance on the FIST. Either representing information in an abstract linguistic code does not relate to performance on the FIST, or it is a necessary but not a sufficient condition for influencing performance.

According to proponents of a third type of verbal-mediation account, it is not only important that information be represented in an abstract linguistic code for it to influence cognition and behaviour, but the specific information that gets represented also needs to be considered (e.g., Dusek, 1978; Furth & Milgram, 1973; Kendler & Kendler, 1961; Luria, 1959; Wheeler & Dusek, 1973).[27] In other words, certain labels may help performance

[27] The fact that I grouped popular verbal-mediation accounts, such as Kendler's (1963), Vygotsky's, (1934/1986, 1978); Whorf's (1956) into the previous category (i.e., as linguistic-code explanations) is not intended to suggest in any way that proponents of these accounts necessarily considered that the specific information that is represented is unimportant, only that they tended to emphasize the general effects of representing information linguistically.

~~because they convey specific information in a linguistically coded representation~~, whereas other labels do not help because they convey irrelevant or useless information. For example, on the basis of their data on labelling effects on memory and clustering, Furth and Milgram (1973) argued that labelling not only has an effect on directing the children's attention to specific information, but it also allows them to discover and use surreptitiously introduced categorical structure of material. Similarly, Dusek (1978) argued that specific labels not only help direct attention to relevant information (while also directing it away from irrelevant information), but labels also have important encoding functions.

On the basis of the theoretical framework that I presented in the Introduction on the role of abstraction in cognitive flexibility and on the basis of the findings in the current experiments, I propose a verbal-mediation explanation that takes into consideration the actual information that is represented (i.e., a linguistic-and-informative-code explanation) to account for both the developmental changes observed on the FIST, on the one hand, and the labelling-related effects observed in certain conditions, on the other. In the next section, I will outline this account.

5.3. Explanation of Current Findings

So far, little has been said about the notion of abstraction except that (a) it was defined as a cognitive process that isolates and extracts information from the environment, and (b) it was a process that was essential for children to do well on Selection 1. However, in the Introduction, I also made the apparently contradictory claims that Selection 1 responses measured abstraction, but that labels would affect Selection 2 performance, the cognitive-flexibility measure, because language carries information abstractly. At this point, it becomes

~~necessary to elaborate on the notion of abstraction and, using Werner's (1948) distinction, it~~ appears to be necessary to differentiate between at least two qualitatively different levels of abstraction: One level at which information is simply detected and extracted from the stimulus information–a kind of selective-attention process–and a second level at which stimulus information is not only detected, but it is actually identified by means of an arbitrary symbolic tag that contextualizes it within a system of concepts. Werner referred to these two levels of abstract as primitive and categorial abstraction, each of which can be distinguished by key characteristics that either limit (in the former case) or permit (in the latter case) the expression of specific kinds of cognitive processes and cognitions (see Table 24 for a summary of these characteristics).

Werner (1948) defined primitive abstractions as instances in which information is extracted from a perceptual array (hence, the justification for the use of the term "abstraction"), but these kinds of abstractions remain closely bound to the perceptual information from which they are generated. In a paradoxical sense, then, primitive abstractions are concrete kinds of abstractions. By continuing to be bound to perceptual information, a primitive abstraction "brings forth qualities which do not stand out in isolation, but suffuse and dominate the totality" (Werner, 1948, p. 237). Therefore, despite the fact that it is possible to respond on the basis of one specific aspect of an object and also to group objects according to this same aspect, there is little cognitive control or choice over the particular aspect that gets isolated, and objects are grouped on the basis of only one

Table 24

Key Characteristics of the Two Hypothesized Levels of Abstraction

Characteristic	Level of Abstraction Primitive Abstraction *("Detection")*	 Categorial Abstraction *("Identification")*
Cognitive Processes	Likely to be a kind of selective-attention process.	Likely to be a linguistically mediated representational process.
Relation to Other Information	Abstracted information is bound to information available in the immediate perceptual environment.	Abstracted information is considered in relation to information contained within existing conceptual knowledge structures.
Control of Information	There is no deliberate choice in the information to which the individual attends.	Information to be abstracted is chosen deliberately by the individual.
Associated Matching Abilities	Match items on overall similarity or on one salient, but undifferentiated, quality (i.e., whatever dimensional value captures attention or dominates the whole; one-track abstractions).	Match items in different ways on different dimensions (i.e., flexible matching).
Expected Performance on FIST	Success on Selection 1 only.	Success on both Selection 1 and 2.

quality (whichever quality dominates the whole).[28] Werner called these one-track abstractions. Not only is there no control over the particular grouping that is formed, but Werner also argues that individuals who form groupings in this way cannot identify precisely the dimension by which they do so. In other words, this rudimentary form of abstraction acts like a selective-attention process. Remarkably, Werner even posited that primitive abstractions predominate up to the age of four years.

Like primitive abstractions, in categorial abstractions, information is also extracted from the perceptual array. However, a categorial abstraction is not bound by the perceptual information from which it is constructed. Instead, it is considered in relation to information organized within existing conceptual knowledge structures. That is, Werner (1948) contends that in categorial abstraction, "the quality (e.g., a color) common to all the elements involved is deliberately detached–mentally isolated, as it were–and the elements themselves appear only as visible exemplifications of the common quality" (p. 243). Because abstraction is determined by the deliberate selection of dimensions (not their perceptually available exemplars), a direct result of forming categorial abstractions is that, unlike "one-track" primitive abstractions, there is choice in the kind of information that is focussed upon.

[28] Alternatively, it is possible that in primitive abstractions, individuals do not match similar items on the basis of one specific dominating aspect of a stimulus, but instead, they respond on the basis of overall similarity calculated as a sum total across all aspects (cf. Evans & Smith, 1988; Smith, 1989a, 1989b, 1993). However, it may be difficult to distinguish between these two possibilities on some tasks including the FIST because they are likely to lead to similar response patterns. For example, if shown a little yellow teapot, a little blue teapot, and a medium blue teapot on the FIST (see Figure 6), children who succeed only on Selection 1 might be able to select the two little ones either (a) because size dominates or (b) because these two items are more similar to each other than the little yellow teapot and medium blue teapot due to the fact that the two little ones are identical on two of the three dimensions, whereas the items in the other pair are identical on only one of the three dimension (i.e., shape).

Presumably, then, individuals who are able to select deliberately the dimension on which they wish to attend should also be able to switch flexibly between attending to one or another dimension.

In short. Werner's notion of categorial abstraction, then, is similar to the general definition of abstraction presented earlier in the Introduction in that in categorial abstraction, information is both extracted from the environment and considered within a system of concepts. Of particular interest to the approach adopted in the current paper, Werner also argued that the transition between primitive and categorial occurred as a result of the advent of language. Specifically, he proposed that, "It is by means of representation through language and through the naming process that the human mentality reaches the level of abstract concept" (p. 254).

With this basic distinction between primitive and categorial abstraction in mind, it becomes easier to reconcile the findings of the current experiments; particular, those pertaining to the discrepancy between performance on Selection 1 and performance on Selection 2. More specifically, by distinguishing between these two levels of abstraction, it is possible to attribute one kind of abstraction ability for succeeding on Selection 1, and another kind that permits success on both kinds of selections. The ability to form primitive abstractions can account for 4-year-olds' successful performance on Selection 1 and their concurrent poor performance on Selection 2. That is, 4-year-olds may have been able to succeed on Selection 1 of the FIST (but not on Selection 2) because they spontaneously formed primitive abstractions. On this account, they detected similarities between items on the basis of dominating qualities, and as a result, they were only able to match items on a specific trial on the basis of the quality that happened to dominate on that particular trial. In other words, they were unable to select items on the basis of other shared similarities on Selection 2, because by forming primitive abstractions, they were limited to only one-track abstractions over which they had no control. In contrast, 5-year-olds may have succeeded on both selections because they approached the task by spontaneously forming and using categorial abstractions.

If this distinction between primitive and categorial abstraction holds for performance on the FIST, two implications for labelling-related effects should follow. First, labelling should not affect the tendency to form primitive abstractions, and as a result, labelling should not affect Selection 1 performance on the FIST. It was not possible to test this hypothesis directly given the way in which labelling manipulations were introduced on the FIST in the current experiments. That is, labels on Selection 1 followed instead of preceding each selection, and therefore, given that the labels themselves actually depended upon Selection 1 responses, it was difficult to assess their own influence on these selections.[29] However, one labelling-related finding in the current experiments actually appears to contradict this first implication. Children who spontaneously produced at least one relevant label in the preliminary trials not only did better on Selection 2 than children who did not produce any, as

[29] There is perhaps one indirect way of assessing whether labels affected performance on Selection 1 in the current experiments. That is, it might have been possible for labels presented on earlier trials to influence how children approached the task on later trials (i.e., intertrial labelling effects as opposed to intratrial effects), and therefore, have had an impact on their Selection 1 performance on later trials. Yet, no such increase in performance was observed across trial blocks as a function of labelling condition. In reality, all 3-year-olds (children who presumably had difficulty with primitive abstractions) tended to get worse across trial blocks rather than improving. Admittedly, however, although the lack of a notable improvement on Selection 1 over trials for children in the relevant-label condition is consistent with the claim that labelling does not affect primitive abstraction, it is far from compelling and needs to be assessed more directly before any real conclusions can be reached.

anticipated, but they also performed markedly better on Selection 1. This finding suggests that spontaneous identification of relevant dimensions may also be important for Selection 1 performance. However, the increase in Selection 1 performance demonstrated by spontaneous labellers may have less to do with specific effects of labelling and more to do with the fact that the nonlabeller group included both "nonabstractors" (children who did not do well on Selection 1) and primitive abstractors, whereas the spontaneous labellers likely included only categorial abstractors. As a group, most 3-year-olds did more poorly on Selection 1 (and Selection 2) than the (4- and) 5-year-olds. Moreover, as a group, most of the 3- (and 4)-year-olds did not spontaneously produce labels, whereas most of the 5-year-olds did. Hence, 3-year-olds made up a significant proportion of the children in the nonlabeller category, and as a result, they likely lowered the overall means of nonlabellers on Selection 1 (and on Selection 2). Hence, the finding that children who spontaneously identified the relevant dimension did better on Selection 1 (and perhaps on Selection 2 as well) than those who did not label spontaneously likely resulted from the fact that the latter group included children who failed to abstract information altogether (predominantly 3-year-olds).

The second implication of the distinction between primitive and categorial abstraction is that forming primitive abstractions is a necessary but not sufficient condition toward forming categorial abstractions, but that it should be relatively easy to induce individuals who spontaneously act on the basis of primitive abstractions to form categorial abstractions when they are provided with appropriate arbitrary tags (i.e., relevant labels). The findings in Experiment 2 that 4-year-olds were no more likely to produce relevant labels spontaneously in the preliminary trials than 3-year-olds (see Table 9), but that they were as likely as the 5-year-olds to identify consistently the irrelevant dimension on Selection 1 (and that those who did so also did better only on Selection 2; see Table 13 together support the idea that despite not spontaneously using categorial abstractions, 4-year-olds did have the necessary dimensional structure in place for situating specific cues within their respective dimensions.[30] It is no surprise, then, that when they were given relevant labels they were able to benefit. What remains to be explained, however, is why they failed to draw spontaneously upon this structure to select items on the FIST if it was already available to them.

As alluded to earlier, a slightly different way of conceptualizing the different processes involved in primitive and categorial abstractions is that in primitive abstraction, individuals need only detect qualities shared by different items, whereas in categorial abstraction, individuals must identify the qualities in question (cf. Karmiloff-Smith's, 1992, notion of explication–representational redescription of implicit knowledge into explicit knowledge). So a crucial difference between primitive and categorial abstractions is that the former is concerned with detection processes, whereas the latter is concerned with identification processes, and each of these processes presumably has different implications for cognition and behaviour (cf. Gibson, 1969). For example, what does it mean to identify an object (or an attribute)? First, it is to denote essential properties of the object in such a way as to be able to consider it separately from its instantiation. One convenient and economical way of accomplishing this is to assign arbitrary tags to objects so that objects that are alike are identified by the same tag, and objects that are different are identified with different tags (cf.

[30] Recall that identifying the irrelevant dimension correctly required that children respond to questions referring to dimensions (e.g. "What colour are these pictures?") with cue terms within that same dimension, although the precise cue term that they used did not matter. Thus, the only requirements for doing well on these questions were that children know each dimensional term, and that they be able to locate specific cues correctly within these dimensions. In other words, they needed to have a well-organized dimensional structure.

Clark, 1987). A consequence of assigning arbitrary tags to identify objects is that tags create distance between individuals and their environment because tags necessarily transcend particular objects in that they can refer not only to individual objects, but also to entire classes of objects that share certain properties in common. As a result, individuals no longer need to reason on the basis of immediately available stimuli (or even on the basis of previously experienced stimuli); they can also reason on the basis of imagined or transformed representations of reality. On this account, then, a process of identification using arbitrary tags is a necessary condition for the emergence of flexibility in both thought and action.

5.4. Limitations of the Current Explanation and Future Directions

The ideas that development entails in part an increase in differentiation and articulation of knowledge structures, on the one hand, and a progressive hierarchicalization of these same structures, on the other, are complementary ideas that permeate several theories of cognitive and perceptual development (e.g., Gibson, 1969; Inhelder & Piaget, 1964; Smith, 1989a; Vygotsky, 1934/1986; Werner, 1948). The findings of the current experiments are pertinent to the latter of these processes. That is, on the basis of Werner's (1948) notion of categorial abstraction, I proposed in the previous section that labels play an important role in accessing information that is hierarchically organized within dimensional structures, or at least withing broader systems of concepts. For example, labels are seen as permitting specific cues of dimensions (e.g., red or blue) to be recognized as exemplars of specific dimensions (e.g., colour), rather than being detected perceptually only as (poorly differentiated) qualities of objects (cf. Karmiloff-Smith, 1984; Werner, 1948). However, despite the fact that it is said to play a vital role in the development of higher cognitive processes in several theories (e.g.,

~~Inhelder & Piaget, 1964, on the development of class inclusion; Smith, 1989a, on the~~ development of dimensional understanding; Zelazo & Frye, 1997, on the development of embedded rules), there is very little direct empirical support for the claim that organizing information within a hierarchical structure is a necessary prerequisite for the emergence of these higher forms of cognitive processes and for the behaviours that are made possible by such processes. That is, there is ample evidence that labels can help children approach cognitive tasks conceptually rather than perceptually (e.g., Bleichfeld, Moely, Rabinowitz, & Turgeon, 1977; Furth & Milgram, 1973; Gentner & Namy, 2000; Morgan & Greene, 1994; Sugimura, 1978; Waxman & Gelman, 1986). Yet, there is no evidence to my knowledge that undisputably demonstrates that the hierarchical nature of the information is critical, and that if this information were organized differently, the same processes or behaviours would not be possible. For example, Zelazo and Frye (1997) argue that at around age of 4 years, children develop the ability to use embedded or hierarchically arranged if-then-if-then rule structures (e.g., "If we are playing the colour game, then if the card is red, then do. . ."). However, as these authors themselves acknowledge (Frye, Zelazo, Brooks, & Samuels, 1996), it is possible that children acquire the ability to use conjoint if-and-if-then rule structures (e.g., "If we are playing the colour game and if the card is red, then do. . .") instead of hierarchically arranged ones. Unfortunately, experiments have yet to be done in which the organizational structure of conceptual information is manipulated experimentally and performance is assessed accordingly.

Likewise, the findings of the current experiments provide little in terms of direct empirical support for the necessity of hierarchicalization. Moreover, the finding in Experiment 3 that there was no difference in the performance between children who were

~~exposed to relevant cue terms and those who were exposed to relevant dimensional terms~~ could be construed either as evidence for the necessity of hierarchically arranged dimensional structures, or as evidence against it. For instance, any label that invokes the concept of a relevant cue (e.g., "blueness") may be sufficient for influencing performance without there being a need to consider it as a cue subordinated within a higher dimension (i.e., "blue as a colour" per se). Therefore, the finding could be interpreted to mean that considering the dimension itself is unnecessary (i.e., that categorial abstractions are unnecessary). Or alternatively, it is possible that, at least by this age, language has developed sufficiently that by labelling a cue, the dimension is necessarily invoked (cf. Gentner & Loewenstein, in press; Waxman & Gelman, 1986). On this account, using the cue or the dimensional term would not matter because both are sufficiently linked that referring to one automatically conjures up the other. The finding in Experiment 2 that children who did well on Selection 2 tended to respond correctly in the irrelevant-label condition to questions about dimensions and their cues provides some indirect support for this latter interpretation (see previous discussions in Sections 3.4. and 5.3.).

Given the finding in Experiment 2 that 4-year-olds in the irrelevant-label condition tended to provide correct cue terms when asked about the irrelevant dimension on each selection, it appears that even though 4-year-olds may have the requisite knowledge for responding on the basis of categorial abstractions, they fail to use this knowledge spontaneously, and instead they respond on the basis of primitive abstractions. As noted in the previous section, it unclear what conditions might prompt children to use this categorial information spontaneously by the time they are 5 years. Perhaps a minimum amount of experience with the medium of language or a minimum number of relevant acquired words are necessary for objects presented in the environment to be identified spontaneously with linguistic symbols (cf. Waxman & Gelman, 1986). It might be possible to obtain clues as to the underlying conditions that prompt this developmental change by studying how children acquire dimensional information in the first place. Linda Smith and her colleagues have been attempting to do just that using longitudinal observational studies of parent-child interactions (Smith & Sandhofer, 2001), experimental studies with adults in which adults are trained in different ways to learn nonsense dimensions (Sandhofer & Smith, in press), and neural network models (Gasser & Smith, 1998; Smith, 1993; Smith, Gasser, & Sandhofer, 1997).

In the previous section, I also argued for the importance of the arbitrary nature of symbols (i.e., labels) in the development of cognitive flexibility. That is, because of their arbitrary nature, I posited that labels provide a way of reasoning using information other than that which is immediately available or has been previously experienced. However, whether arbitrariness is a necessary aspect of labels for these to be effective in improving performance on Selection 2 has not been demonstrated empirically. Perhaps nonarbitrary symbols that convey relevant information (e.g., some sort of iconic cue) would be as effective (but see, DeLoache, 2000, for limits in the extent to which concrete objects can be useful as symbols). The findings of the current experiments are silent in this respect, and therefore, it remains to be determined experimentally whether or not arbitrariness is a necessary property of labels for these to be effective in inducing better performance on Selection 2.

One further problem in attributing changes in performance on the FIST to developments in abstraction abilities is that this explanation focusses only on the change that occurs between 4 and 5 years, and it says nothing about potential developments that may occur earlier that might explain the changes in performance found between 2 and 3 years in

Experiment 1, and those found between 3 and 4 years (in Experiments 1 and 2). Admittedly, the purpose of the present experiments was to attempt to determine the role of labelling in cognitive flexibility, and therefore the 4- to 5-year-old change is the more relevant one. Nonetheless, earlier changes will have to be explained eventually, if not only in the hope of better explaining the later changes themselves. Is it that children who failed to select items consistently on Selection 1 necessarily failed to form primitive abstractions at all, or is it something particular about the items on the FIST that made them appear as though they were incapable of abstracting similarities between items? The 3-year-olds in Experiment 1 were able to select matching items in the criterial trials one year before they could do so consistently on the FIST. One obvious difference between items presented in the criterial trials and those presented in FIST is that pairs of matching items used in the criterial trials were identical, whereas pairs of matching items used in the FIST were <u>nonidentical</u>. Smith (1984) showed that preschoolers select matching items on the basis of absolute identity prior to selecting matching items on the basis of part identities (see Section 5.5.2., for a more detailed description). Jacques (1995) also documented age-related differences in performance on comparable versions of a sorting task that were identical to each other in every respect except for their use of identical and nonidentical items (see Section 5.5.2., for a more detailed description). Another aspect of the items in the FIST was that two possible matches could be made on any trial. As a result children needed to decide on Selection 1 between two potential, yet conflicting, matches. It is possible that the conflict inherent in the stimuli rather than the nonidentical nature of the stimuli can explain 3-year-olds' difficulty on Selection 1. Perhaps if presented with a version of a FIST trial in which it was possible to make only one match between nonidentical items (e.g., a little yellow teapot, a medium <u>blue</u> teapot, a large <u>blue</u>

~~teapot), then 3-year-olds might be able to select matching pairs of nonidentical items~~ correctly. The nature of the underlying weakness responsible for the difficulty of 3-year-olds' difficulty on the FIST remains to be determined.[31]

Existing research, including the results of the experiments presented in this paper, can leave no doubt that language–and labelling, in particular–can affect performance on cognitive tasks, at least in experimental settings. However, it is one thing to show that language manipulations introduced within an experimental setting can affect certain cognitive processes, but quite another to prove that language is also the element responsible for the development of these same processes in children's natural environments. Obviously no experiment or set of experimental results can definitely identify the underlying causes for the emergence of certain processes in the course of ontogenetic development. However, longitudinal observational designs used in conjunction with experimental designs might be particularly useful for understanding the natural relation that exists between thought and language throughout the early course of human development. Moreover, the entire range of cognitive processes that can be affected by language manipulation (positively or negatively) also remains to be defined and consequently, explained. As others have argued, to understand the exact role that language plays in human cognition, it is as important to know when language helps, when it does not, and when it may actually hinder performance (e.g., Brandimonte & Gerbino, 1993). Perhaps by using a variety of theoretical and empirical approaches to investigating the nature of the relation between language and thought and by using these approaches in such a way as to permit the systematic elimination of specific

[31] The difficulty of the 2-year-olds is perhaps best attributed to difficulty with the task instructions given their difficulty on the criterial trials.

explanations, it might eventually be possible to make educated guesses about the exact nature of the relation.

5.5. Relevance to Other Work

Aside from their relevance to the somewhat older literature on language and thought (e.g., Bruner, 1973; Kendler, 1963; Vygotsky, 1934/1986; 1978; Werner, 1948; Whorf, 1956), the current findings also dovetail well with recent lines of research (e.g., Deák & Bauer, 1995; Davidson & Gelman, 1990; Gentner & Namy, 2000; Smith, 1984; Smith, 1989a; Smith & Sandhofer, 2001; Waxman & Gelman, 1986; Zelazo, Frye, & Rapus, 1996). As an example, the relation between the current research and work by Zelazo and Frye and their colleagues on the development of embedded rule use will be highlighted briefly to illustrate some of the broader implications of the current work. In addition, Piaget's view of the role of language on the development of class inclusion will also be considered in this section, in part because of the obvious links between the tasks used to assess class inclusion and the FIST, and in part because of the impact that Piaget's views have had on the field.

5.5.1. Rule Use in Preschoolers

The results from the FIST converge well with those from another task that superficially resembles the FIST, the Dimensional Change Card Sort (DCCS). In the DCCS (e.g., Frye et al., 1995; Zelazo et al., 1996; see Zelazo & Frye, 1997; Zelazo & Jacques, 1996, for reviews), children are presented with two target cards (e.g., a red car and a blue flower) that remain visible throughout the task and children are asked to sort test cards (e.g., blue cars and red flowers) that match one of the target cards on one dimension (e.g., shape) and match the other target card on the other dimension (e.g., colour). In a preswitch phase, children are

~~first told explicitly to sort test cards by one dimension. After a number of trials, children are~~ then asked to switch and sort test cards by the other dimension in a postswitch phase. Despite being told the relevant rules on each trial (e.g., "Red ones go here and blue ones go there."), the majority of 2-year-olds fail to use even the preswitch rules correctly except in a redundant version of the task in which the test cards are identical to the target cards (Jacques, 1995). In contrast, 3-year-olds tend to sort correctly on the preswitch phase of the standard version, but sort perseveratively in the postswitch phase by continuing to use the preswitch rules. Finally, the majority of 4- and 5-year-olds are able to switch and sort the test cards according to the postswitch rules (e.g., Frye et al., 1995; Zelazo et al., 1996). Furthermore, findings from several studies suggest that poor performance on this task results from limitations in flexible thinking, and not from difficulty with inhibitory response control (e.g., Jacques, Zelazo, Kirkham, & Semcesen, 1999; Zelazo et al., 1996).

Table 25 summarizes the achievements of children at each age on the DCCS and on the FIST. Of particular interest, despite the fact that the two tasks superficially resemble each other, the table indicates that there appears to be about a one-year lag on the FIST in preschoolers' ability to succeed on aspects of that task that seem analogous to those on the DCCS. That is, 2-year-olds are able to sort cards correctly on the preswitch phase of a redundant version of the DCCS (in which the target and test cards are identical) one year earlier than children are able to select pairs of identical cards consistently on the criterial trials. In addition, 3-year-olds are able to sort nonidentical cards on the basis of one dimension on the DCCS one year earlier than children are able to perform well on Selection 1 of the FIST, which also requires that they select pairs of nonidentical matching items correctly on the basis of one dimension. Lastly, 4-year-olds are able to sort flexibly according

~~Table 25~~

Summary of Performance of Different Age Groups on Two Measures of Cognitive Flexibility in Preschoolers (Dimensional Change Card Sort and Flexible Item Selection Task) by Type of Task (i.e., Deductive vs. Inductive).

	Kind of Task	
Age Group	Deductive Task (Dimensional Change Card Sort)	Inductive Task (Flexible Item Selection Task)
2-year-olds	Pass Preswitch of Redundant	Fail All
3-year-olds	Pass Preswitch of Standard	Pass Criterial Trials
4-year-olds	Pass Postswitch of Standard	Pass Selection 1
5-year-olds	Pass Postswitch of Standard	Pass Selection 2

Note. Passing a given part of a task meant that children also passed the parts of the particular task that their younger counterparts passed, but that they failed the parts of the same task that the older children passed.

to two dimensions on the DCCS one year before children are capable of performing well on both selections of the FIST, which also requires that children select matching items flexibly on the basis of two dimensions.

In addition to differences in the procedures for administering the tasks, there are several important difference between these two superficially similar tasks. One important difference between the FIST and the DCCS is that the FIST is an inductive as opposed to a deductive task. That is, on the DCCS children are told by which dimension to sort on every trial, whereas on the FIST children are required to detect relevant dimensions themselves. Thus, although both tasks require that children use dimensional information and that they be able to switch flexibly between different dimensions, only the FIST requires that children abstract relevant dimensional information for themselves. Perhaps the one-year décalage

~~between performance on the FIST and on the DCCS reflects the added requirement on the~~ FIST that children both abstract relevant dimensions for themselves (as opposed to being told which dimensions to use) and use that information flexibly. In addition, however, because the DCCS is a deductive task, the tasks also differ in that the DCCS items (and the rules) are always labelled on the basis of whichever dimension is relevant for sorting on a particular trial (e.g., "Red ones go here, blue ones go there. Where does this red one go?"). Given the improvements of 4-year-olds in the relevant-label conditions in the current series of experiments, it is quite possible that the one-year décalage between performance on these two tasks is due to the widespread and consistent use of relevant labels on the DCCS, but not (always) on the FIST. Future work should attempt to match more closely the procedures for administering both tasks in order to determine whether it is the deductive or inductive natures of the tasks, the presence or absence of relevant labels, or both that leads to successful performance on the DCCS at an earlier age than on the FIST.

5.5.2. Piaget's View of the Role of Language in the Development of Class Inclusion

As mentioned previously, several major theories of development postulate a critical role in cognitive development to the hierarchicalization of knowledge structure (e.g., Gibson, 1969; Vygotsky, 1934/1986; Werner, 1948), and Piaget's theory on the development of class inclusion (Inhelder & Piaget, 1964) is no exception. However, whereas Vygotsky and Werner attributed the ability to organize information hierarchically to language development, Piaget (e.g., Inhelder & Piaget, 1964; Piaget, 1964/1967) held that language only plays a minor role in the process. Piaget's argument is premised on the idea that language cannot possibly account for the development of class inclusion because it is not directly linked to the emergence of preschoolers' initial abilities to classify objects into graphic and then, into nongraphic collections, abilities that Inhelder and Piaget (1964) proposed were early precursors of the development of class inclusion. As a result, Piaget (Inhelder & Piaget, 1964; Piaget, 1964/1967) argued that because language is not present at the inception of this entire developmental process, its role is likely only one of facilitating this development, not permitting it.

Perhaps language may not be the first of the essential elements to develop that is believed to be necessary for the development of class inclusion (or for higher forms of intellectual processes in humans, more generally), but its role may nonetheless be as vital to the development of this higher form of thought. For example, no one would contest the fact that sugar (natural or otherwise) is an essential ingredient in making beer, despite the fact that it is the last ingredient to be added to the mixture. In fact, it not only facilitates the process, it actually permits it: Without sugar, beer would simply not be beer. Moreover, the role of sugar in the beer-making process should not be considered less important relative to that of the other ingredients simply because these other ingredients are incorporated earlier in the process. By the same token, despite the fact that language may exert its influence at a relatively late point in development, it may still play as essential a role in the development of specific mental processes as other cognitive precursors that happen to emerge earlier in development. Therefore, Piaget (1964/1967) may very well be correct in that language is not involved as early as other cognitive abilities in the development of class inclusion (e.g., the ability to form nongraphic collections), but language may nonetheless be essential for permitting its emergence. In short, the importance of language should not be questioned on the grounds that it only shows its influence relatively late in the process of cognitive development.

~~On a more general note, the ability to solve increasingly complex problems in~~ development likely depends on the precise orchestration of various cognitive processes. However, this does not mean that these processes cannot exist independently at earlier points development. In fact, Vygotsky (1929; 1934/1986) himself argued that language and cognition emerge from independent sources (i.e., the natural and cultural lines of development). According to him, early in development, thought can occur without language and language can exist without thought. However, at a certain point in development, these two independent processes merge, and their newly developed interdependence allows for new possibilities in human cognitive and linguistic processing, literally creating a revolutionary change in human thought. The current work focusses primarily on the point in development at which language and abstraction appear to become connected (i.e., the development of categorial abstraction), one consequence of which is proposed to be the emergence of a new degree of cognitive flexibility, which in turn permits qualitative changes in the control of thought and behaviour. The concern of the current work is not with whether language or abstraction appears first in development. In fact, from a Wernerian perspective, primitive abstractions are viewed as thoughts devoid of linguistic meaning (Werner, 1948).[32]

[32] By the same token, it may also be possible for language to be produced without abstraction necessarily taking place (cf. Smith & Sandhofer, 2001). This point has not been addressed in the current paper, although see Deacon (1997) who argues that early in development, children use labels as indexes for things rather than as symbols in the proper sense of the term (cf. Werner & Kaplan, 1963).

CHAPTER VI

CONCLUSIONS

In language, by the selection and representation of certain aspects of a thing through the medium of sound, an act of creation–the creation of a mental concept of a thing–is involved. In this respect the function of language is no different from that of any other creative activity, particularly artistic creative activity.

- Heinz Werner (1948, p. 257).

A series of three experiments were conducted to determine whether language contributes to the emergence of cognitive flexibility in preschoolers. First. the Flexible Item Selection Task (FIST)–a task adapted from the Visual-Verbal Test (Feldman & Drasgow. 1951)–was developed to assess abstraction and cognitive flexibility in preschoolers. The results of Experiment 1 revealed that (a) 2-year-olds failed to understand basic task instructions; (b) 3-year-olds performed poorly on Selection 1, suggesting that they had difficulties with abstracting a common dimension from nonidentical (and conflicting) items; and (c) 4-year-olds did well on Selection 1 but did significantly worse than 5-year-olds on Selection 2, suggesting that they had problems with cognitive flexibility. It was also hypothesized that the age-related changes in cognitive flexibility between 4 and 5 years might be due to underlying changes in language development, or more precisely, in children's ability to represent spontaneously information into a linguistic code. Experiment 2 was then conducted to test this claim by using labelling manipulations on the FIST. The results of Experiment 2 confirmed the general pattern of age changes found in Experiment 1, and they also revealed that not only does Selection 2 performance relate to receptive language development in general, but that it is also influenced by labelling on Selection 1. When 4-year-olds were asked to provide labels on Selection 1 that were relevant to their selections, their Selection 2 performance was significantly improved compared to 4-year-olds who were not asked to label or who were asked to label irrelevant aspects of the stimuli. Moreover, within both the relevant-label and irrelevant-label conditions, children who made fewer labelling errors on Selection 1 did better on Selection 2 than those who made more errors. Furthermore, across all conditions, children who spontaneously labelled the relevant dimension in the preliminary trials did better on both selections than those who did not.

Finally, Experiment 3 was then conducted for two additional reasons: to determine whether specific labels presented in a controlled manner on Selection 1 actually cause improvements on Selection 2, and to determine which kinds of labels help improve 4-year-olds' Selection 2 performance. In that experiment, the experimenter–rather than children–selected items and provided predetermined labels on Selection 1. Results showed

~~that labels that referred to the relevant dimension (i.e., the dimension on which items~~ matched each other; e.g., size) helped, but those that referred to the irrelevant dimension (i.e., the dimension that did not vary across the 3 items; e.g., shape) did not. This same pattern of results held regardless of whether the experimenter labelled the dimension itself (e.g., "same size") or the cue (e.g., "both big").

In short, this series of experiments attempts to specify whether language can influence the development of one aspect of human cognition, the ability to represent information flexibly. As Vygotsky (1934/1986, 1978) elegantly argued in his various writings, language provides an ideal vehicle for actualizing thought. The particular view exposed in this paper is that the emergence of flexible thought may be a corollary of the development in humans of a higher form of abstraction in which objects and their attributes are represented within a broader system of concepts with the use of arbitrary linguistic symbols. This process of identifying objects using arbitrary tags provides individuals with a representational device that explicitly separates the representation (the tag) from its referent (the object). Before such a separation is established, the representation and the referent fail to be sufficiently differentiated from each other to allow individuals a means of manipulating the representation independently. On this account, then, arbitrary symbols may act by eliminating constraints imposed by the representational medium on the number of representations that can be considered about a particular object or event.

The idea that language affects cognition is not new, nor are findings that labelling-manipulations can dramatically improve performance on cognitive tasks. In the past century, Vygotsky (1934/1986; 1978), Werner (1948), Whorf (1956), Kendler (1963), Bruner (1973), and several others have popularized the notion of verbal mediation, and by doing so, they

~~generated fierce theoretical debates and stimulated numerous experimental investigations. In~~ fact, to account for the findings of the current study, I adopted a verbal-mediation account that emphasizes different levels of abstraction, an account proposed by Werner (1948) more than half a century ago. Yet, despite the fact that it accounts for the current findings reasonably well, it is still speculative, and whether or not it will hold up under further investigation remains to be determined. However, perhaps the most important contribution of the current work is that it calls attention to the need for more research that not only describes links between language and cognition, but that also attempts to determine precisely how language might come to serve different aspects of the human intellect.

References

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. Developmental Psychology, 35, 1311-1320.

Berg, E. (1948). A simple objective test for measuring flexibility in thinking. Journal of General Psychology, 39, 15-22.

Bleichfield, B., Moely, B. E., Rabinowitz, F. M., & Turgeon, V. F. (1977). The use of categorical concepts by young children in the solution of a multiple-item discrimination learning task. Child Study Journal, 7, 29-47.

Brandimonte, M. A., & Gerbino, W. (1993). Mental image reversal and verbal recoding: When ducks become rabbits. Memory and Cognition, 21, 23-33.

Bruner, J. S. (1973). The course of cognitive growth. In J. S. Bruner, Beyond the information given (J. M. Anglin, Ed.; Ch. 19, pp. 312-323). New York: W. W. Norton.

Bruner, J. S., & Kenny, H. J. (1966). On multiple ordering. In J. S. Bruner, R. R. Olver, & P. M. Greenfield (Eds.), Studies in cognitive growth (Ch. 7, pp. 154-167). New York: John Wiley & Sons.

Cantor, J. H., & Spiker, C. C. (1976). The effects of labelling dimensional values on setting differences in shift performance of kindergarten children. Memory and Cognition, 4, 446-452.

Chelune, G. J., & Baer, R. A. (1986). Developmental norms for the Wisconsin Card Sorting Test. Journal of Clinical and Experimental Neuropsychology, 8, 219-228.

Chelune, G. J., & Thompson, L. L. (1987). Evaluation of the general sensitivity of the Wisconsin Card Sorting Test among younger and older children. Developmental

Neuropsychology, 3, 81-90.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), Mechanisms of language aquisition [sic] (pp. 1-33).

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Davidson, N. S., & Gelman, S. A. (1990). Inductions from novel categories: The role of language and conceptual structure. Cognitive Development, 5, 151-176.

Deacon, T. W. (1997). The symbolic species: The co-evolution of language and the brain. New York: W. W. Norton.

Deák, G. O., & Bauer, P. J. (1995). The effects of task comprehension on preschoolers' and adults' categorization choices. Journal of Experimental Child Psychology, 60, 393-427.

Delis, D. C., Squire, L. R., Bihrle, A., & Massman, P. (1992). Componential analysis of problem-solving ability: Performance of patients with frontal lobe damage and amnesic patients on a new sorting test. Neuropsychologia, 30, 683-697.

DeLoache, J. S. (2000). Dual representation and young children's use of scale models. Child Development, 71, 329-338.

Dempster, F. N. (1992). The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. Developmental Review, 12, 45-75.

Dempster, F. N. (1993). Resistance to interference: Developmental changes in a basic processing mechanism. In M. Howe & R. Pasnak (Eds.), Emerging themes in cognitive development (Vol. 1, pp. 3-27). New York: Springer-Verlag.

Dickerson, D. J. (1970). Effects of naming relevant and irrelevant stimuli on the discrimination learning of children. Child Development, 41, 639-650.

Drasgow, J., & Feldman, M. (1957). Conceptual processes in schizophrenia revealed by the Visual-Verbal Test. Perceptual and Motor Skills, 7, 251-264.

Duncker. K. (1945). On problem-solving. Psychological Monographs, 58, (5; pp. 113).

Dunn, L. M., & Dunn, L. M. (1981). PPVT: Peabody Picture Vocabulary Test-Revised: Manual for forms L and M. Circle Pines, MN: American Guidance Services.

Dusek, J. B. (1978). The effects of labeling and pointing on children's selective attention. Developmental Psychology, 14, 115-116.

Esposito, N. J. (1975). Review of discrimination shift learning in young children. Psychological Bulletin, 82, 432-455.

Evans, P. M., & Smith, L. B. (1988). The development of identity as a privileged relation in classification: When very similar is just not similar enough. Cognitive Development, 3, 265-284.

Feldman, M. J., & Drasgow, J. (1951). A Visual-Verbal Test for schizophrenia. Psychiatric Quarterly, 25(Suppl.), 55-64.

Flavell, J. H. (1970). Developmental studies of mediated memory. In H. W. Reese & L. P. Lipsitt (Eds.), Advances in child development and behavior (Vol.5) (pp. 181-211). New York: Academic Press.

Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance-reality distinction. Cognitive Psychology, 15, 95-120.

Frye, D., Zelazo, P. D., Brooks, P. J., & Samuels, M. C. (1996). Inference and action in early causal reasoning. Developmental Psychology, 32, 120-131.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. Cognitive Development, 10, 483-527.

Furth, H. G., & Milgram, N. A. (1973). Labeling and grouping effects in the recall of pictures by children. Child Development, 44, 511-518.

Gasser, M., & Smith, L. B. (1998). Learning nouns and adjectives: A connectionist account. Language and Cognitive Processes, 13, 269-306.

Gentner, D., & Loewenstein, J. (in press). Relational language and relational thought. In J. Byrnes & E. Amsel (Eds.), Language, literacy, and cognitive development. Hillsdale, NJ: Erlbaum.

Gentner, D., & Namy, L. L. (2000). Comparisons in the development of categories. Cognitive Development. Special Issue, 14, 487-513.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), Perspectives on language and thought: Interrelations in development (pp. 225-277). Cambridge: Cambridge University Press.

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 3½-7 years old on a Stroop-like day-night test. Cognition, 53, 129-153.

Gibson, E. J. (1969). Principles of perceptual learning and development. New York: Appleton-Century-Crofts.

Glucksberg, S., & Weisberg, R. W. (1966). Verbal behavior and problem solving: Some effects of labeling in a functional fixedness problem. Journal of Experimental Psychology, 71, 659-664.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational

~~change and its relation to the understanding of false belief and the appearance-reality~~ distinction. Child Development, 59, 26-37.

Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. Journal of Experimental Psychology, 38, 404-411.

Harnishfeger, K. K., & Bjorklund, D. F. (1993). The ontogeny of inhibition mechanisms: A renewed approach to cognitive development. In M. Howe & R. Pasnak (Eds.), Emerging themes in cognitive development (Vol. 1, pp. 28-49). New York: Springer-Verlag.

House, B. J. (1989). Some current issues in children's selective attention. In H. W. Reese (Ed.), Advances in child development and behavior, Vol. 21. (pp. 91-119). San Diego, CA: Academic Press.

Inhelder, B., & Piaget, J. (1964). The early growth of logic in the child: Classification and seriation (E. A. Lunzer & D. Pepert, Trans.). New York: Harper & Row, Publishers. (Original work published in 1959)

Jacques, S. (1995). The development of rule use during the preschool period. Unpublished master's thesis, University of Toronto, Ontario, Canada.

Jacques, S., Zelazo, P. D., Kirkham, N. Z., & Semcesen, T. K. (1999). Rule selection versus rule execution in preschoolers: An error-detection approach. Developmental Psychology, 35, 770-780.

Karmiloff-Smith, A. (1984). A new abstract code or the new possibility of multiple codes. [Commentary on D. Premack's article]. The Behavioral and Brain Sciences, 6, 149-150.

Karmiloff-Smith, A. (1992). Beyond modularity: A developmental perspective on cognitive science. Cambridge, MA: MIT Press.

Kendler, H. H., & Kendler, T. S. (1961). Effect of verbalization on reversal shifts in children. Science, 134, 1619-1620.

Kendler, T. S. (1963). Development of mediated responses in children. In J. C. Wright & J. Kagan (Eds.), Basic cognitive processes in children. Monograph of the Society for Research in Child Development, 28 (2, Serial No. 86; pp. 33-52).

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Kobayashi, E. L., & Cantor, J. H. (1974). The effects of dimensional naming upon children performance in a modified optional shift problem. Memory and Cognition, 2, 401-405.

Kunen, S., & Duncan, E. M. (1983). Do verbal descriptions facilitate visual inferences? Journal of Educational Research, 76, 370-373.

Levine, B., Stuss, D. T., & Milberg, W. P. (1995). Concept generation: Validation of a test of executive functioning in a normal aging population. Journal of Clinical and Experimental Neuropsychology, 17, 740-758.

Luria, A. R. (1959). The directive function of speech in development and dissolution. Part I. Development of the directive function of speech in early childhood. Word, 15, 341-352.

Luria, A. R. (1961). The role of speech in the regulation of normal and abnormal behaviour (J. Tizard, Ed.). New York: Pergamon Press.

Milner, B. (1963). Effects of different brain lesions on card sorting. Archives of

Neurology, 9, 100-111.

Milner, B., & Petrides, M. (1984). Behavioral effects of frontal-lobe lesions in man. Trends in Neuroscience, 7, 403-407.

Morgan, J. T., & Greene, T. R. (1994). An analysis of categorization style in preschoolers. Psychological Reports, 74, 59-66.

Nelson, K. (1996). Language in cognitive development: The emergence of the mediated mind. Cambridge, England: Cambridge University Press.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). Applied linear statistical models: Regression, analysis of variance, and experimental designs (3rd ed.). Homewood, IL: Irwin.

Paivio, A. (1969). Mental imagery in associative learning and memory. Psychological Review, 76, 241-263.

Pennington, B. F., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. Journal of Child Psychology and Psychiatry and Allied Disciplines, 37, 51-87.

Piaget, J. (1967). Language and thought from the genetic point of view. In J. Piaget, Six psychological studies (Ch. 3, pp. 88-98; D. Elkind, Ed.; A. Tenzer, Trans.). New York: Vintage Books. (Original work published in 1964)

Premack, D. (1984). The codes of man and beasts. The Behavioral and Brain Sciences, 6, 125-137.

Roberts, K., & Jacob, M. (1991). Linguistic versus attentional influences on nonlinguistic categorization in 15-month-old infants. Cognitive Development, 6, 355-375.

Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991). The 'windows task' as a measure of strategic deception in preschoolers and autistic subjects. British Journal of

Developmental Psychology, 9, 331-349.

Sanders, B. (1971). Factors affecting reversal and nonreversal shifts in rats and children. Journal of Comparative & Physiological Psychology, 74, 192-202.

Sandhofer, C., & Smith, L. B. (1999). Learning color words involves learning a system of mappings. Developmental Psychology, 35, 668-679.

Sandhofer, C., & Smith, L. B. (in press). Why children learn color and size words so differently: Evidence from adults' learning of artificial terms. Journal of Experimental Psychology: General.

Siegel, S. M. (1957). Discrimination among mental defective, normal, schizophrenic and brain damaged subjects on the Visual-Verbal concept formation test. American Journal of Mental Deficiency, 62, 338-343.

Smith, L. B. (1984). Young children's understanding of attributes and dimensions: A comparison of conceptual and linguistic measures. Child Development, 55, 363-380.

Smith, L. B. (1989a). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 146-178). Cambridge, England: Cambridge University Press.

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

Smith, L. B. (1993). The concept of same. In H. W. Reese (Ed.), Advances in child development and behavior, Vol. 24, (pp. 215-252). San Diego, CA: Academic Press.

Smith, L. B., Gasser, M., & Sandhofer, C. M. (1997). Learning to talk about the properties of objects: A network model of the development of dimensions. In R. L. Goldstone, D. L. Medin, & P. G. Schyns (Eds.), The psychology of learning and motivation:

~~Vol. 36. Perceptual learning. (pp. 219-225). San Diego, CA: Academic Press.~~

Smith, L. B., & Sandhofer, C. (2001, April). How word learning determines dimensional representations. In S. Jacques & P. D. Zelazo (Chairs), Language as a tool for thought. Symposium presented at the Biennial Meeting of the Society for Research in Child Development in Minneapolis, MN.

Stuss, D. T., Benson, D. F., Kaplan, E. F., Weir, W. S., Naeser, M. A., Lieberman, I., & Ferrill, D. (1983). The involvement of orbitofrontal cerebrum in cognitive tasks. Neuropsychologia, 21, 235-248.

Stuss, D. T., Eskes, G. A., & Foster, J. K. (1994). Experimental neuropsychological studies of frontal lobe functions. In F. Boller, & J. Grafman (Eds.), Handbook of Neuropsychology (Vol. 9; pp. 149-185). Amsterdam: Elsevier.

Sugimura, T. (1978). Effects of pretraining concept names on conceptual sorting task in children. Japanese Psychological Research, 20, 29-38.

Tikhomirov, O. K. (1978). The formation of voluntary movements in children of preschool age. In M. Cole (Ed.), The selected writings of A. R. Luria (pp. 229-269). New York: M. E. Sharpe.

Vygotsky, L. S. (1929). II. The problem of the cultural development of the child. Journal of Genetic Psychology, 36, 415-434.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1986). Thought and language (A. Kozulin, Ed.). Cambridge, MA: MIT Press. (Original work published in 1934)

Waxman, S., & Gelman, R. (1986). Preschoolers' use of superordinate relations in classification and language. Cognitive Development, 1, 139-156.

Welsh, M. C., Pennington, B. F., & Groisser, D. B. (1991). A normative-developmental study of executive function: A window on prefrontal function in children. Developmental Neuropsychology, 7, 131-149.

Werner, H. (1948). Comparative psychology of mental development. New York: Science Editions.

Werner, H., & Kaplan, B. (1963). Symbol formation: An organismic-developmental approach to language and the expression of thought. New York: John Wiley & Sons.

Wheeler, R. J., & Dusek, J. B. (1973). The effects of attentional and cognitive factors on children's incidental learning. Child Development, 44, 253-258.

Whitehill, R. P. (1969). Some effects of varying verbal labeling conditions on reversal- and nonreversal-shift concept attainment. Developmental Psychology, 1, 770.

Whorf, B. L. (1956). Language, thought, and reality: Selected writings of Benjamin Lee Whorf (J. B. Carroll, Ed.). Cambridge, MA: MIT Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13, 103-128.

Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. Review of General Psychology, 1, 198-226.

Zelazo, P. D., & Frye, D. (1997). Cognitive complexity and control: A theory of the development of deliberate reasoning and intentional action. In M. Stamenov (Ed.), Language structure, discourse, and the access to consciousness (pp. 113-153). Amsterdam: John Benjamins.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. Cognitive Development, 11, 37-63.

Zelazo, P. D., & Jacques, S. (1996). Children's rule use: Representation, reflection and cognitive control. In R. Vasta (Ed.), Annals of Child Development (Vol. 12, pp. 119-176). London: Jessica Kingsley.

Appendix A

Glossary of Terms.

Note. All examples are based on the example presented in Figure 6.

All-Items Response: A type of response in which children incorrectly select all three items.

Correct-Pair Response: A type of response in which children correctly select a matching pair of items.

Cue: A particular exemplar of a dimension (e.g., small for size; yellow for colour; teapot for shape).

Dimension: One of the four (Experiment 1; colour, shape, size, and number) or three (Experiments 2 and 3; colour, shape, and size) variables on which items could vary in a trial set.

Dominant Cue: The cue of a relevant dimension on which two of the three items match in a trial set.

Induced Labels: Task-relevant utterances that children provide in response to the experimenter's queries.

Irrelevant Cue: The cue of the dimension that is irrelevant in a trial set (i.e., the dimension that does not vary across the three items).

Irrelevant-Cue Condition: The condition in Experiment 3 in which the experimenter referred to the irrelevant dimension by the dominant cue (e.g., "teapot") after selecting items on Selection 1.

Irrelevant Dimension: The dimensions (two in Experiment 1; one in Experiments 2 and 3) that do not vary across the three items presented in a trial set.

Irrelevant-Dimension Condition: The condition in Experiment 3 in which the experimenter referred to the irrelevant dimension by its dimensional label (e.g., "thing") after selecting items on Selection 1.

Irrelevant-Label Condition: The condition in Experiment 2 in which the experimenter asked children about the irrelevant dimension after each selection (e.g., "What thing are these pictures?").

Irrelevant Labels: Task-relevant utterances that children provide that refer to the irrelevant dimension in someway.

Item / Card: One of the three stimuli that appear in a trial set on the FIST. In Experiment 1, the word "card" was used instead of "item" because number was used as a dimension (therefore, the term "item" could have been misinterpreted).

~~No-Cue Condition:~~ ~~The condition in Experiment 3 in which the experimenter referred~~ to items in a nondescript manner (e.g., "they are the same in one way") after selecting items on Selection 1.

No-Items Response: A type of response in which children incorrectly refuse to select any of the items.

No-Label Condition: The condition in Experiment 2 in which the experimenter did not asked children anything after each of their selections.

Nondominant Cue: The cue of a relevant dimension on which the remaining item differs from the matching pair (e.g., medium).

One-Item Response: A type of response in which children incorrectly select only one item on Selection 1.

Other Labels: Task-relevant utterances that children provide that do not refer to any of the dimensions used in the experiments.

Other One-Item Response: The subtype of one-item responses that includes all instances of one-item responses on Selection 2 other than remaining-item responses.

Relevant-Cue Condition: The condition in Experiment 3 in which the experimenter referred to the relevant dimension by the dominant cue (e.g., "blue") after selecting items on Selection 1.

Relevant Dimension: A dimension that varies in a trial set and according to which items could be selected.

Relevant-Dimension Condition: The condition in Experiment 3 in which the experimenter referred to the relevant dimension by its dimensional label (e.g., "size") after selecting items on Selection 1.

Relevant-Dimension Pair: The two dimensions that are relevant in a trial set: the pivot item matches one of the other items on one of these dimension and matches the remaining item on the other.

Relevant-Label Condition: The condition in Experiment 2 in which the experimenter asked children about the relevant dimension after selection (e.g., "Why do these pictures go together?").

Relevant Labels: Task-relevant utterances that children provide that refer to the relevant dimension in someway (i.e., the dimension according to which they select items).

Remaining-Item Response: The subtype of one-item responses on Selection 2 in which children select the item that they or the experimenter did not select on Selection 1.

~~Same-Pair Response:~~ ~~A type of Selection 2 response in which children select the same~~ matching pair that they (Experiments 1 and 2) or the experimenter (Experiment 3) selected on Selection 1.

Selection 1: The first set of item(s) that children select on a given trial in Experiments 1 and 2, or the first pair of items that the experimenter select on a given trial in Experiment 3.

Selection 2: The second set of item(s) that children select on a given trial in Experiments 1 and 2 or the only set of item(s) that children select on a given trial in Experiment 3.

Spontaneous Labels: Unprompted task-relevant utterances that children provide (i.e., utterances that children provided without first being asked by the experimenter).

Pivot Item: The item that needs to be selected twice in a trial set because it matches one of the other item on one dimension and the remaining item on the other dimension.

Pivot-Item Placement: The placement of the pivot item presented in a trial set. In Experiment 1, it could appear in one of three cards (i.e., Card 1, Card 2, or Card 3), and in Experiments 2 and 3, it could appear in one of three windows (i.e., Window 1, Window 2, or Window 3).

Trial Set: A set of three items presented simultaneously on each trial of the FIST.

Wrong Labels: Task-relevant utterances that children provide that refer to one of the dimensions, but not the irrelevant dimension or the dimension by which items are selected.

Wrong-Pair Response: A type of response in which children incorrectly select a nonmatching pair of items.

Appendix B

Counterbalancing Details for Each Version (Paper and Computerized Versions) of the Flexible Item Selection Task

Counterbalancing Information for the Paper Version of the Flexible Item Selection Task Used in Experiment 1

Table 1

Counterbalancing Information for Each Trial Set Presented in the Flexible Item Selection Task in Experiment 1 as a Function of Block and Trial Number

Block / Trial Number	Relevant Dimensions	Dominant Cues	Nondominant Cues	Irrelevant Cues	Pivot-Card Placement
Block 1					
Trial 1	size / number	medium / three	small / one	orange / socks	Card 2
Trial 2	shape / number	socks / one	phone / two	purple / large	Card 1
Trial 3	colour / shape	orange / fish	purple / socks	small / one	Card 2
Trial 4	colour / number	purple / two	pink / three	fish / medium	Card 3
Trial 5	shape / size	phone / large	fish / medium	pink / two	Card 3
Trial 6	colour / size	pink / small	orange / large	phone / three	Card 1
Block 2					
Trial 7	shape / number	fish / three	phone / two	orange / medium	Card 2
Trial 8	shape / size	socks / small	fish / medium	purple / one	Card 1
Trial 9	colour / number	orange / one	pink / three	phone / small	Card 1
Trial 10	colour / size	purple / medium	orange / large	socks / two	Card 2
Trial 11	size / number	large / two	small / one	pink / fish	Card 3
Trial 12	colour / shape	pink / phone	purple / socks	large / three	Card 3

Note. The actual cards presented on each trial can be recreated from the information provided in this table. For example, on Trial 1, children were shown, one medium orange pair of socks, three medium orange pair of socks, three small orange pair of socks, and the pivot item (i.e., three medium orange pair of socks) was placed in the centre position (i.e., Card 2).

Counterbalancing Information For the Computerized Version of the Flexible Item Selection Task Used in Experiments 2 and 3

Table 2

Relevant-Dimension Pairs and Window-Placement Combinations Used in the Computerized Flexible Item Selection Task

Relevant-Dimension Pairs	Window Placement
size / colour	12 / 23
size / colour	23 / 13
size / colour	13 / 12
colour / size	12 / 23
colour / size	23 / 13
colour / size	13 / 12
colour / shape	12 / 23
colour / shape	23 / 13
colour / shape	13 / 12
shape / colour	12 / 23
shape / colour	23 / 13
shape / colour	13 / 12
shape / size	12 / 23
shape / size	23 / 13
shape / size	13 / 12
size / shape	12 / 23
size / shape	23 / 13
size / shape	13 / 12

Note. The first dimension indicated in the first column appeared in the first window placement indicated in the second column (e.g., for size / colour and 12 / 23, items that matched each other in terms of size were located in Windows 1 and 2, whereas items that matched each other in terms of colour were located in Windows 2 and 3; for colour / size and 12 /23, the reverse applied). The window number that appears twice in bold font in the second column is the window that contained the pivot item.

Table 3

Items Used in the Computerized Flexible Item Selection Task (and Number of Times Each Appeared in Parentheses)

small blue boat (1)
small blue shoe (1)
small blue teapot (2)
small red boat (2)
small red shoe (2)
small red teapot (3)
small yellow boat (3)
small yellow shoe (2)
small yellow teapot (2)
medium blue boat (2)
medium blue shoe (2)
medium blue teapot (2)
medium red boat (3)
medium red shoe (2)
medium red teapot (2)
medium yellow boat (2)
medium yellow shoe (1)
medium yellow teapot (2)
large blue boat (3)
large blue shoe (3)
large blue teapot (2)
large red boat (1)
large red shoe (2)
large red teapot (1)
large yellow boat (1)
large yellow shoe (3)
large yellow teapot (2)

Note. Despite the fact that each item did not appear equally often, each cue (e.g., small) appeared equally often (i.e., 18 times).

Table 4

Trial Sets Used in the Computerized Flexible Item Selection Task (and Relevant Dimensions in Parentheses)

Trial-Set	Window 1	Window 2	Window 3
A	small blue shoe	large yellow shoe	small yellow shoe
	(size)	(colour)	(colour / size)
B	small yellow boat	medium yellow boat	medium red boat
	(colour)	(colour / size)	(size)
C	small red teapot	small blue teapot	medium blue teapot
	(size)	(size / colour)	(colour)
D	medium red teapot	large red teapot	medium yellow teapot
	(colour / size)	(colour)	(size)
E	small red shoe	large blue shoe	large red shoe
	(colour)	(size)	(size / colour)
F	large blue boat	large yellow boat	medium blue boat
	(size / colour)	(size)	(colour)
G	small yellow teapot	small red boat	small red teapot
	(shape)	(colour)	(colour / shape)
H	small blue boat	small blue teapot	small yellow boat
	(colour / shape)	(colour)	(shape)
I	large yellow shoe	large yellow teapot	large blue teapot
	(colour)	(colour / shape)	(shape)
J	large red shoe	large blue shoe	large blue boat
	(shape)	(shape / colour)	(colour)
K	medium yellow boat	medium red boat	medium yellow shoe
	(shape / colour)	(shape)	(colour)
L	medium red teapot	medium blue shoe	medium red shoe
	(colour)	(shape)	(shape / colour)
M	medium yellow teapot	small yellow boat	small yellow teapot
	(shape)	(size)	(size / shape)
N	medium blue teapot	large blue teapot	medium blue shoe
	(shape / size)	(shape)	(size)

(table continues)

Trial-Set	Window 1	Window 2	Window 3
O	small red boat	small red teapot	large red boat
	(size / shape)	(size)	(shape)
P	large blue shoe	medium blue boatlarge blue boat	
	(size)	(shape)	(shape / size)
Q	small red shoe	medium red shoe	medium red boat
	(shape)	(shape / size)	(size)
R	large yellow teapot	large yellow shoe	small yellow shoe
	(size)	(size / shape)	(shape)

Note. The first row within each trial set contains the actual items used in the trial set and the second row contains the dimension(s) that were relevant for each item. As a result, the column with two relevant dimensions contains the pivot item for that particular trial set.

~~Table 5~~

Quasi-random Presentation Orders of the 18 Trial Sets (and Relevant-Dimension Pairs in Parentheses) Used in the Computerized Flexible Item Selection Task

	Quasi-random Orders					
Trial	Order 1	Order 2	Order 3	Order 4	Order 5	Order 6
1 (Demo[a])	C (zc)	F (zc)	L (cs)	K (cs)	O (sz)	R (sz)
2 (Practice[b])	M (sz)	I (cs)	N (sz)	E (zc)	J (cs)	D (zc)
3 (Practice)	H (cs)	P (sz)	B (zc)	Q (sz)	A (zc)	G (cs)
4	B (zc)	C (zc)	H (cs)	G (cs)	N (sz)	Q (sz)
5	P (sz)	L (cs)	M (sz)	D (zc)	L (cs)	E (zc)
6	K (cs)	Q (sz)	C (zc)	R (sz)	B (zc)	J (cs)
7	E (zc)	J (cs)	P (sz)	J (cs)	M (sz)	N (sz)
8	I (cs)	E (zc)	G (cs)	A (zc)	K (cs)	H (cs)
9	Q (sz)	O (sz)	Q (sz)	O (sz)	C (zc)	B (zc)
10	D (zc)	H (cs)	A (zc)	I (cs)	H (cs)	L (cs)
11	L (cs)	M (sz)	J (cs)	B (zc)	P (sz)	O (sz)
12	N (sz)	A (zc)	O (sz)	L (cs)	F (zc)	I (cs)
13	F (zc)	N (sz)	I (cs)	P (sz)	G (cs)	C (zc)
14	G (cs)	G (cs)	D (zc)	C (zc)	E (zc)	P (sz)
15	A (zc)	B (zc)	R (sz)	H (cs)	R (sz)	K (cs)
16	R (sz)	K (cs)	F (zc)	M (sz)	D (zc)	F (zc)
17	J (cs)	R (sz)	K (cs)	F (zc)	I (cs)	M (sz)
18	O (sz)	D (zc)	E (zc)	N (sz)	Q (sz)	A (zc)

Note. The letters assigned to each trial set are the same letters assigned to trial sets in the previous table (Table 4). For example, the first trial set presented in Order 1 is Trial-Set C, which is a small red teapot, a small blue teapot, and a medium blue teapot (see Table 4). Abbreviations for the relevant-dimension pairs are zc for size and colour, cs for colour and shape, sz for shape and size.

[a] Demo refers to the demonstration trial. [b] Practice refers to the practice trials.