

# 6

---

## The Dual Components of Developing Strategy Use

### *Production and Inhibition*

Deanna Kuhn  
Maria Pease

Contemporary researchers who study the development of cognitive strategies address a distinctly different set of issues than did those who approached the topic in its infancy in the 1970s. The evolution in study of this topic can be traced from early assumptions that the capacity to behave strategically did not develop until later in childhood to the contemporary recognition that even infants can be strategic. Moreover, the study of strategy development has become much more complex. Rarely do we see over time a simple transition from the application of one kind of strategy to a problem to the application of a new, different strategy. Instead, it is now clear, individuals have a repertory of strategies they bring to a new situation, some more adequate or advanced than others. The task of microgenetic analysis over time, in a context of repeated encounters with the problem, is thus to examine the nature of strategy *selection*, which itself evolves over time (Kuhn & Phelps, 1982; Kuhn, 1995, 2001; Siegler, 2006). The now well-replicated finding is that more advanced modes become more frequent and less advanced ones less frequent, although in an uneven and not entirely predictable way. The period of time in which a mixture of more and less advanced strategies are applied variably may be prolonged.

The implication is that strategy development involves much more than learning to execute a strategy. For the evolution just described to take place, two distinct challenges must be met: The less advanced (and likely more habitual) mode of response must be repeatedly inhibited and the more advanced (and initially weaker) mode of response must be consolidated and strengthened. The question we examine here is this: During the often extended periods of transition observed in microgenetic studies, how are these two challenges related to one another?

Two possibilities seem viable. One is that the two are inversely related, that is, occurrence of the advanced response mode in a given instance makes occurrence of the less advanced mode less likely (the context being one in which exhibiting one does not preclude also exhibiting the other). The other possibility is that the two challenges are met independently—in other words, advanced responses must be executed and less advanced ones must be inhibited, but the one occurrence has no influence on the other one. Whichever of these possibilities is correct, there are two distinct tasks to be accomplished, each with its own set of challenges, if change is to occur. One is increased selection and execution of the better strategy. The other is stronger inhibitory control of the inferior strategy.

These are the two components of strategy development, and their connection to one another, that we examine in the research described in this chapter. Doing so requires us to address all of the themes of this volume. Metacognition, we claim, is central to strategy selection. And the instructional implications of our topic are significant. How are the multiple challenges of strategy development met in instructional contexts? And how are these developmental challenges best supported?

### **THE PROBLEM CONTEXT: UTILIZING STRATEGIES OF INVESTIGATION AND INFERENCE IN INQUIRY**

The problem context in which we examine these questions is the complex, multifaceted one of scientific inquiry, although we focus on the inference phase of the inquiry process, thus also situating the task in the research literature on inductive multivariable causal inference (Kuhn & Dean, 2004). In self-directed scientific inquiry (see Lehrer & Schauble, 2006, or Zimmerman, 2007, for review of studies), the individual has access to a database and is asked to plan and execute an investigation and to draw and justify inferences regarding the relations among variables depicted in the database. Typically, multiple potential independent variables may influence a dependent variable, and the task is to examine the database and make inferences regarding which of the variables bear a causal relation to the outcome and which do not.

Here we focus on the conclusions individuals draw on the basis of their investigation, as these constitute the culmination of the inquiry process. We divide them into the two broad categories of valid judgments and invalid judgments (Schauble, 1990; Kuhn, Schauble, & Garcia-Mila, 1992; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995). *Valid judgments* are judgments the individual draws on available evidence to justify, in a manner adequate to support the judgment. (Specific examples are presented later.) *Invalid judgments* are those lacking justification adequate to support them. Valid judgments (that a variable is causal or noncausal) are therefore always correct, whereas invalid judgments may be incorrect or correct (regarding the variable's true causal status). In the multivariable causal context described, a valid judgment requires the individual to have accessed from the database and compared at least two instances that differ with respect to only a single variable (what has come to be known as a *control-of-variables* strategy), allowing an inference to be made regarding how variation in that variable affects outcome. The strategy application that leads to a valid judgment therefore requires intention and planning, to identify appropriate instances to compare to one another, to secure them from the database while withholding any inferential judgment, and then to analyze the pattern of outcomes as the basis for making a judgment of causality (that the focal variable makes a difference) or noncausality (that it does not).

An invalid causal (or noncausal) judgment, in contrast, can be made quickly and intuitively, by observing no more than a single instance and outcome. When justification for such a judgment is solicited, the most common one is co-occurrence (or association): Because a particular level of a variable was present when the outcome occurred, that variable is implicated as having played a role in the outcome. Occasionally, an invalid judgment may make reference to a previous instance in which both variable and outcome were absent, but no comparative analysis is undertaken across instances (especially one that would identify uncontrolled variables). The most common type of invalid judgment, however, is one that ignores the evidence entirely and is based on retrieval of the respondent's previous knowledge or beliefs regarding the content at hand. (Examples of each of these types are presented shortly). The reasoning required to produce invalid judgments of any of these types is therefore minimal. Each of the types has been found to occur among both children and adults but to diminish with age and with experience with problems that entail investigatory and inference skills (Kuhn et al., 1995). Microgenetic analyses of performance over time reveal the typical pattern of prolonged periods of mixed usage of both valid and invalid inference strategies, with a gradual increase over time in the proportion of use of valid strategies (Kuhn et al., 1992, 1995; Schauble, 1990, 1996).

In the context of interpreting a single outcome from the database, an individual can thus make both valid and invalid judgments, in so doing presumably drawing on multiple kinds of inference strategies. Of five variables that are identified, with levels of each occurring in conjunction with an outcome across a succession of instances, for example, an individual might make the valid judgment that a particular variable is causal on the basis of a comparison of the outcome in the current instance to a previous one in which the level of only this variable differed and the outcome varied (i.e., a controlled comparison). At the same time, as has been documented to happen frequently, in responding to this instance the individual might also identify a second variable as causal, but on the basis only that a level of this variable also was present in conjunction with the outcome being examined and therefore must have contributed to it. Other than declaring a variable causal or noncausal, a third option with respect to each of the variables is to suspend judgment and claim that the causal status of that variable is not yet certain.

Hence, in evaluating a given instance (an outcome in conjunction with different levels of the five identified variables), while only one judgment is made about any one variable, multiple judgments (of causality, of noncausality, or of uncertainty)—valid or invalid—may be made regarding the variables identified in the instance. In subsequently evaluating another instance, these judgments (regarding a variable's causal status) may change. Judgments have been observed to fluctuate as individuals evaluate successive instances (Kuhn et al., 1995; Schauble, 1990).

## **A MICROGENETIC INVESTIGATION**

In the context of the scientific inquiry problem we have described, change can be examined not only in the knowledge an individual acquires about the causal system but also in the strategies of investigation and inference by means of which this knowledge is acquired (Kuhn, 1995; Kuhn et al., 1995). Researchers who have used the microgenetic method report a similar pattern of change. At all points multiple strategies are available and applied, but change occurs in the form of a shifting frequency distribution (Kuhn, 2001; Kuhn & Phelps, 1982; Kuhn et al., 1995) or overlapping waves (Siegler, 1996, 2006). That is, with continued engagement less effective strategies come to be used less frequently and more effective strategies begin to be used more frequently.

The data we bring to bear on this question here are microgenetic (Kuhn, 1995; Siegler, 2006) data—that is, they entail repeated observations of the same individuals engaged in the same or similar problems over time, allowing examination of patterns of change across time. The

data are drawn from a larger 3-year longitudinal study in which we follow the development of inquiry skills among students beginning in their fourth-grade year as they encounter a sequence of problems of increasing complexity (Kuhn & Pease, 2008). The specific analyses presented here, addressed to the specific question we have identified, were not included in the report of that study as they were not central to the longitudinal developmental questions that were the focus of that work.

Our purpose in examining microgenetic data in the present work, then, is not the typical one of examining patterns of change over time. Instead, we turn to such data to address the particular question identified above: whether occurrence of more and less advanced response modes operate independently or are (inversely) related to one another. This is a different question from that of how they change over time. One type of judgment may become more frequent and another type less frequent over time, but this does not tell us whether one of these trends in some way governs or influences the other or whether the two trends take place independently of one another. Repeated-observation data involving individuals working on the same or similar problems over time are necessary to address our question as the question is one about variation in responses to the same kind of problem on different occasions.

One other feature of our research design that warrants noting at the outset is that participants' problem-solving activity is situated in a social context. During most sessions, students work on the task in pairs. We regard this feature as advantageous in any case, since cognition very frequently occurs in a social context. But it also stands to provide a second, less direct kind of evidence regarding the independence of the two components of strategy change. Other people can serve as external influences on individual cognition. In particular, the thinking they display is likely to have an influence on an individual's propensity to rely on one or the other mode of response. Moreover, it is possible that this external influence functions differently in the respective cases of the two different modes.

As our participants worked most of the times with a series of changing partners (except for initial and subsequent individual assessments), we sought to examine how the social context of working with a same-level, higher level, or lower level peer influenced a participant's propensity to make judgments of the two types. Conceivably, this influence of social context on performance may be different for the two kinds of judgments. One, for example, as we in fact speculated might be the case, may be more susceptible to social influence than those regarded as of a more advanced type. If any such differences (in the effect of social context) across the two kinds of judgments do in fact emerge, they stand to serve as additional evidence of a second, less direct type, regarding the independence of the two modes.

The 34 fifth-grade students reported on here began participation in our larger, longitudinal study of the development of inquiry skills when the students were fourth graders and continued through their sixth-grade year (Kuhn & Pease, 2008). Students were from an urban independent school serving a socioeconomically and ethnically diverse population. As would be expected among this age group, all 34 met a criterion of being in the process of developing scientific inquiry skills. Specifically, they showed variable usage (across occasions) of effective and ineffective strategies, as detailed below. One participant was eliminated because he showed no variation (i.e., exclusive use of ineffective strategies) at all sessions.

Students worked with inquiry software for one or sometimes two 45-minute periods per week, except when occasional special school activities or field trips intervened. Students worked in pairs, with pair composition varying across sessions, except for the initial one or two sessions allocated to initial assessment of individual skill levels, and a later final assessment carried out individually for the same purpose. The sessions on which the present analyses are based began in late October of the fifth-grade school year and continued into early May. Due to school absences and other reasons students had to be away from class, the number of sessions a student participated in varied across students, from a low of 8 to a high of 15 (mean = 10.53).

A sample of one version of the software, *Earthquake Forecaster*, is presented in Figures 6.1–6.4. *Earthquake Forecaster*, and several other parallel programs are multimedia inquiry software programs created with Adobe Director multimedia authoring software as Flash files (Kuhn & Dean, 2005; Kuhn, Katz, & Dean, 2004). The program requires students to assess the causal status of five dichotomous variables in contributing to the level of earthquake risk. The introduction to the program explains the importance of developing means to predict earthquakes in order to protect others and maintain safety. To accomplish this, students must learn which features do and do not make a difference. Of the five features that students investigate in *Earthquake Forecaster*, two have no effect and three have simple (noninteractive) causal effects.

After the initial introduction, students are asked to choose what they will find out about in their first selection of an instance (or case) to examine (see Figure 6.1). Students identify whether they are or are not finding out about a feature by clicking the feature picture(s) corresponding to their choice(s). Then, students construct an instance of their own choosing, by selecting the level of each feature (see Figure 6.2). These choices yield an outcome displayed in the form of a gauge representing the earthquake risk level. Students are then asked to make and justify any causal or noncausal inferences they believe to be justified regarding the

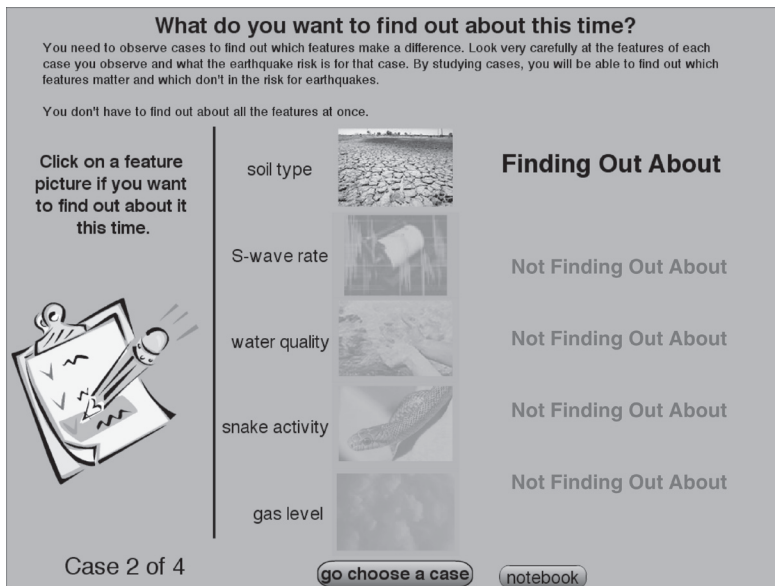


FIGURE 6.1. Find out screen.

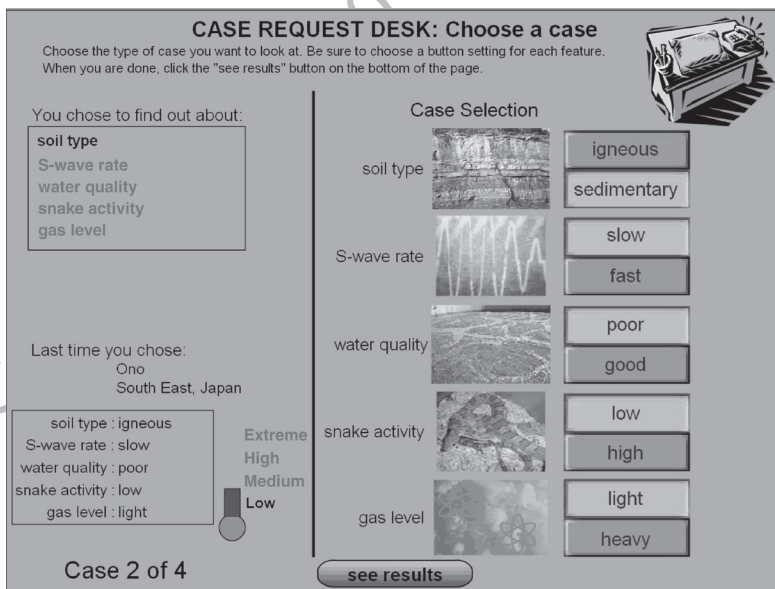


FIGURE 6.2. Case request screen.




status of any of the features (Figure 6.3). Or, for each feature, they have the option of suspending judgment (Figure 6.3). The final screen prompts the student to enter any notes they wish to (Figure 6.4).

Each of the screens shown in Figures 6.1–6.4 is depicted as it would appear during the course of the second instance the student chooses for investigation. For second and subsequent instances, the screen includes not only the outcome for the current instance the student is investigating but also shows the outcome for the instance chosen immediately preceding this one. After the student answers questions regarding the outcome of the fourth instance and is prompted to make any additional notes that may be desired, the program thanks the student for participating and shuts down.

After the initial one to two sessions assessing individual skill levels, students began working in pairs on different versions of the software that were structurally equivalent to *Earthquake Forecaster*. The pair made a single joint response at each prompted point in the program, and this response was taken as the response for each of the individuals that made up the pair. The work was done in a 45-minute class that met twice a week for most of the school year. The class was described to students explicitly as a class in inquiry, which was defined for the class as ways of asking

### Case Files: See Results and Draw Conclusions

Look at the risk for this case. What did you find out this time?  
When you are done, click the "continue" button at the bottom of the page.










Case Results	What do these results show?	
This Case Tokyo South East, Japan  soil type : sedimentary    Extreme S-wave rate : slow            High water quality : poor snake activity : low         Medium gas level : light  <div style="text-align: right;">   Low         </div>	Did this feature make a difference?      What told you so?	
Last time you chose: Ono South East, Japan  <div style="border: 1px solid gray; padding: 5px; width: fit-content;">             soil type : igneous            Extreme              S-wave rate : slow            High              water quality : poor              snake activity : low         Medium              gas level : light   <div style="text-align: right;">   Low           </div> </div>	<div style="display: flex; align-items: center;">  <div style="margin-left: 10px;"> <b>soil type</b>  <input type="button" value="Yes"/> <input type="button" value="No"/> <input type="button" value="Not Sure"/> </div> <div style="margin-left: 20px; border: 1px solid gray; padding: 5px;">             I found out this does not makes a difference because...           </div> </div>	<div style="display: flex; align-items: center;">  <div style="margin-left: 10px;"> <b>S-wave rate</b>  <input type="button" value="Yes"/> <input type="button" value="No"/> <input type="button" value="Not Sure"/> </div> <div style="margin-left: 20px; border: 1px solid gray; padding: 5px;">             I'm not sure if this makes a difference because...           </div> </div>
	<div style="display: flex; align-items: center;">  <div style="margin-left: 10px;"> <b>water quality</b>  <input type="button" value="Yes"/> <input type="button" value="No"/> <input type="button" value="Not Sure"/> </div> <div style="margin-left: 20px; border: 1px solid gray; padding: 5px;">             I'm not sure if this makes a difference because...           </div> </div>	<div style="display: flex; align-items: center;">  <div style="margin-left: 10px;"> <b>snake activity</b>  <input type="button" value="Yes"/> <input type="button" value="No"/> <input type="button" value="Not Sure"/> </div> <div style="margin-left: 20px; border: 1px solid gray; padding: 5px;">             I'm not sure if this makes a difference because...           </div> </div>
	<div style="display: flex; align-items: center;">  <div style="margin-left: 10px;"> <b>gas level</b>  <input type="button" value="Yes"/> <input type="button" value="No"/> <input type="button" value="Not Sure"/> </div> <div style="margin-left: 20px; border: 1px solid gray; padding: 5px;">             I'm not sure if this makes a difference because...           </div> </div>	
<b>Case 2 of 4</b>	<input type="button" value="continue"/>	

FIGURE 6.3. Results and conclusions screen.



**Notebook**

Do you want to put anything in your notebook about what you found out?

**Session:**  
8/5/2004

I learned that soil type makes no difference because I compared a case with sedimentary soil type and one with igneous soil type (with everything else the same) and there was no difference in earthquake risk.

---

Notes:                      Notebook

\_\_\_\_\_

\_\_\_\_\_

FIGURE 6.4 Notebook screen.

questions and seeking answers. In working with a partner, students were instructed not to divide the task (i.e., for one student to make responses to one segment of the program and the other student to another) and sufficient adult “coaches” circulated among students to ensure this did not happen. Students were instructed instead to discuss each question or choice with their partner and not to respond until agreement had been reached between them.

The first program student pairs worked on was *Avalanche Hunter*. Wind type, snow-type cloud cover, soil, and slope were the five binary variables potentially having causal effects on avalanche risk. Each content version of the software also contained a prediction module that students worked on, to apply one’s learning (by predicting outcomes from different variable constellations), but here we focus on just the inquiry strategies themselves and in particular the inference phase of the inquiry process. Work with *Avalanche Hunter* continued from late October to mid-December, by which time the majority of students had achieved a high degree of mastery, although, as detailed below, they still showed less than 100% consistent optimal strategy usage.

Other more advanced forms of the software were elaborations of the structure of the basic program. These enhanced the challenge of students’ inquiry by introducing more complex forms of evidence. Beyond the scaf-

folding provided by the software itself, the two adult “coaches” supervising the sessions provided one further scaffold in the form of encouragement to find out about one variable at a time (as necessary, among students who did not formulate this intention without assistance). This scaffold was introduced as earlier work (Kuhn & Dean, 2005) had shown it to be highly effective in structuring students’ activity and enhancing progress in investigatory and inference strategies.

Following winter vacation, when inquiry sessions resumed in mid-January, a new form of *Avalanche Hunter* was introduced, one in which one variable (cloud cover) had twice as large an effect as the other causal variables, and students were asked to indicate whether any of the variables were more important than any others. By mid-February two-thirds of the students had mastered this problem and were ready to move on (they had correctly identified all causal and noncausal effects using appropriate methods and justifications for inferences), while the remaining one-third did not meet this criterion and were provided more practice with the basic software. The latter group thus switched to new content to maintain their interest: the *Ocean Voyage* program (in which ancient ships varying on five dimensions vary in the success of their voyages), which did not contain any further structural advance. During this same period, the more advanced group also worked with *Ocean Voyage*, but in their case a more advanced probabilistic version of *Ocean Voyage* was introduced, one in which the outcome for a particular constellation of variable levels was not constant but rather took the form of a distribution with one outcome (voyage distance) most frequent but adjacent outcomes of lesser and greater distance also occurring but with lower frequency.<sup>1</sup> Students thus had to compare results over multiple trials with the same constellation (of variable levels) in order to make informative comparisons between two constellations.

In late March there occurred for all students a phase of individual assessment, returning to the basic structure of *Earthquake Forecaster*. The purpose was to assess how much progress each student had made individually, in the absence of the influence of working with a peer. Students individually required between one and two sessions to complete the *Earthquake Forecaster* program (both investigation and prediction modules) at least once.

Following completion of the individual assessment, and a brief vacation, at the end of April and through mid-May, all students encountered a final new data structure, presented within the *Earthquake Forecaster* content, in which two of the three causal variables interacted with one another.<sup>2</sup> Students returned to working in pairs and worked with the interaction database from one to four times depending on the time available.

## Identification of Strategies and Classification of Judgments

In one cycle of the program, the participant (or pair of participants) had the opportunity to examine four instances. A valid judgment is not possible until a second instance is examined for comparison with the first. Hence, valid judgments become possible following examination of the second instance. A second valid judgment becomes possible following examination of the third instance (since the third can be compared to the first or second), and a third valid judgment becomes possible following examination of the fourth instance. If the individual or pair continue on the same occasion to engage in a second iteration of the program, the fifth instance they examine allows for the possibility of another valid judgment, and so on. The number of possible valid judgments at a single session therefore ranged from a low of one (since participants occasionally failed to complete a cycle at a given session) to a high of 11, with a median of three.

Each instance, beginning with the first, in contrast, allowed for the possibility of 5 invalid judgments, since an invalid judgment of causality or noncausality could be made about each of the five variables for each of the four instances in the cycle. The range of possible invalid judgments per session thus ranged from a low of 5 to a high of 60, with a median of 20 (4 judgments  $\times$  five variables). Additional evidence regarding a judgment came from the justification the individual (or pair) offered for it. The four principal types of justifications for determinate inferences appear in Table 6.1. For ease of comparison, the examples in Table 6.1 all refer to the same variable and to a judgment of causality. Justifications of noncausal judgments are parallel except that no difference (in outcome) is present and the respondent accordingly concludes that the variable does not make a difference to the outcome.

In order to generate the fourth justification type, note, the student would have had to construct the two instances in order to compare them and draw the appropriate inference. In the case of the first three types, no such intentional construction of instances is necessary. For Type 1, no instances of evidence are invoked to support the judgment. For Type 2, any single instance will suffice, and in Type 3 just about any two instances, with no fixed relation to one another, will suffice.

It is on this basis, then, that we regarded the fourth type as signaling a more reflective, analytic type of processing. Generation of a controlled comparison is unlikely to happen by chance (and, indeed, rarely occurred in the absence of the appropriate justification). Even once the evidence has been generated, the student must recognize its relevance, make the relevant comparison, and draw the appropriate conclusion. Although the

**TABLE 6.1. Types of Justifications for Determinate Judgments (of Causality or Noncausality)**

Justification type	Example
1. Absence of evidence-based justification	“The heavy gas level means high risk, because the gas has bad chemicals in it.”
2. Single-instance justification	“The heavy gas level increases the risk, because here you have heavy gas and the risk is high.”
3. Cross-instance uncontrolled comparison	“The heavy gas level increases the risk, because here you have heavy gas and the risk is high. Before, when the level of everything was good, the risk was low.”
4. Cross-instance controlled comparison	“In this instance only the gas level changed, compared to the last instance, and the risk increased. So the gas level makes a difference.”

first three types in Table 6.1 arguably involve some level of reasoning, it is neither complex nor effortful and can be accomplished by the sort of covariation assessment that even infants are capable of (Alloy & Tabachnik, 1984).

Note we do not include indeterminacy judgments (“not sure”) in the analysis since the kind of processing underlying them is likely to vary. An indeterminacy judgment might arise from a close analysis of the available evidence and recognition that the evidence is insufficient to permit an inference regarding causality. Or it might arise from a nonreflective subjective sense of uncertainty. Typical justifications for indeterminacy judgments, for example, “I’m not sure yet,” are often difficult to distinguish in this respect. Accordingly, only determinate judgments (the variable makes a difference or doesn’t make a difference) were coded. Two indices were calculated for each individual (or pair) at each session, as the basis for further analysis. One was the proportion of valid determinate judgments (the proportion being the number of valid judgments divided by the number of possible valid judgments). The other was the proportion of invalid determinate judgments (the number of invalid judgments divided by the number of possible invalid judgments).

### **Patterns of Change over Time**

The general pattern of change evident in earlier studies (Kuhn et al., 1995; Schauble, 1990) appeared in the present work as well, when children worked most of the time with partners. Examining change first of all in terms of qualitative patterns, at the initial individual assessment six of 34 students made all possible valid judgments (the exact number possible varying slightly across individuals depending on how many instances they constructed) and showed no invalid judgments, thus per-

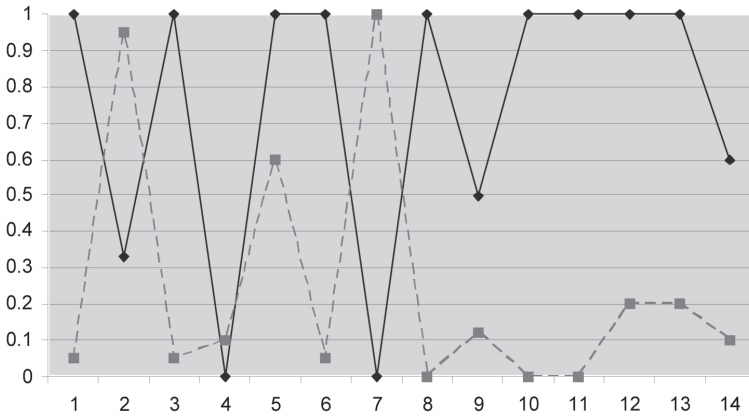
forming at ceiling on both of these dimensions.<sup>3</sup> (None of these six maintained this record, however, when they went on to work with partners.) At the individual posttest, the number of students performing at ceiling on both dimensions increased to nine of 34. Nineteen of the 34 showed at least one of these achievements (maximum possible valid judgments or no invalid judgments), compared to 11 at the pretest. Of the remaining 15, who did not achieve ceiling performance on either dimension, nine showed progress on both dimensions (increasing valid inferences and decreasing invalid ones) and an additional three showed progress on one or the other. Among those students who showed progress only on one of the two dimensions, most progressed on the dimension of reduction of invalid judgments.

Quantitative analysis of the change data confirmed that the group as a whole made significant progress on both dimensions, with repeated-measures analysis of variance yielding a significant effect of time (initial vs. final assessment) with respect to proportion of valid judgments (which increased over time) and proportion of invalid judgments (which decreased over time). Proportion of valid judgments increased from a mean of .392 to a mean of .794 across the two assessments,  $F(1,29) = 20.73, p < .05$  (partial eta squared = .417). Also significant was the effect of time with respect to the proportion of actual invalid judgments to possible invalid judgments (the latter number depending on the number of instances examined). This proportion decreased from a mean of .381 to a mean of .113 across the two assessments,  $F(1, 29) = 23.67, p < .05$  (partial eta squared = .449).

An illustration of one student's change over time appears in Figure 6.5. A second student's record appears in Figure 6.6. In each the solid line represents valid inferences and the dotted line invalid inferences. Individual sessions occurred on occasions, 1, 11, and 12 for Anna and on occasions 1, 10, and 11 for Sasha. On all other occasions, participants worked with a partner. As reflected in Figures 6.5 and 6.6, performance is highly variable over time. This variability can be attributed to a combination of the student's own intraindividual variability (as documented in earlier research in which participants worked alone) and variability attributable to the influence of the partner. In both the cases shown, variability diminishes over time, but does not disappear, as valid judgments increase in frequency and invalid judgments decrease.

### **Connections between Applications of Superior Strategies and Inhibition of Inferior Strategies**

We turn now to the central question posed in the present study—the relation between appearance of valid judgments and appearance of invalid



**FIGURE 6.5.** Anna's performance over time. Varying effect sizes were introduced at Session 6 and probabilistic effects at Session 8. Interaction effects were introduced at Session 12. Solid line depicts proportion of valid judgments that were made (relative to the total number of valid judgments possible, which varied based on the number of instances the student constructed). Dotted line depicts proportion of invalid judgments (again, relative to the total number of invalid judgments possible).

judgments. We first looked for any evidence that patterns of performance over time differed for the two kinds of judgments. Such differences would be suggestive of independence in their functioning. Examining the charts of performance over time (like those shown in Figures 6.5 and 6.6) for each of the participants, a participant's variability over time appeared to be somewhat greater in making valid judgments than in making invalid judgments. To verify this difference, we computed for each participant the standard deviation (in proportion of valid judgments) across all of that participant's sessions, first for valid judgments and then for invalid judgments. This analysis supported our observation. For 28 of the 34 participants, standard deviation was higher for valid judgments than invalid judgments. Median standard deviations across participants (based on percentage scores from 0 to 1.00 for each participant) were .41 for valid judgments and .28 for invalid judgments, a significant difference,  $F(1, 33) = 39.28, p < .001$  (partial eta squared = .543).

The next question we asked is whether such a relation emerges in the individual data, when participants are working alone. For this purpose we examined first the pretest data and then the posttest data to ascertain whether a relation appeared. For each we examined the relationship between students' pretest (or posttest) scores for valid judgments and pretest (or posttest) scores for invalid judgments. We included only those

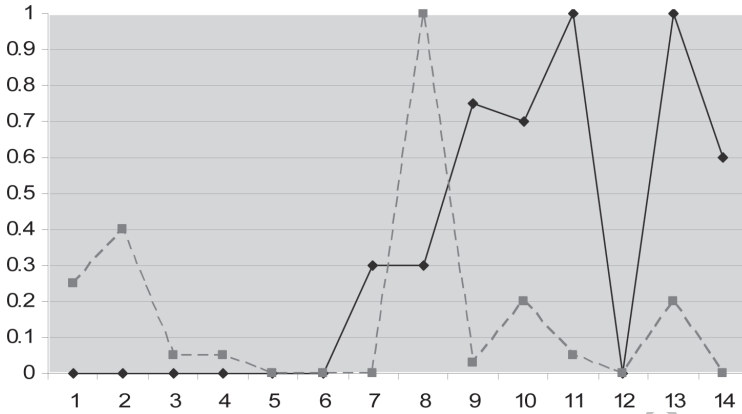


FIGURE 6.6. Sasha's performance over time.

participants who showed variability on both dimensions (across the multiple instances the participant evaluated in this assessment). Omitted were those (identified above) who had reached asymptote of perfect performance on both dimensions. The data take the form of the same percentages illustrated in Figures 6.5 and 6.6.

Using these percentages, comparisons can be made across individuals and across judgment types within individuals. Within individuals, a negative association would be predicted if the two judgment types are related, that is, high likelihood of making valid judgments, presumably driven by an analytic system, would be associated with low likelihood of making invalid judgments, presumably driven by a heuristic system. Because these percentages can be assumed to have no more than ordinal properties, the nonparametric gamma index of association was calculated for each participant. The gamma statistic  $G$ , first discussed by Goodman and Kruskal, is appropriate for measuring the relation between two ordinally scaled variables (Siegel & Castellan, 1988).

For the pretest individual data, the association between proportion of valid judgments and proportion of invalid judgments was negative, as expected, but did not reach an .05 level of significance.<sup>4</sup> For the posttest individual data, this association similarly was negative but did not reach an .05 level of significance. A scatter plot for the posttest data is shown in Figure 6.7. The scatter plot for the pretest appears very similar and is not shown. As seen there, deviations from an inverse association are frequent—some individuals make a high proportion of valid judgments but also make a high proportion of invalid judgments, while others make a low proportion of both kinds of judgments.<sup>5</sup>



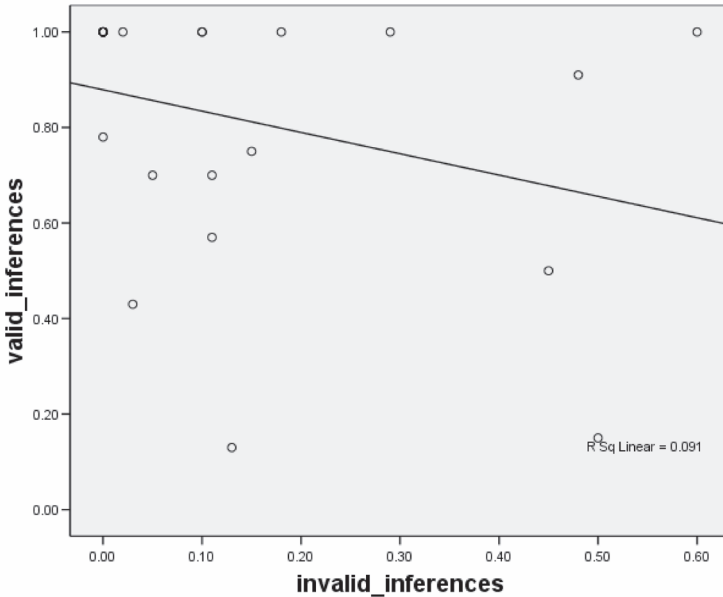


FIGURE 6.7. Relations between valid and invalid judgments in individual post-test data.

We turned next to the intervening sessions when participants worked with a partner. It is possible that a relation between the two kinds of judgments emerges only here, when the influence of a partner increases the variability in a participant's judgments. We thus examined performance over time while students worked with partners and investigated whether any relation appears between a given participant's level of functioning in making valid judgments and level of functioning in making invalid judgments. For this analysis, each participant's record of performance over time was examined individually. Because we had information about each participant's level of functioning when working alone, we were less interested in an absolute level of performance and rather whether this level in the dyadic context was higher, lower, or equivalent to the participant's own level when working alone.

Accordingly, for each dyadic session the proportion of a participant's valid judgments was compared to the same proportion when the participant worked alone,<sup>6</sup> and categorized as either higher, lower, or equal to the solitary level. A parallel categorization was made for invalid judgments. The majority of participants showed varied records in this respect, on occasions performing at a level equivalent to their individual level, on

others below it, and on others above it. This variability was influenced by how often an individual's pairing was with a more able, less able, or equally able partner, as we go on to examine.

For each participant, the ordinal gamma statistic was again employed to examine the relation between an individual's level of functioning relative to partner in making valid judgments and level of functioning making invalid judgments. The gamma statistic showed a significant (inverse) relation at the .05 level for only six of the 34 participants. When the .01 level of significance is used, only three of the 34 are significant. We can thus draw the same essential conclusion we did in examining records of individual performance. The relative frequency of valid judgments and relative frequency of invalid judgments do not appear to be related.

### **Does Social Influence on Production versus Inhibition Differ?**

These findings led us to ask the question of what might be related to the variability over time in a participant's level of functioning in making the two kinds of judgments. In particular we were curious about the likely influence of the partner. Does a partner affect the two kinds of judgments in the same way? To examine partners' influence, on every occasion in which a participant worked with a partner, we identified the partner's level of functioning as higher than, equal to, or lower than the participant's, separately for valid judgments and invalid judgments. Partner's level of functioning was identified in the same way as was the participant's level of functioning but the comparison determining the designation of high, low, or equal in this case was between the participant's and the partner's individual level of functioning.<sup>6</sup> Since the relation of the participant's and the partner's level was a matter of chance (partners were not assigned to represent particular degrees of mismatch), most participants' records contained a mixture of the three types (partner higher than, equal to, or lower than participant), although in a few cases not all three types appeared. For the sample as a whole, the median percentage of occasions at which the partner was more able was 43.5%, was equally able was 20%, and was less able was 25%.<sup>7</sup> These percentages are similar when broken down by type of judgment (valid or invalid).

Using the same gamma statistic and analytic procedure described above, for valid judgments we found statistically significant positive relations for 68% of the individual participant's records<sup>8</sup> between the level of the participant's functioning (assessed as his or her individual level) and the level of the partner relative to the participant. For invalid judgments, the gamma coefficient was less often significant but still well above a chance level—35% of participants showed a significant positive relation

between the level of the participant's functioning and the level of the partner's functioning relative to the participant. Significant gamma coefficients (which have a potential range of  $-1.00$  to  $+1.00$ ) ranged from .78 to 1.00 across the sample. These high positive associations signify that a higher functioning partner tended to improve a participant's performance (relative to his or her individual level), while a lower functioning partner tended to weaken the participant's performance.

These results, confirming that partners did have an influence on one another, led us to examine finally the interesting question of the relative degree of social influence and hence performance variability for the two types of judgments, valid versus invalid, as well as for the two types of influence: a higher functioning partner (with the potential to improve one's performance) and a lower functioning partner (with the potential to weaken one's performance). To conduct this analysis, we computed for each participant the percentage of dyadic sessions in which their performance improved (relative to individual level) when working with a higher level partner and the percentage in which it declined when working with a lower level partner, separately for valid judgments and for invalid judgments.

These data were subjected to a two-way repeated-measures analysis of variance with judgment type (valid vs. invalid) one factor and partner level (higher or lower than participant) as the other. (Cases in the "equivalent" category were not examined.) The dependent variable was the proportion of instances in which the participant's performance shifted (relative to solitary level) in the direction of the partner (i.e., was higher in the case of a superior-performing partner or was lower in the case of an inferior-performing partner). The means for the four resulting cells appear in Table 6.2. As reflected there, both partner level and judgment type were shown to have an effect. The effect of judgment type was significant,  $F(1, 33) = 9.63$ ,  $p < .004$  (partial eta squared = .226), as was the effect of partner type,  $F(1,33) = 4.31$ ,  $p < .046$  (partial eta squared = .116). The interaction between the two was nonsignificant,  $F(1,33) = .162$ ,  $p = .69$ . The numbers in parentheses in Table 6.2 are the respective medians. While very similar to the means, they establish that the patterns reflected in Table 6.2 are not the product of only a few extreme-scoring participants.

In sum, these analyses suggest that the social context of working with a partner does influence an individual's performance. A partner is more often influential in raising a participant's functioning than in lowering it. A partner's influence (either positive or negative), moreover, is more pronounced in the case of invalid judgments (which need to be inhibited to improve performance) than it is in the case of valid judgments (which need to be constructed to improve performance).<sup>9</sup>

**TABLE 6.2. Proportion of Occasions in Which a Partner Influenced Participant's Level, by Judgment Type and Direction of Partner Mismatch**

	Participant improved with superior partner	Participant declined with inferior partner
Valid judgments	38.88 (30)	27.88 (24)
Invalid judgments	49.76 (50)	35.06 (30)

*Note.* Improvement is defined as a higher level of functioning than that shown by the participant when performing alone. Decline is defined as a lower level than that shown by the participant when performing alone. Numbers in parentheses are the respective medians.

## CONCLUSIONS: TOWARD A DUAL-PROCESS MODEL

The various analyses we report here all support the independence model, both those that directly examine the relation between judgments associated with two response modes and those that show differential effects of other variables, notably social influence, on judgments associated with the respective modes. These findings warrant replication with different populations of different age levels and with different kinds of tasks. The task we employed, however, is a generic one (that can employ any content) and represents the kind of multivariable causal induction that people engage in commonly in natural contexts. Equally important, it allows for multiple different kinds of judgments to be made in response to a single problem cue—a valid judgment can be made that one variable plays a causal role, based on appropriate evidence, while at the same time causal or noncausal judgments are made regarding other variables that are invalid due to lack of the necessary evidence. It is this characteristic that has allowed us to examine the question of how the propensity to make one kind of response (valid inference) is related to the propensity to make another kind of response (invalid inference).

If the two response modes we have identified are independent, at a minimum we need a model in which their distinct functioning is represented. Rather than increasing strength of one mode, in a given problem context, in any way *causing* decreasing strength of the other (or decreasing strength of one causing increasing strength of the other), two distinct challenges must be represented. One involves constructing, accessing, and implementing one mode. The other involves gaining awareness of, monitoring, and inhibiting the other mode when it is inappropriate.

Our results are consistent with growing attention in the study of cognitive development to the role of response inhibition (Harnishfeger & Bjorklund, 1993; Kuhn, 2006; Kuhn & Franklin, 2006; Williams, Ponesse, Schacher, Logan, & Tannock, 1999). Traditionally, the focus in cognitive development research has been on the attainment of new forms

of cognition. Recognition based on microgenetic work of the coexistence of multiple forms highlights the need to gain control of and relinquish the less sophisticated or adaptive mode of operation, as well as to attain and consolidate the more advanced form—two distinct kinds of change, both of which are facilitated by practice (Brace, Morton, & Munakata, 2006; Kuhn et al., 1995; Kuhn & Franklin, 2006; Kuhn & Pease, 2006; Siegler, 2006). But engagement and practice by themselves are not sufficient. A model that incorporates the dual challenges of production and inhibition requires a metalevel operator distinct from operations that occur at the performance level (Kuhn, 2001). Constructing, implementing, and monitoring the more advanced operation is a distinct task from inhibiting the less advanced response. Each of these tasks, we would argue, requires a metalevel operator that governs the performance operators. If so, further specifying the nature of this metalevel operation becomes an important objective.

Our findings are also relevant to the growing literature in cognitive psychology on dual-process systems (Evans, 2003; Evans & Over, 1996; Sloman, 1996). The two kinds of judgments that our task yields may not map perfectly onto the theoretical constructs of heuristic and analytic processing modes. In particular, a small proportion of responses we classified as invalid judgments arguably might have entailed some degree of analytic processing that went astray and failed to yield a valid judgment. (The reverse error, classifying as valid a judgment that was produced heuristically, is highly unlikely.) Broadly, however, the production of valid judgments can be hypothesized to require an analytic operator, and the inhibition of invalid judgments can be hypothesized to not require an analytic operator and to arise from a heuristic system.

In reviewing the dual-processing literature, Evans (2003) emphasizes the need to better understand how the two systems interact. Several authors have addressed the question at a theoretical level. Taking the position that the two are closely linked, Klaczynski (2001, 2004, 2005), for example, proposes that the analytic system serves two functions. It does the cognitive work necessary to generate and execute the higher order response and in addition it inhibits the alternative heuristic response. Stanovich (1999, 2004), in contrast, subscribes to the alternative possibility that production of an analytic response does not increase or decrease the probability of an additional heuristic response to the same situation, and, similarly, a heuristic response does not affect the probability of an additional analytic response. He describes the heuristic systems as “not under the control of the analytic processing system” (2004, p. 37) and able to “sometimes execute and provide outputs that are in conflict with the results of a simultaneous computation being carried out by analytic processing” (p. 37), although he does later note that the analytic sys-

tem is capable in certain situations of “overriding” the heuristic system. Given the modest amount of empirical evidence that has been brought to bear on dual-systems models relative to the theoretical interest they have engendered, the question we have identified—whether a formulation like Klaczynski’s, in which the analytic system controls and inhibits the heuristic system, or one like Stanovich’s, in which the two systems are largely independent, is more correct—thus seems a fundamental one to address via empirical investigation. The work described here is one such example and one that clearly favors one alternative over the other.

Another aspect of the present work that warrants note is its social dimension. Cognition is fundamentally and most often a social activity that takes place in a social rather than an isolated context and is not only influenced by but indeed constructed within this context. In the educational literature, the benefits of “cooperative learning”—which means essentially having children work in small groups—has long been regarded as a beneficial practice, despite the only modest amount of research evidence available regarding how students interact in such groups and what kinds of cognitive processes, beneficial or not, are involved (Damon, 1984; Damon & Phelps, 1989; Dimant & Bearison, 1991; Resnick, Levine, & Teasley, 1991; Resnick & Nelson-LeGall, 1997). As noted earlier, we had participants work in pairs because of its presumed facilitative effects, rather than to study the social process per se, and also because this social context better resembles the natural one in which cognition develops. In the specific case of the task employed here, however, we do have evidence of the superior progress made in a pair versus solitary condition. In an earlier study, students worked simultaneously over a period of months on one content version of the task alone and on another content version with a partner; intraindividual comparisons showed the majority of participants making more progress on the task they were engaged in with a partner (Kuhn, 2001).

Although much more evidence is needed, the present results can be taken as good news in the sense that a partner appeared to influence a child to function at a higher level more often than the partner influenced the child to function at a lower level. Although we did not observe partners’ social interaction itself except anecdotally, the better idea appears to have more often won out. Moreover, our results suggested that partners had an important influence on the second of the two processes postulated in the dual-process model and the one that in general has received much less attention: inhibiting the less effective mode of functioning (and hence, as we have noted, serving as further evidence of independent operation of the two processes). Although social process data must be examined more extensively before broad conclusions can be drawn, our findings suggest that the more valuable influence of a social context on thinking may lie

less in invoking new ideas than it does in making evident the weaknesses of existing ones.

Finally, we need to put the findings we have described in an educational context. Two broad implications warrant noting. First, strategy development is more than a simple matter of acquiring expertise. There are now many examples in the literature of students who acquire strategic skill that does not benefit them. For many, sometimes multiple reasons, they do not utilize the skill they have acquired. A metalevel manager must be invoked and investigated, as it is at this level that performance is determined. As we have undertaken to illustrate here, this manager must monitor and control multiple potential actions, not just one. Second, in real-life educational settings, these processes do not take place in a vacuum. There are external, as well as internal, influences on a student's strategy production, as well as strategy inhibition. This is likely a good thing. It is in a rich social context of deciding what to do that deliberation over alternatives is most likely to come into play.

### NOTES

1. Specifically, the variable of captain's age (young or old) yields a distribution of outcomes, rather than a single consistent outcome. The most frequent outcome (60% of instances) is level 1 for the young captain and level 2 for the old captain. However, in 20% of instances, the young captain yields a level 0 outcome and in 20% a level 2 outcome. Similarly In 20% of cases the old captain yields a level 1 outcome and in 20% a level 3 outcome. Thus, students must generate multiple instances and compare these distributions (for young and old captain) in order to identify the effect. Comparison of only two instances may be misleading.
2. The interacting variables are snake activity and gas level. Snake activity has an effect only when gas level is heavy.
3. Included in this category are three participants (two at the pretest and one at the posttest) who showed all possible valid inferences and only a single additional invalid inference made in the context of a large number of correct judgments of indeterminacy (that the evidence was inadequate to make an inference regarding that variable). These isolated incorrect judgments, it was reasoned, could be attributed to momentary inattention on the participant's part or to a data recording error.
4. This is so even though constraints exist that dictate some degree of inverse correlation between the two values. If, on a typical occasion, an individual makes judgments about four instances, each involving five variables, there exist 20 opportunities to make a judgment. Yet in any single one of these 20 cases, the individual cannot make both a valid judgment and an invalid one.
5. Substitution of the conventional parametric Pearson  $r$  statistic does not



- change this outcome. The  $r$  coefficient reached significance neither for the pretest data nor for the posttest data.
6. The participant's level of functioning at a given session was taken as the participant's individual level at the individual assessment closest in time to this session. This was the participant's pretest level for sessions occurring during the 2 months immediately following the pretest (following which there was an extended holiday break) and the participant's posttest level for sessions occurring during the 3 months preceding and following the posttest.
  7. Although pairing was done randomly, the somewhat higher proportion of instances in which a participant works with a higher level peer can be explained by the fact that the higher functioning participants overall had slightly better attendance and also tended to complete a greater number of instances per session.
  8. For three participants in the case of valid inferences and three participants in the case of invalid inferences, the gamma statistic could not be computed because of lack of variance in one or the other variable. The sample size for these analyses is therefore 31 rather than 34.
  9. Note, this result does not contradict the finding reported above that significant associations with partner level were more frequent for valid inferences than invalid inferences. A number of factors could influence the gamma statistic of association for each participant, notably the distribution (and hence variance) of the three kinds of partner mismatch (higher, lower, equivalent). The number of times a participant was matched with each kind of partner was a matter of chance. Hence, some individuals had limited variance across the three types. What is notable about these gamma coefficients, then, is the number of participants for which they are significant, rather than those for which they are not. What the ANOVA results indicate, in contrast, is that when a partner mismatch occurred it was more likely to influence the participant in the case of invalid than valid inferences.

## REFERENCES

- Alloy, L., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, *91*, 112–149.
- Brace, J., Morton, J. B., & Munakata, Y. (2006). When actions speak louder than words: Improving children's flexibility in a card-sorting task. *Psychological Science*, *17*, 665–669.
- Damon, W. (1984). Peer education: The untapped potential. *Journal of Applied Developmental Psychology*, *5*, 331–43.
- Damon, W., & Phelps, E. (1989). Strategic uses of peer learning in children's education. In T. Berndt & A. Ladd (Eds.), *Peer relationships in child development* (pp. 135–157). New York: Wiley.
- Dimant, R., & Bearison, D. (1991). Development of formal reasoning during successive peer interactions. *Developmental Psychology*, *27*, 277–284.

- Evans, J. St. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Science*, 7, 454–459.
- Evans, J. St., & Over, D. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.
- Harnishfeger, K., & Bjorklund, D. (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. Foundations*, pp. 28–49. New York: Springer-Verlag.
- Klaczynski, P. (2001). The influence of analytic and heuristic processing on adolescent reasoning and decision making. *Child Development*, 72, 844–861.
- Klaczynski, P. (2004). A dual-process model of adolescent development: Implications for decision making, reasoning, and identity. In R. Kail (Ed.), *Advances in child development and behavior* (Vol. 31, pp. 73–123). San Diego: Academic Press.
- Klaczynski, P. (2005). Metacognition and cognitive variability: A dual-process model of decision making and its development. In J. Jacobs & P. Klaczynski (Eds.), *The development of decision making in children and adolescents*. Mahwah, NJ: Erlbaum.
- Kuhn, D. (1995). Microgenetic study of change: What has it told us? *Psychological Science*, 6, 133–139.
- Kuhn, D. (2001). Why development does (and doesn't) occur: Evidence from the domain of inductive reasoning. In R. Siegler & J. McClelland (Eds.), *Mechanisms of cognitive development: Neural and behavioral perspectives*. Mahwah, NJ: Erlbaum.
- Kuhn, D. (2006). Do cognitive changes accompany developments in the adolescent brain? *Perspectives on Psychological Science*, 1, 59–67.
- Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education*, 91, 710–726.
- Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, 16, 866–870.
- Kuhn, D., & Franklin, S. (2006). The second decade: What develops (and how)? In D. Kuhn & R. Siegler (Eds.) & (W. Damon & R. Lerner (Series Eds.)), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (6th ed., pp. 953–993). Hoboken NJ: Wiley.
- Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Society for Research in Child Development Monographs*, 60(4, Serial No. 245).
- Kuhn, D., Katz, J., & Dean, D. (2004). Developing reason. *Thinking and Reasoning*, 10, 197–219.
- Kuhn, D., & Pease, M. (2006). Do children and adults learn differently? *Journal of Cognition and Development*, 7, 279–293.
- Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction*, 26, 512–559.
- Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. In H. Reese (Ed.), *Advances in child development and behavior* (Vol. 17). New York: Academic Press.

- Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction*, 9, 285–332.
- Lehrer, R., & Schauble, L. (2006). Scientific thinking and scientific literacy: Supporting development in learning contexts. In W. Damon & R. Lerner (Series Eds.) & K. A. Renninger & I. Sigel (Vol. Eds.), *Handbook of child psychology* (6th ed., Vol. 4). Hoboken NJ: Wiley.
- Resnick, L., Levine, H., & Teasley, S. (1991). *Perspectives on socially shared cognition*. Washington, DC: American Psychological Association.
- Resnick, L., & Nelson-Le Gall, S. (1997). Socializing intelligence. In L. Smith, J. Dockrell, & P. Tomlinson (Eds.), *Piaget, Vygotsky and beyond*. London: Routledge.
- Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, 49, 31–57.
- Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, 32, 102–119.
- Siegel, S., & Castellan, J. N. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.
- Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.
- Siegler, R. (2006). Microgenetic studies of learning. In W. Damon & R. Lerner (Series Eds.) & D. Kuhn & R. Siegler (Vol. Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (6th ed., pp. <x-<x). Hoboken NJ: Wiley.
- Slooman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Stanovich, K. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Stanovich, K. (2004). *The robot's rebellion*. Chicago: University of Chicago Press.
- Williams, B., Ponesse, J., Schacher, R., Logan, G., & Tannock, R. (1999). Development of inhibitory control across the life span. *Developmental Psychology*, 35, 205–213.
- Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, 27, 172–223.