Judgment Biases in a Simulated Classroom—
A Cognitive-Environmental Approach

Klaus Fiedler, Eva Walther, Peter Freytag, and Henning Plessner
University of Heidelberg, Germany

Biases in judgments and decision making that are commonly explained in terms of the judge's prior expectancies can originate in unbiased environmental learning. Applying a recently developed sampling approach to decision making to the assessment of student achievement in a simulated classroom, we investigate systematic biases in teacher's judgment of student achievement. What might appear to reflect teacher expectancies based on students' overall ability level, gender stereotypes, or naive behavioral theories can be explained by ordinary learning rules underlying decision making, such as increasing accuracy with increasing sample size of observations in a probabilistic environment. Ability (% correct answers) and motivation levels (% raising hands) of 16 students in eight subject matters were manipulated. In Experiment 1 the judged difference between smart and poor students increased with sample size, due to participation rate or unequal teacher attention. In Experiment 2, verification biases in hypothesis testing about the relative assets of boys and girls in language and science were shown to be independent of gender stereotypes, but mediated instead by differential size of stimulus samples. Experiment 3 showed that judges' sensitivity to environmental base rates can mimic expectancy biases resulting from scripted knowledge.

Imagine a female student who shows little motivation in math, and who is known to be rather poor at other disciplines. How might the teacher's judgment processes affect the evaluation of this particular student's performance in
Would teachers tend to overestimate or underestimate her performance, relative to her objective rate of correct responses? By all intuition, any bias should be to her disadvantage. Inferences from other subject matters ought to be negative, low motivation suggests lack of interest, and common gender stereotypes hardly consider maths a female domain. Nevertheless, we will show that judgement tendencies may actually favor such a student.

The aim of the present research was to open new perspectives on prominent issues of judgment and decision under uncertainty. Our focus is on judgment tendencies that have been traditionally interpreted as reflecting systematically biased cognitive processes, using teachers’ assessment of student achievement as a sensible applied domain (cf. Jussim, 1989). Revisiting the issue of teacher expectancy biases within a cognitive-ecological framework, we try to convey the insight that biased judgments need not reflect biased processes. To be sure, when teachers’ subjective estimates deviate systematically from objectively manipulated performance rates, it is appropriate to refer to a “bias” as distinguished from an unsystematic “error.” However, whereas the surplus meaning of the term “judgment bias” entails the assumption that distorted cognitive or motivated processes can be blamed, we highlight the notion that systematic judgment tendencies can result from unbiased processes, as a normal by-product of adaptive behavior in a probabilistic world. Even when no wishful thinking, prejudice, faulty expectation, or resource limitation is involved, judgment biases will occur as an inevitable consequence of the stimulus environment to which judges are exposed.

The dialectic notion of biases without biased processes is not as hard to understand as it might first appear. To illustrate, hardly anybody would question that judgments of the Self deviate systematically from judgments of other people. Self-referent judgments are more differentiated (Sande, Goethals, & Radloff, 1988), less polarized (Linville & J ones, 1980), and more sensitive to contextual information (Watson, 1982) than other-referent judgments. Now, would we predict that all self-other differences disappear when information is processed rationally and in an unbiased fashion? Certainly not. Systematic differences remain just because the samples of stimulus observations about the Self and about others are highly unequal, skewed, and selective. Because different individuals assess the same entities from different angles, relying on different stimulus samples, their judgments can diverge in systematic, predictable ways. Note, however, that “biases” within such an ecological framework must not be equated with faulty, inefficient, or illogical cognitive processes. Just as a “bias” in signal detection analysis (i.e., a flexible threshold shift, as an adaptive function of the stimulus context), a bias in evaluative judgment can reflect a sensitive ecological adaptation process.

Research Strategy

Investigating these issues in the institutionalized setting of teachers’ achievement assessment in the classroom seems to be a useful strategy. Performance assessment calls for natural judgments of probability (e.g., of correct responses)
and can be compared to clearly defined performance criteria. Moreover, judgments provide the basis for important and consequential decisions about the students' academic career. Most previous research in this area has relied on data from real school classes (cf. Jussim, Madon, & Chatman, 1994; Rosenthal & Rubin, 1978). Such a research strategy, however, while maximizing realism, may not be optimal for isolating different causal influences on teacher judgments. In our attempt to gain more experimental control, we therefore employ a different strategy, analyzing the acquisition of information about artificial classes in a simulated mini-world. Participants are asked to play the role of a teacher. Their task is to assess the performance of 16 students (eight girls and eight boys) whose ability (proportion of correct responses) and motivation (proportion of raising hand) in eight different subject matters is controlled by the computer program. The simulated classroom setting is represented graphically on the computer screen, showing 16 desks with the students' names in different locations (Fig. 1). Teachers select questions from pull-down menus, call upon one of the students who raises his/her hand, and so assess all 16 students' performance across several sessions. Photographs of students are available in addition to their names to support the complex impression formation task. Resulting judgments of ability, motivation, and other aspects are analyzed as a function of the students' true parameters—and other theoretically relevant variables.

Such a simulation approach has several advantages. Prior knowledge can be ruled out and the entire process of judgment formation can be studied from the very first to the last observation, along with a computerized record of every stimulus detail. The assignment of sitting positions and performance

![What are the characteristics of a fable?](Fig. 1. Screen-shot from the computer program used in all experiments.)
parameters to individual students can be counterbalanced, and the relations
between student gender, performance, and gender-typed subject matters (lan-
guage vs science) can be controlled. Yet, despite this high degree of experimental
control, the task setting is experienced as rich and seminaturalistic. Participants report to form “real” personal relationships with “their” students. Most importantly, the interactive dynamics of a simulated classroom afford an opportu-
nity to operationalize learning-based decision in a truly cognitive–ecological
framework. As several scholars have emphasized, in order to understand cogni-
tive processes within the individual, it is essential to describe the stimulus
environment impinging on the individual (Brunswik, 1955; Gibson, 1979; Giger-

Types of Expectancy Biases

The literature on achievement assessment is characterized by an ongoing
debate between the realist position emphasizing basic accuracy (e.g., Jussim,
1989, 1991) and the constructivist position that teacher expectations often
override real performance (e.g., Rosenthal & Jacobson, 1968; Wineburg, 1987).
As we shall see, both positions can be reconciled within a cognitive–ecological
approach (cf. Jussim & Eccles, 1992); biases may arise in spite of basically
accurate processes.

Where judgments have been shown to deviate systematically from objective
performance measures, these biases are commonly attributed to preexisting
expectancies. Thus, in explaining the most prominent phenomenon of self-
fulfilling prophecies (Crano & Mellon, 1978; Jussim, Eccles, & Madon, 1996;
Kukla, 1993; Rosenthal, 1991; Rosenthal & Jacobson, 1968; Rosenthal & Rubin,
1978), it is assumed that the teacher’s prior expectations solicit expected behav-
ior in the target person (student) that finally fulfills and justifies the initial
expectancy. Less popular is the notion of self-defeating prophecies (Kukla, 1993)
in which expectations solicit disconfirming or reactant behavior. Even when
the target’s objective behavior remains unchanged, expectancies may distort
the judge’s subjective perception of the target’s behavior (Darley & Fazio, 1980;
Fiedler, Walther, & Nickel, 1999; Miller & Turnbull, 1986; Plessner, 1999;
Snyder, 1984). By default, theoretical explanations of teacher judgment biases
in academic settings refer to expectancies as a key theoretical construct (cf.
Jussim et al., 1994), just as in other domains of expert judgment (e.g., Anderson
& Kellam, 1992; Dowling & Graham, 1976).

More specifically, three different types of influence of expectancies or prior
knowledge on decision making can be distinguished: (a) overgeneralization of
prior experience with decision targets; (b) naïve theories of decision target’s
behavior or functioning; and (c) stereotypes associated with decision targets.
With students as targets, these expectancy types can be illustrated as follows.

Overgeneralization. Any preliminary classification of students as smart
versus poor, based on an early stage of learning, may be overgeneralized in
subsequent observations, as in the classical halo-effect (Cooper, 1981). Initial
observations may create expectancies that bias subsequent assessment. Students who are initially recognized as generally smart may then be treated more positively and may be perceived and judged more favorably than students initially classified as poor.

Naïve theories of student behavior. Teachers' inferences may be based on heuristic cues. As ability is harder to observe than motivation, teachers may infer ability from motivational cues (e.g., Brattesani, Weinstein, & Marshall, 1984). Given a class of 16 students of whom only one can answer each question, there are 16 times more opportunities to observe motivation (who raises his/her hand) than ability (answering a question). Thus, following a common naïve theory, teachers may assume that students who often raise their hands a lot know more than students who remain passive most of the time.

Social stereotypes. Last not least, expectancies may derive from stereotypical knowledge of the social categories to which individual students belong. Gender stereotypes, in particular, afford a common source of achievement-related expectations (e.g., Jussim, 1989), aside from socioeconomic class (Rosenthal & Jacobson, 1968). Gender-typing identifies maths and science as disciplines in which boys outperform girls, whereas girls are expected to mostly exhibit their relative assets in arts and language. Due to these stereotypical associations, the same performance in maths and science may appear more intelligent and competent when exhibited by boys than girls, and the same responses in arts or language may appear more favorable in girls than boys. Similarly, teachers may selectively remember student performance that is consistent with prevailing gender stereotypes.

Previous research has been quite creative in isolating various processes through which prior expectancies can be confirmed in social interaction— from lop-sided information search (Mynatt, Doherty, & Tweney, 1977; Schulz-Hardt, Frey, Moscovici, & Lüthgens, 2000; Snyder, 1984) and conversational rules (Zuckerman, Knee, Hodgins, & Miyake, 1995) to the differential diagnosticity of expected versus unexpected hypothesis tests (Trope & Thompson, 1997). In all these paradigms, however, prior expectancies are considered the ultimate source of judgment bias. Hardly any theoretical account has considered the possibility that biased judgments may originate in unbiased stimulus learning, quite independent of the judges' prior expectations.

A COGNITIVE-ECOLOGICAL SAMPLING APPROACH

We do not question the potential significance of genuine expectancy effects originating in stereotypes, naïve theories, or overgeneralized distinctions. However, we point out an alternative and equally intriguing source of biases that are basically stimulus-driven rather than expectancy-driven. This alternative approach derives from a recently developed sampling approach to decision making under uncertainty (Fiedler, 2000; Fiedler, Brinkmann, Betsch, & Wild, 2000). The basic tenet of this approach is that judgments are often rather accurate relative to the actually observed stimulus samples. However, exactly
because judgments are so sensitive to inductive learning, any bias in the sample of ecological data gathered from an individual’s perspective will carry over to subsequent judgments.

A comprehensive overview of the sampling approach and its psychological implications is given elsewhere (Fiedler, 2000; Fiedler et al., 2000). Suffice it here to delineate the sampling framework as it applies to the present task setting. We start from the Brunswikian assumption that student ability, just like other distal entities, is not amenable to immediate perception. Ability must be inferred or construed from whatever sampled data happen to be available. For convenience, ability might be conceived as a population parameter, that is, the proportion of correct responses in the (unlimited) population of all responses from that student. To “perceive” this latent attribute, an estimate must be based on a limited sample of observations. Samples provide the interface between the distal environment and cognitive judgment processes, whether samples are gathered in the external world or in memory (Hastie & Park, 1986; Juslin & Olsson, 1997; Tversky & Kahneman, 1973).

Let us distinguish the predictor and the criterion of a judgment problem, analogous to an experimental design. Judging student ability means to consider different students as a predictor, or independent variable, and ability as a criterion, or dependent variable. Neither predictor information nor criterion information is usually sampled randomly. On the predictor side, teachers gather for various reasons more information about certain students than about others. On the criterion side, too, high ability and low ability are not equally likely to be observed. Predictor sampling as well as criterion sampling processes can thereby lead to systematic biases. For instance, when latent ability is rather low (e.g., at a level of 0.20 correct responses), sampling may obscure the low population base rate because the students raise their hands only when they know the correct answers, or actively evade the teacher’s attention when they do not know the correct answer. As a consequence of such a criterion sampling effect, the effective stimulus sample may seriously overestimate the “true” latent ability. When exposed to such a biased sample of observations, teachers should overestimate ability. Note that in this scenario, judgments inflate low base rates not because they neglect base rates but, on the contrary, because they are sensitive to actually observed base rates. In fact, the bias that results from being exposed to an ecology that exhibits some data and conceals other will increase with the accuracy of judgments, relative to the effective stimulus input. Experimental evidence for rather strong biases of that kind comes from recent studies on diagnostic decision making (Fiedler et al., 2000), showing that the stimulus samples drawn from the universe were biased to a greater extent, and actually justified even stronger biases, than those observed in the resulting judgments. Thus, one of the distinct and theoretically challenging implications of the sampling approach is that inflated probability judgments of low base rate events may not reflect a base rate neglect (Borgida & Brekke, 1981; Kahneman & Tversky, 1972) but the judges’ sensitivity to unusual base rates in the sample. Such an alternative account of the base rate fallacy even
holds for judgments based on naturally observed event frequencies (cf. Gigerenzer & Hoffrage, 1995).

Another intriguing implication is that variations in predictor sample size alone can lead to systematic over- and underestimations, even when samples are not biased toward specific criterion values. Thus, when there are clearly unequal numbers of observations about different students, the superior memory strength and reliability of large as opposed to small samples is sufficient to produce systematic effects. To illustrate this point, consider two students, A and B, with exactly the same latent ability of say, 0.80 correct answers (in the population of all possible answers). Assuming an unbiased teacher, both correct and incorrect student answers have the same chance of being noticed, encoded, and evaluated by the same fair criteria. However, a larger sample is available on student A than B because B has been absent for some time. The expected distribution of correct and incorrect answers (given 80% ability) could thus be $12/1$ (correct answers) and $3/2$ (wrong answers) for A compared with $4/1$ and $1/2$ for B. Although both samples reflect the same distal ability parameter, the $12/1$, $3/2$ sample is psychologically more “significant” and more likely to lead to favorable evaluations than the smaller $4/1$, $1/2$ sample. Teachers will assess the actually existing ability level (i.e., the high rate of correct responses) more reliably when the sample is large, giving rise to more accentuated, less regressive judgments of student A than B (see Fiedler, Walther, & Nickel, 1999; Fiedler, Kemmelmeier, & Freytag, 1999; Kaplan, 1981; Shavitt & Sanbonmatsu, 1999).

Just as judges cannot be blamed for biased criterion samples (higher visibility of certain outcomes), stronger inferences drawn from large (as opposed to small) predictor samples need not reflect irrational processes or cognitive deficits. Small samples are less reliable and less likely to reveal the properties that show up clearly in large samples. In environmental learning, a fully normal, adaptive rule is that behavioral responses increase with the amount of stimulus information (i.e., the number of trials). Accordingly, even when the ability of students A and B is equally high, 15 observations on A justifies stronger inferences in a probabilistic world than only 5 observations on B. Normative information-processing devices (such as the Bayes theorem or a t test) are susceptible to this kind of environmentally generated bias.

Thus, variations in sample size alone can create a biased picture of the distal world. As we move through the social environment, we encounter targets representing high and low levels on various attribute dimensions. The sensitivity with which these upward and downward peaks are recognized depends on sample size. The number and density of available observations may vary dramatically across the social environment, and so does our ability to “see” latently existing information. In the classroom, unequal samples can reflect manifold factors, such as the students’ presence, degree of participation, sitting position, physical appearance and salience, and reaction speed of their contributions, and the teacher’s relationship to students, his/her expectations and hypotheses to be tested, his/her selective memory, and many other factors. Even when teachers might be blamed for some of these factors, they are completely
“innocent” for other. In either case, the effect is the same. Completely unbiased information processing devices, such as connectionist judgment algorithms (Dougherty, Gettys, & Ogden; 1999; Fiedler, Kemmelmeier, & Freytag, 1999), will produce biased judgments when fed with unequal stimulus samples.

OVERVIEW OF PRESENT RESEARCH

In the experiments to be reported, we offer revisited accounts for all three types of expectancy biases in terms of predictor sampling or criterion sampling processes. In Experiment 1, we first demonstrate that a universal regression effect works against the overgeneralization of smart versus poor students. True performance differences are generally underestimated, due to imperfect environmental learning, the implication being that smart students are underestimated and poor students are overestimated. However, this generalized regression effect, which overrides a halo effect in favor of expectedly smart students, is then moderated by a predictor sampling effect. As samples of observations increase, regression should decrease so that the judged ability of smart students should increase whereas the judged ability of poor students should decrease.

While Experiment 1 speaks to expectancies of the overgeneralization type, Experiment 2 is concerned with expectancies derived from stereotypes. Rather than leaving it up to teachers to gather information in an arbitrary fashion, teachers are asked to test specific hypotheses about the assets of boys and girls in different subject matters. When the task is to test stereotype-consistent hypotheses (i.e., that boys are good in science and that girls have their assets in language), teachers’ positive testing strategies (Fiedler, Walther, & Nickel, 1999; Klayman & Ha, 1987) should produce larger samples for boys in science and for girls in language lessons than vice versa. Only when such positive-testing occurs, differential predictor sampling should tend to confirm the gender stereotype. Holding ability constant, girls’ talents should be more visible in language and boys’ assets should be more apparent in science. However, crucially, when teachers are asked to test counter-stereotypical hypotheses, positive testing focuses on girls in science and on boys in language, resulting in counter-stereotypical judgment biases of the same magnitude. Experiment 2 not only offers an alternative view on stereotyping but also a distinctive account of positive testing effect (Klayman & Ha, 1987).

Finally, the focus of Experiment 3 is on expectancies based on scripted naive theories, with judgment biases deriving from criterion sampling. Ability of students is held constant and only motivation (rate of raising hands) is allowed to vary. In this situation, when teachers are nevertheless forced to judge ability, they will very likely use scripted knowledge, inferring complete performance (correct response) from incomplete observations (raising hands). However, we show that this inference is sensitive to environmental base rates. When criterion sampling produces constantly high ability feedback (i.e., raising hands is mostly followed by correct responses), high motivation is used to infer high ability. In contrast, in a constant low ability environment, raising hand is often paired with failure so that motivation should be used as a proxy for low ability.
Thus, the utilization of scripted expectancies is not guided by a blind heuristic but by sensitive environmental learning.

To demonstrate these biases resulting from adaptive learning rather than biased, expectancy-driven processes, we intentionally create demanding task conditions, calling for multiple judgments of different students in various disciplines based on a few learning sessions. Such high cognitive load ought to facilitate the sort of heuristic shortcuts that are commonly assumed to induce expectancy-based guessing. Being unable (or unmotivated) to effectively encode all the abundant information input, teachers might resort to quick and easy (not to say “dirty”) heuristics. Judgments under high uncertainty might thus be misled by simplifying generalizations, stereotypes, or scripted theories.

We do not pretend that expectancy biases of this kind do not exist. However, as mentioned above, we advance the alternative, complementary assumption that biases are not confined to cognitive misers but may even result from unbiased processes, driven by high accuracy motivation (Chaiken, Liberman, & Eagly, 1989). Based on considerable evidence for the accuracy of inductive-statistic assessment (Greene, 1984; Hasher & Zacks, 1984), especially when stimuli can be assessed in natural frequency format (Gavanski & Hui, 1992; Gigerenzer, 1991; Gigerenzer & Hoffrage, 1995), and based on our own recent evidence (Fiedler et al., 2000), we assume that well-motivated teachers should also exhibit this basic accuracy (cf. Jussim & Eccles, 1992). They should not only succeed in assessing the average performance level of individual students but also in making finer distinctions within students between different disciplines, and not confuse motivation and ability. However, this basic accuracy (Jussim, 1989, 1991) holds relative to the samples observed, as distinguished from the latent parameters of the environment (i.e., population). Sampling variation should leave sufficient latitude for the occurrence of systematic biases. The very accuracy with which judgments reflect properties of the effective stimulus samples should thus produce systematic deviations from the latent parameters.

**EXPERIMENT 1**

The first experiment is described in greater detail than subsequent ones, which follow the same general method. To repeat, participants took the role of a teacher of 16 students (sixth-graders) over up to eight lessons. After each lesson, they rated each student’s ability, motivation, and potential ability. The actual differences in motivation and ability between students and disciplines were manipulated (see below). We suspected that judgment accuracy would be generally high, in spite of the complexity of the task. Nevertheless, systematic biases should result from sampling effects. Experimental variables that affect sample size—such as student participation rate, focus of hypothesis tests, and number of inference cues—should thus produce systematic biases, even when the translation process from samples to judgments is accurate.
Method

Overview and Design

Each student in the stimulated school class represented a unique cell of a within-judge design involving the factors student sex (boy vs girl), ability level (high vs low) or motivation level (high vs low), and congruency (high vs mixed vs low) with gender stereotypes regarding achievement in languages (English or German, stereotypical asset of girls) versus sciences (Maths or Physics, stereotypical asset of boys). This was accomplished by the systematic variation of an ability parameter (probability of giving correct answers) and a motivation parameter (probability of raising hands). The parameter matrix (Table 1) assigns particular ability (or motivation) levels to the eight male and eight female students. (Capital letters A and M refer to ability and motivation, respectively). In addition to general performance differences between smart students ($0.5 < A < 0.8$) and poor students ($0.2 < A < 0.5$) students, the parameter matrix also reveals variation within male and female students across subject matters (language vs science), providing the basis for manipulating stereotype congruency.

The only between-judge factor referred to the source of performance variation, which was either A or M. For one group of teachers, students differed only in A, as described in Table 1, while all M parameters remained constant.

| TABLE 1 |
| Parameter Matrix Describing the Ability (Probability of Correct Responses) or Motivation (Participation Rate) of the 16 Target Students in the Experiment 1 |

<table>
<thead>
<tr>
<th>Discipline type</th>
<th>German</th>
<th>English</th>
<th>Mathematics</th>
<th>Physics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>Lesson 1</td>
<td>Lesson 2</td>
<td>Lesson 1</td>
<td>Lesson 2</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>2</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td>3</td>
<td>.5</td>
<td>.5</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td>4</td>
<td>.8</td>
<td>.8</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>5</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td>6</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>7</td>
<td>.2</td>
<td>.2</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>8</td>
<td>.5</td>
<td>.5</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td>10</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>11</td>
<td>.8</td>
<td>.8</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>12</td>
<td>.5</td>
<td>.5</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td>13</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>14</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td>15</td>
<td>.5</td>
<td>.5</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td>16</td>
<td>.2</td>
<td>.2</td>
<td>.5</td>
<td>.5</td>
</tr>
</tbody>
</table>
at $M = 0.5$. For example, the second matrix row pertains to a boy with generally high ability ($0.5 < A < 0.8$, i.e., 50 to 80% correct answers). His ability pattern is incongruent with the gender stereotype (better in language than in science), but semantically consistent in that it varies only between but never within disciplines types (language and science). The motivation of this student remains invariant ($M = 0.5$). For the other group of teachers, in contrast, the same differences pertain to motivation ($M$ as in Table 1), whereas ability is held constant ($A = 0.5$).

Participants

A total of 30 male and female native speakers (most of them students at the University of Heidelberg) participated for payment (DEM 75.00). An equal number of participants were randomly assigned to both groups of the source of variation factor ($A$ variable vs $M$ variable).

Procedure

Participants were received individually. Written instructions made them familiar with the general setting and with the basic features of the computer program. Participants were asked to take the role of a teacher who is about to take over a class of sixth-graders. They were informed that there would be eight lessons (two each in the languages English and German, and the sciences Physics and Maths), and that they would be asked to rate all students individually after each lesson. Participants learned that they would teach two lessons in each of the first four experimental sessions.

Instructions also described the pull-down menus employed for the selection of questions. Teachers learned that they could call upon only those students who actually raised their hands, and that feedback would be given about whether an answer was right or wrong. An opportunity was given to ask questions about the general procedure, before the experimenter started the first session.

The computer program started with a description of all submenus available at the start of every experimental session. These submenus comprised re-reading the extended version of the general instruction, looking at photographs of individual students, inspecting today's schedule of lessons, and scanning the item pools of specific lessons (each consisting of 20 items). Items had been taken from regular sixth-graders' textbooks (plus some items generated by the authors themselves) and were of apparently equal difficulty, according to a pretest. Participants were reminded that they would be asked to judge ability and motivation after each lesson. Afterward participants spent 10 to 15 min looking through item pools and photographs, before they decided to start the first lesson.

Giving lessons. In each experimental session, participants gave lessons in one subject out of each of the two discipline types, sciences and languages. The following constant ordering of lessons was used: German/fables and Physics/
heat (session 1), Maths/set theory and English/grammar (session 2), Maths/measurement units and German/orthography (session 3), and English/speech exercises and Physics/electricity (session 4).

The lessons subroutine involved participants in a sequence of question-asking trials and a final performance judgment. Every trial began with the complete menu of the 20 items available for that lesson. Teachers selected one item per trial by moving the cursor up and down the item list and striking the return key when the chosen item was highlighted. The computer then displayed the classroom, showing 16 students sitting in different locations, with rectangles symbolizing the individual students' desk. Those students that raised their hands in response to the selected question were marked (cf. Fig. 1). It was only from these students, selected by a random generator at a rate specified by the parameter M, that a response could be solicited. Teachers selected one student by moving the cursor onto the student's name and then striking the return key. The student's desk was immediately replaced by a box with one of two possible captions: either the word "right" or the word "wrong" blinking for three seconds (cf. Fig. 1), indicating whether the student had delivered a correct answer (randomly generated with probability A). After the feedback, the computer returned to the item menu and the teacher selected the next question. Teachers could ask the same question repeatedly if they wanted to compare several students' ability. The program allowed for different subsets of students raising their hands when the same question was asked repeatedly. To forestall inconsistent response patterns, repeated answers of the same student to the same item yielded the same feedback. The program kept participants in the ask-and-call-upon loop for 20 min and then switched to the dependent measures irrespective of how many trials a participant had run. The whole session was registered and stored for later data analyses, including the temporal order with which specific questions had been asked, the choice latencies, the student that had been called upon, and whether the answer was correct.

Manipulation of students' behavior. As evident from Table 1 above, performance (either A or M, dependent on the experimental group) varied over students as well as lessons. The experimental variables were defined in terms of these parameters. The students' general level of ability/motivation was either high (students 1–4 and 9–12, with parameters in the range of 0.8 to 0.5) or low (students 5–8 and 13–16, with parameters in the range (0.2 to 0.5). Each student's performance pattern across lessons could be classified by stereotype congruency, as congruent (students 1, 5, 9, 13), incongruent (students 2, 6, 10, 14), or mixed (students 3, 4, 7, 8, 11, 12, 15, 16). As already mentioned, the resulting student profiles could be semantically consistent with the similarity structure of disciplines (equal performance within language and science and differences only between; students 1, 2, 5, 6, 9, 10, 13, 14) or inconsistent (variation within language and science; students 3, 4, 7, 8, 11, 12, 15, 16). Note that only semantically consistent patterns can be congruent or incongruent with the gender stereotype, whereas inconsistent students are stereotypically mixed. The allocation of student names to sitting positions (desks) was held
constant (as in Fig. 1) but the assignment of parameter conditions (matrix rows) to the eight boys and the eight girls was randomized for each teacher.

Dependent measures. At the end of each lesson, teachers rated all students for ability and—in order to differentiate between various aspects of performance—also for motivation in that lesson and for their expectations about performance (potential ability). To reduce same-response tendencies across the three ratings, participants first rated all students for ability, then in a second run for motivation, and finally for potential ability. All ratings were delivered on 50-point graphical ratings, in response to the prompt “Please estimate, for all students, the percentage of all answers in this lesson that were correct” (ability), “... the percentage of questions when a student had raised his/her hand?” (motivation), and “... what percentage of correct answers a student would have given even when he/she did not raise his/her hand?” (potential ability). The sitting position determined the order of student ratings, from the upper left to the lower right. Thus, the design confounds all four principally uncontrollable factors, sitting position, name, appearance, and rating position. However, as assignment of parameter values to individual students was randomized, this should not undermine the results.

When participants had completed all dependent measures, they were interviewed carefully to see whether they had any suspicion about the “real objective” of the experiment and to gather demographical data. Finally, participants were thanked, debriefed, and paid.

Results

We confine the reported analyses to the two main ratings of ability and motivation; the third rating, concerning potential ability, was largely redundant with the ability ratings and did not add any further evidence. Preliminary analyses revealed no differences as a function of participant gender; all analyses are therefore pooled over this factor.

Percentage judgments on the graphical 50-step scale were transformed, for convenience, to the same proportion scale (between 0 and 1) as the parameter matrices. Data analyses can be conducted with teachers or students as the unit of analysis. Thus, one can enter judgments of all 16 students, averaged over all teachers (of the same condition), and then treat these student scores as a random variable. Alternatively, one can use teachers as the unit of analysis, averaging first over subgroups of (e.g., high vs low level of ability/motivation; male vs female) students, treating teachers’ scores as a random variable. We will sometimes switch between both modes of analysis. For instance, to evaluate the accuracy of the judgments, it is natural to use students as the unit of analysis, because the distal parameters pertain to students. In contrast, the total group of 16 students is too small to allow for multiple splits by level, sex, stereotype congruency, and semantic consistency. Such analyses are therefore based on individual teacher's scores for various types of students.
Basic Accuracy and Fairness

For a first overall test of judgment accuracy, the mean judgments (pooled over teachers) of the 16 students were analyzed in a three-factorial analysis of variance (ANOVA) involving performance level (A/M high vs low) as a between-student factor and discipline types (language vs science) and source of variation (A-variable vs M-variable) as within-student factors. To obtain the discipline type factor, judgments were averaged across all four language lessons (two German and two English) and across all four science lessons (two Physics and two Maths). The bar chart in Fig. 2 gives the mean ratings as a function of these factors, with regard to ability and motivation.

The ANOVA for ability (A) judgments yields a significant main effect for performance levels, $F(1, 14) = 25.52, p < .001$, reflecting markedly higher A judgments for the eight high-level students than for the eight low-level students. This main effect is qualified by a levels $\times$ source of variation interaction, $F(1, 14) = 9.41, p < .01$. Judged A differences between smart and poor students are correctly confined to the A-variable condition. Thus, judges can differentiate clearly between variation in A and M. Separate two-way ANOVAs confirm that the levels main effect is apparent for the A-variable condition, $F(1, 14) = 42.08, p < .001$, but not for the M-variable condition, $F(1, 14) = 1.02$.

The same ANOVA performed on motivation (M) ratings also yields a motivation levels main effect, $F(1, 14) = 77.19, p < .001$, and a levels $\times$ source of variation interaction, $F(1, 14) = 40.62, p < .001$, restricting the levels main effect...
effect to the condition in which the variation is actually due to M. In separate ANOVAs, the levels main effect is obtained only when the variation is in M, $F(1, 14) = 73.32, p < .001$, as opposed to A, $F(1, 14) = 0.05$. Two minor, unpredicted effects emerge for M judgments, a source of variation main effect, $F(1, 14) = 5.57, p < .05$, and a discipline type $\times$ source of variation interaction, $F(1, 14) = 5.92, p < .05$, indicating somewhat enhanced ratings in science for constant M.

Regressiveness

In these overall analyses, evaluations of individual students are quite sensitive to actually existing differences (Fig. 2). Nevertheless, judgments do deviate systematically from the objective parameters due to a pervasive regression effect—a typical sign of incomplete environmental learning. Actually high ability (0.8) is underestimated and actually low ability (0.2) is overestimated, with no systematic deviation for intermediate (0.5) ability (see Fig. 3). Consistent with normal learning principles, regression is strongest in the first session and decreases with learning experience (i.e., with growing samples), as evident from Fig. 3. This pattern is not in line with a perpetuating expectancy effect based on overgeneralizations of the initial classification of smart versus poor students. Rather, smart students suffer and poor students profit from the pervasive regression effect.

Reliability

The gradual learning of distal ability parameters over growing samples of observations is also apparent in measures of interjudge reliability. Treating the 15 judges in each condition like items of a test, it is possible to compute the internal consistency with which different judges converge in discriminating between students (Rosenthal, 1987). For A ratings, the effective interjudge reliability starts at a level of 0.26 and 0.42 in the first two lessons, and their

![FIG. 3. Mean judgments of ability in Experiment 1 by parameter A and session.](image)
systematic agreement raises to 0.88 and 0.79 in the last two lessons (in the A-variable condition). Interestingly, M ratings are rather reliable from the beginning, starting from 0.67 and 0.82 and ending at 0.85 and 0.73. This high initial level reflects the larger data base and higher reliability of M than A judgments. All students’ motivation (i.e., whether they raise their hands) is visible for every question the teacher directs at the class, but ability can be observed only in one student per question.

Sampling Errors

Sampling effects provide a crucial mediator for understanding the deviations of teacher judgments from the latent A and M parameters. Consider first the fact that sample statistics may deviate from the population parameters. The average correlation, within individual judges across the $16 \times 8$ matrix cells, between the A parameters and the actually obtained sample statistic for each student in each lesson amounts to 0.68. A judgments in turn correlate 0.47 on average with the actually observed sample statistics. The resulting average correlation between A parameters and A judgments is only 0.35. M judgments, by comparison, correlate similarly with M parameters (0.41) and samples (0.42) because the M samples correlate almost perfectly with the M parameters (0.98), due to the larger number and reliability of M observations.

Sample-Size Effects

Our major predictions relate judgment biases to variations in sample size. The regressive tendency to underestimate high performance and to overestimate low performance should decrease with increasing sample size. Of two equally smart students, the one with the larger sample of observations should be judged smarter. In contrast, of two equally poor students, the one with the larger sample should appear even poorer. Sample size should have little influence on judgments of intermediate students. Note that this prediction of enhanced discrimination with increasing amount of information (i.e., sample size) reflects a genuine learning effect and cannot be attributed to heuristic guessing, which should be maximal under high uncertainty (i.e., small samples).

The sample-size argument amounts to testing a three-way interaction, involving ability levels (high vs low), sample size (large vs small), and an informativeness factor that contrasts those cells of the parameter matrix where the student’s high or low performance is actually manifested (parameter $\neq 0.5$) to those cells where performance differences among students are invisible (parameter $= 0.5$). The sampling account—as opposed to a common expectancy account—predicts that teacher ratings should discriminate between smart and poor students only where differences actually exist, and only when samples are large enough.

Accordingly, we calculated four repeated measures for each student: (a) across all lessons in which this student had an informative parameter and across all
teachers who drew a large sample on this student, (b) across informative lessons and across teachers drawing small samples; (c) across uninformative lessons and teachers drawing large samples; and (d) across uninformative lessons and teachers drawing small samples. In this analysis, “large” and “small” samples are defined as above versus below the individual teachers’ median sample size. Results are summarized in Fig. 4.

As predicted, existing A differences (in the A-variable condition) are extracted more readily when samples are large rather than small. In the ANOVA of A ratings, a main effect for levels, $F(1, 14) = 48.72$, $p < .001$, replicates the clear discrimination between smart and poor students. Two-way interactions indicate that this main effect is confined to those informative lessons where high versus low ability is actually manifested (in $A > .5$), levels $\times$ informativeness, $F(1, 14) = 68.97$, $p < .001$, and where samples are large rather than small, levels $\times$ sample size, $F(1, 14) = 42.36$, $p < .001$. Most importantly, the crucial three-way interaction, $F(1, 14) = 27.49$, $p < .001$, confirms that smart and poor students are properly discriminated only when samples are large and informative.

The same ANOVA performed for M judgments, given a variable M, yields a strong levels main effect, $F(1, 14) = 112.46$, $p < .001$, and a levels $\times$ informativeness interaction, $F(1, 14) = 82.54$, $p < .001$, indicating that judged M differences are largely confined to those lessons where M differences actually exist. Notably, the M levels $\times$ sample-size interaction is also significant, $F(1, 14) = 7.33$, $p < .05$. This means that letting students respond renders teachers more sensitive to motivational differences than merely passively observing them while they raise their hands or not.

Pragmatic Confusion

Although there was little confusion of A and M in the overall analysis (see above), a more refined analysis reveals several noteworthy interactions of motivation and ability. When only M varies, A judgments produce a levels $\times$ sample-size interaction, $F(1, 14) = 6.86$, $p < .05$. Motivated students who raise their hands more often than justified by their performance, receive lower A

![Fig. 4](image)

**Fig. 4.** Mean judgments of ability in Experiment 1 by performance level, sample size, and informativeness when $A$ is variable.
ratings than less motivated students (0.51 vs 0.54) when rich samples make the discrepancy apparent. In contrast, when samples are impoverished, judges begin to confuse motivation and ability, giving higher A judgments when M is high than low (0.52 vs 0.46). Thus the uncertainty of impoverished A samples facilitates the use of M as a proxy to A.

A sample-size main effect in M ratings, $F(1, 14) = 6.33, p < .05$, reflects another intriguing pragmatic illusion; teachers misattribute their self-selected samples to the preferred students’ motivation. That is, they believe that students at whom they have directed many questions have raised their hands more often than students they have themselves neglected. A similar sample-size main effect is obtained for M ratings in the A-variable condition, $F(1, 14) = 13.04, p < .01$. Teachers tend to attribute high M to students from which they have themselves solicited large samples.

**Stereotype Expectancies**

An analysis of the role of gender stereotypes must be confined to those eight students (with a semantically consistent performance pattern) whose performance was either congruent or incongruent with gender stereotypes (see Table 1). Consider first the A ratings in the M-variable condition. An ANOVA (using judges as unit of analysis) involving performance levels (M high vs low) × stereotype congruency (congruent vs incongruent students) yielded a levels × stereotype congruency interaction, $F(1, 30) = 7.50, p < .05$. In ratings of counter-stereotypical targets, a high (vs low) motivation level serves to decrease A ratings (0.51 vs 0.55). Apparently, when students raise their hands “too” often in the wrong domain, the modest impression caused by only 50% correct answers gets even worse. This does not hold for stereotype-congruent situations (0.53 vs 0.51), where enhanced motivation tends to be used as a cue to enhanced ability.

M judgments in the M-variable condition yield an interaction, too, $F(1, 30) = 6.06, p < .05$. Actual differences between high and low M students are extracted more readily for stereotype-congruent (0.58 vs 0.45) than stereotype-incongruent targets (0.55 vs 0.50). Together, these findings suggest that motivational differences are slightly less likely to be discounted when they are supported by common gender stereotypes.

**Discussion**

The findings reported thus far provide some initial support for the implications of the sampling approach and the underlying environmental learning framework. In the complex, cognitively demanding task setting, there was no breakdown of systematic information processing due to capacity limitations. Unlike a cognitive miser who prefers shallow, heuristic processing most of the time (Fiske & Taylor, 1991) and who is guided by top-down guessing rather than data assessment, the typical teacher in the simulated classroom assessed the actually observed data quite accurately. Despite the complexity of 128
judgments (16 students × 8 disciplines) on several dimensions, they succeeded in differentiating smart and poor students, motivated and unmotivated ones. These differences were not overgeneralized but restricted to those specific lessons where the high or low performance of students was actually manifested. The average teacher's ability ratings correlate 0.47 with the actual sample statistics to which they have been exposed, after only about 20 min experience in each lesson. Aggregated over teachers, this correlation raises to 0.77.

With regard to the first type of teacher expectancies—overgeneralizations of global differences between smart and poor students—we did not find evidence for an a priori expectancy effect of the halo type. If anything, regressive judgments created unfairness in the opposite direction. Smart students tended to be underestimated, whereas poor students received too favorable judgments. However, as regression depends on sample size (reliability), discrimination between smart and poor students was enhanced as sample size increased. Thus, instead of a global-expectancy bias in favor of students initially classified as smart, what we find is a relative decline of the disadvantage of smart students with increasing sample size. The major determinant of this judgment bias, sample size, is an environmental property rather than a biased process within the teacher, due to prior expectations. A similar conclusion holds for the influence of gender stereotypes that is generally small and visible only when samples are impoverished.

EXPERIMENT 2

Stereotypes are considered a major source of expectancies and self-fulfilling processes (cf. Brophy, 1983). For instance, gender stereotypes may suggest that boys are better than girls in maths and science whereas girls outperform boys in language and maybe arts. We do not want to contest the prior-expectancy role of stereotypes. However, the purpose of Experiment 2 is to highlight an alternative process by which gender stereotypes can produce a bias in the evaluation of boys and girls. This alternative process relies solely on different sampling behavior that may be—but need not be—solicited by stereotypical beliefs. Teachers who believe that boys dominate in science may sample more observations about boys than girls in maths, or more observations about girls than boys in language lessons. Such a focused sampling strategy alone, commonly referred to as positive testing (Klayman & Ha, 1987; Laughlin, Magley, & Shupe, 1997), should induce a systematic bias on subsequent judgments, according to the sample-size principle introduced in Experiment 1. Large samples should enhance reliability and attenuate the basic regression effect, facilitating the discrimination of smart versus poor students. Given large samples for boys in maths or science, judgments of smart boys should increase and judgments of poor boys should decrease, relative to smart and poor girls. Conversely, the discrimination between smart and poor girls should be accentuated in languages, assuming larger samples for girls than boys.

However, importantly, the alternative process we are proposing is not intrinsically biased toward prior stereotypical beliefs, but a learning function of
variable sample size. The stereotype-congruent prediction holds only as long as sampling behavior is determined by the gender stereotype. The basic independence from prior expectancies can be demonstrated when teachers are asked to test counter-stereotypical hypotheses, that is, whether girls have assets in maths and science and boys in language. Positive testing (i.e., larger samples for girls in maths and for boys in language) should then foster accentuated ratings of girls in science and boys in language. This pattern is clearly distinct from a common expectancy effect, which would invariantly predict enhanced or accentuated judgments of boys and girls in their gender-typed domains.

That sample-size effects can override expectancies could be already demonstrated in recent experiments (cf. Fiedler, Walther, & Nickel, 1999). The present research not only constitutes an attempt to apply this replicable finding (see also Fiedler, Kemmelmeier, & Freytag, 1999; Semin & Strack, 1980) to the explanation of teacher judgment biases but also highlights important ecological constraints. Our cognitive–environmental framework predicts clearly specified boundary conditions for the predicted judgment biases. Granting that judgment biases are actually mediated by sample size, the predicted effects should be obtained only where teachers actually engage in positive testing. To meet this condition, it is essential that the environment support the sampling process. That is, a sufficient number of boys and girls must raise their hands in the respective domain to allow for positive testing. Otherwise, the process leading to sample-based judgment biases should be inhibited. The influence of focused attention should vanish if the stimulus environment does not support the information search process (cf. Fiedler, Armbruster, Nickel, Walther, & Asbeck, 1996).

Accordingly, teachers in Experiment 2 were instructed to test specific hypotheses that were either consistent or inconsistent with common gender stereotypes. In the consistent condition, teachers tested the hypothesis that girls show enhanced performance in language and boys in science. However, in a repeated-measures design, each judge was also to test counterstereotypical hypotheses, namely that girls have their assets in science and boys in languages. By this manipulation, the hypothesis focus and its influence on sample size were detached from the impact of a priori stereotypes. The A parameter matrix included both boys and girls who confirm and disconfirm gender stereotypes, although the entire gender groups did not differ in either language or science. M parameters were manipulated to either facilitate or inhibit positive testing (in terms of the number of boys and girls who raise their hands).

**Method**

**Participants and Design**

A total of 32 men and women participated in the experiment either for course credit or for payment (DEM 50.00). An equal number of participants were randomly assigned to eight conditions resulting from three between-subject factors, environmental support (positive testing supported by M parameters vs not), lesson order (languages in the first vs second experimental session),
and allocation of stereotype-congruent and incongruent hypotheses to disciplines (e.g., testing that girls are good at English and Physics vs girls good at German and Maths). The combinations of hypotheses to be tested and lesson orders are given in Table 2.

The last two factors were not of theoretical interest but merely intended to control for interactions involving gender, subject matter, and lesson order. As all results were unaffected by these variations, data were collapsed over these technical factors. As in Experiment 1, various within-judges contrasts were included in the data analyses (see Results), pertaining to performance level (smart vs poor), gender (boys vs girls), or disciplines (language vs science).

Materials and Procedure

The A parameter matrix was modified slightly to include only stereotype-congruent and incongruent conditions, but no mixed conditions. As shown in Table 3, within all subgroups of four smart (A = 0.8 or 0.5) and poor (A = 0.2 or 0.5) boys and girls, the between-disciplines variation confirmed the gender stereotype for one pair (i.e., girls being better in language and boys on science) and disconfirmed the stereotype for the other pair. Thus, the overall A pattern neither confirmed nor disconfirmed the stereotype.

The M parameter matrix was manipulated in the following way to either support positive testing or not. In the nonsupportive environment, the participation rate for all students in all lessons was held constant at M = 0.5. In the environment supposed to support positive testing, motivation (raising hand rate) was high (M = 0.8 or 0.5) for girls and low for boys (M = 0.2 or 0.5) in those lessons in which the hypothesis to be tested focused on girls, but M was higher for boys than girls in lessons with a hypothesis focus on boys. Within each pair of students who shared the same gender and parameter pattern (see last paragraph), one received the informative M parameter (i.e., 0.8 and 0.2) and the other received the uninformative parameter (0.5 and 0.5).

The general procedure remained invariant, except for changes in instructions necessary to introduce the specific hypotheses to be tested. Teachers were now provided with information about the gender group’s assets stemming from the alleged former teacher of the class. To introduce a hypothesis testing perspective, the following paragraph was inserted in the general instructions:

<table>
<thead>
<tr>
<th>Session</th>
<th>Group A</th>
<th>Group B</th>
<th>Group C</th>
<th>Group D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>German C</td>
<td>German I</td>
<td>Maths C</td>
<td>Maths I</td>
</tr>
<tr>
<td>2</td>
<td>Physics I</td>
<td>Physics C</td>
<td>English I</td>
<td>English C</td>
</tr>
<tr>
<td>3</td>
<td>Maths C</td>
<td>Maths I</td>
<td>German C</td>
<td>German I</td>
</tr>
<tr>
<td>4</td>
<td>English I</td>
<td>English C</td>
<td>Physics I</td>
<td>Physics C</td>
</tr>
</tbody>
</table>

Note: C, hypotheses consistent with gender stereotype; I, inconsistent.

TABLE 2
Sequence of Lessons in Different Subjects in Experiment 2
TABLE 3
Parameter Matrix Describing the Ability (Probability of Correct Responses) of the 16 Target Students in Experiment 2

<table>
<thead>
<tr>
<th>Discipline type</th>
<th>Students</th>
<th>German</th>
<th>English</th>
<th>Mathematics</th>
<th>Physics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1</td>
<td>.8</td>
<td>.8</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.8</td>
<td>.8</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.5</td>
<td>.5</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>.5</td>
<td>.5</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.5</td>
<td>.5</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>.5</td>
<td>.5</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>.2</td>
<td>.2</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>.2</td>
<td>.2</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>Female</td>
<td>9</td>
<td>.8</td>
<td>.8</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>.8</td>
<td>.8</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>.5</td>
<td>.5</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>.5</td>
<td>.5</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>.5</td>
<td>.5</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>.5</td>
<td>.5</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>.2</td>
<td>.2</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>.2</td>
<td>.2</td>
<td>.5</td>
<td>.5</td>
</tr>
</tbody>
</table>

You already had a talk with your predecessor about your pupils’ strengths. In individual lessons you are now going to test the supposition that...

...girls show good performance in German.
...girls show good performance in Physics.
...boys show good performance in Mathematics.
...boys show good performance in English.

The specific gender and subject combinations varied according to the experimental condition. Participants were told that the individual hypotheses would be repeated at the beginning of the lessons in the respective disciplines. Participants taught only one lesson per subject; extended teaching over more lessons might have counteracted the hypothesized positive-testing strategies.

Results and Discussion

For ease of presentation, we report separate analyses for stereotype-consistent and -inconsistent hypotheses and for motivation environments that support versus inhibit positive testing. Recall that each judge tested one consistent and one inconsistent hypothesis in both the language domain and the science domain.

Positive Testing

First, a positive-test score was computed for each teacher in each discipline, defined as the difference between the proportions of hypothesis-matching questions and mismatching questions. For instance, if the hypothesis refers to girl's
superiority in maths, the positive-test score would be the proportion of maths questions asked to girls minus the proportion asked to boys. These scores provided the input to ANOVAs including the factors ability level (high-A vs low-A students) × discipline type (language vs science) × student sex (male vs female), conducted separately for environments that support or inhibit positive testing.

When the motivation parameters $M$ support the hypothesis being tested (i.e., when plenty of girls raise their hands to test a hypothesis about girls), positive testing is evident in marked discipline type × student sex interactions. For stereotype-consistent hypotheses, judges asked girls more often (7.45) than boys (5.05) in language and more boys (7.35) than girls (5.15) in science, $F(1, 15) = 61.50, p < .001$. When inconsistent hypotheses were being tested, a reversed positive-testing effect is equally strong, such that teachers asked more boys (7.36) than girls (5.14) in language and more girls (7.52) than boys (4.97) in science, $F(1, 15) = 59.49, p < .001$. Positive testing does not interact with other factors.

As predicted, positive testing is greatly reduced or even eliminated when the environment does not support the hypothesis being tested, confirming the important role of environmental support. When the cooperation rate of boys and girls in language and science inhibits the sampling of hypothesis-congruent information, the discipline type × student sex interaction disappears for consistent hypotheses, $F(1, 15) = 0.84$, and reduces to $F(1, 15) = 7.36, p < .05$, for inconsistent hypotheses.

From Positive Testing to Judgment Biases

To test the major theoretical prediction concerning the mediational role of positive testing in the formation of judgment biases, we conducted discipline type (language vs science) × positive test (present vs absent) × student sex ANOVAs. Positive testing was assumed to be present (absent) in those teachers who had gathered a higher number (a lower or equal number) of questions matching the hypothesis than mismatching questions. For convenience, we reduce the complexity by reporting separate analyses for different environmental boundary conditions.

Environments Supporting Positive Testing

Let us first consider an environment in which students’ motivation facilitates positive testing. Let us further concentrate on judgments of cases (student-discipline combinations) where the actual ability is high ($A = 0.8$) and the hypotheses being tested are consistent with gender stereotypes. Mean $A$ judgments (Table 4) under these conditions show a marked confirmation bias. Girls receive higher ratings than boys in language and lower ratings in science, and this pattern is almost exclusively due to those cases in which judges actually engage in positive testing. This pattern is evident in a discipline type × student sex interaction, $F(1, 15) = 10.66, p < .01$, and most crucially, a significant
TABLE 4
Judgments of Ability When Ability Is High (0.8) by Environments Support of the Hypothesis Being Tested in Experiment 2

<table>
<thead>
<tr>
<th>Environment</th>
<th>Consistent with stereotype</th>
<th>Inconsistent with stereotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive testing</td>
<td>No pos. testing</td>
</tr>
<tr>
<td>Supporting</td>
<td>Girls</td>
<td>Boys</td>
</tr>
<tr>
<td>Language</td>
<td>0.72</td>
<td>0.52</td>
</tr>
<tr>
<td>Science</td>
<td>0.56</td>
<td>0.64</td>
</tr>
<tr>
<td>Inhibiting</td>
<td>Girls</td>
<td>Boys</td>
</tr>
<tr>
<td>Language</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>Science</td>
<td>0.58</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Note. Underlining indicates those cases that are focused in the hypotheses being tested.

three-way interaction, \( F(1, 15) = 6.42, p < .05 \), restricting the effect to the presence of positive testing. An additional student sex main effect, \( F(1, 15) = 5.11, p < .05 \), reflects somewhat more benevolent judgments of girls than boys.

Of special theoretical value is the condition in which the hypotheses being tested are stereotype-inconsistent so that alternative interpretations in terms of expectancies are ruled out. As apparent in Table 4, the same sampling-dependent bias is evident in a significant three-way interaction, \( F(1, 15) = 10.27, p < .01 \). When judges engage in positive testing, girls receive higher ratings than boys in science (!) and lower ratings in language (!), and this pattern disappears when judges do not gather positive-test samples. No other result is significant.

When we consider judgments of cases (student±discipline combinations) with low ability (A = 0.2), the qualitative pattern of means is reversed but much weaker; the three-way interactions do not become significant. This means that judges are more contingent on observations of accomplishments than on failures, which may reflect the fact that hypothesis testing instructions focused on assets rather than deficits (see above). Importantly, however, when we compute each judge’s ratings of high-A cases minus his/her ratings of low-A cases, the crucial three-way interaction appears for judgments of ability when testing consistent hypotheses, \( F(1, 15) = 6.04, p < .05 \), and when testing inconsistent hypotheses, \( F(1, 15) = 11.59, p < .01 \).

M judgments reflect the judges’ general sensitivity to the fact that, in a supporting environment, motivation is enhanced for those disciplines that are in the focus of hypothesis testing, as manifested in discipline type x student sex interactions for both consistent hypotheses, \( F(1, 15) = 11.52, p < .01 \), and inconsistent hypotheses, \( F(1, 15) = 21.82, p < .001 \). Motivation is correctly recognized as elevated for those gender–discipline combinations that are under focus (mean = 0.625) and as reduced for the complementary combinations (0.482), indicating that the M manipulation was effective. Moreover, discipline type x student sex x positive test interactions, \( F(1, 15) = 9.23, p < .01 \), and
F(1, 15) = 7.69, p < .05, indicate that this variation in motivation is more clearly detected when positive testing is used.

Environments Inhibiting Positive Testing

In contrast, when there is no environmental support for judges’ positive testing, that is, when students’ cooperation rates do not provide affordances for sample-size effects, the influence of hypothesis focus on A ratings disappears largely (see bottom of Table 4). The crucial three-way interactions are no longer significant (Fs < 1.6). Although in this condition we can still distinguish positive and negative testing at a qualitative level, the quantitative sample-size effect is too weak (see above) to produce noticeable judgment biases.

EXPERIMENT 3

Experiment 3 addresses the last type of expectancy effects, originating in naive theories or scripts about the genesis of correct (vs incorrect) student responses. Especially when student ability is hard to discern while motivation is more visible, teachers may utilize M cues to infer A, relying on scripted knowledge that raising hands is often the first step in an episode resulting in a correct (or incorrect) answer. If so, a constructive version of an expectancy bias may be obtained. Judgments of manifestly observed performance (i.e., proportion of correct answers) may be confused with inferences derived from scripted knowledge. Even when incomplete episodes of a student’s raising hand does not receive as much weight as complete episodes of a correct response, the teacher’s effective performance sample may increase with the number of M cues, based on a scripted inferences sequence: student raises hand → teacher notices → teacher asks the student → student responds → response is correct.

To provide the stage for such a constructive (i.e., inference-based) version of an expectancy effect, the ability level of all students was held constant and only their motivation was allowed to vary. However, in different experimental environments, ability was either constantly high (A = 0.7) or constantly low (A = 0.3). When teachers are asked to rate student ability in the absence of any variation in A, chances should be high that M variation intrudes into A judgments. The precise pattern of this effect should be revealing about the process underlying the expectancy effect. The traditional heuristic account would predict that under high uncertainty (i.e., when differences in A are restricted to sampling variation), judges should resort to simplifying heuristics such as the assumption that a student who raises her hand is likely to know the correct response. Heuristics are assumed to be applied in a routinized manner; that is, students who raise their hand often should receive higher ability judgments. In contrast, extending the environmental learning approach to constructive inferences predicts a more differentiated pattern. M cues will be used differentially in high-A and low-A environments. In a smart base rate environment (A = 0.7), judges learn to associate M (raising hand) with success, whereas in a poor environment, they should associate M with a high number
of failures. Accordingly, high-M students should receive more favorable ratings than low-M students when $A = 0.7$, but high M should lead to more negative A ratings than low M when $A = 0.3$, as the base rate of failure after raising hands is high.

Such an adaptive, base rate-sensitive use of scripted knowledge is different from the notion of heuristic shortcuts. Our revisited environmental-learning approach suggests that the utilization of environmental base rates can also be understood within the sampling framework. Accordingly, the virtual stimulus sample for each student consists of (a) the subsample of observed complete performance episodes and (b) the subsample of incomplete inferred performance episodes. The first component (a) does not discriminate between students because $A$ is constant; the expected level for the entire class is either high ($A = 0.7$) or low ($A = 0.3$). However, the second component (b) should induce a systematic bias. Depending on whether teachers have learned to associate M with success or failure, increasing the number M-based inferences should either increase or decrease ratings of A. If judgments are in this way sensitive to virtual sample size, the predicted impact of M on A (polarized A judgments for high M students) should be restricted to those informative lessons in which students actually show their high versus low motivation (i.e., where M is manifested in sample size). Polarized judgments of high-M students should not generalize over all lessons, as one would expect if crude heuristics replace piecemeal information processing.

Method

Participants and Design

A total of 30 male and female native speakers (most of them students at the University of Heidelberg) participated for payment (DEM 75.00; approximately $40). An equal number of participants were randomly assigned to the two levels of the only between-subjects factor, ability level ($A$ set to a constant value of 0.3 vs 0.7). For all participants, the M parameter varied across students, according to the same parameter matrix as in Experiment 1 (see Table 1). Various contrasts between dependent measures were included as within-subjects factors in the data analyses.

Procedure

The same procedure was used as in Experiment 1, except for necessary changes to the underlying parameter matrices. Thus, while the experimental situation remained unchanged at surface level, the latent parameters of students’ performance underwent important changes.

Results

It was predicted that when ability is invariantly high ($A = 0.7$), motivated students should receive higher A judgments than less motivated students.
Conversely, when ability is constantly low \((A = 0.3)\), motivated students should receive generally lower judgments. This pattern should be confined to informative lessons where differences in M are actually present in virtual samples.

An overall ANOVA of A ratings included three factors, ability baserate \((A = 0.7 \text{ vs } A = 0.3)\) × discipline types (languages vs science) × motivation levels (high vs low M), the last factor being between students and the first two factors within students. A huge ability base rate main effect, \(F(1, 14) = 238.45, p < .001\), reflects higher ratings when the A base rate is 0.7 rather than low 0.3. As usual, however, judgments were regressive (mean ratings 0.609 vs 0.366), failing to extract the full size of the A manipulation.

More importantly, the ability baserate × motivation level interaction, \(F(1, 14) = 5.79, p < .05\), confirms the main prediction that in a high-ability environment \((A = 0.7)\) the judged ability of high-M students is higher (0.626) than for low-M students (0.593), whereas in low-ability environments \((A = 0.3)\), A ratings for high-M students exhibit an opposite bias (0.35 compared with 0.39 for low-M). This pattern is consistent with an adaptive learning process as distinguished from a stimulus-insensitive heuristic. The only other effect is due to the base rate × discipline types interaction, \(F(1, 14) = 5.81, p < .05\), reflecting stronger discrimination for sciences than for languages.

Further evidence for the stimulus learning approach comes from more refined analyses that include motivation level and two other factors: informative lessons (when high or low M is manifested at \(M = 0.7\) or 0.3) and uninformative lessons (when the same students’ M becomes 0.5; see Table 1) and judgments based on large versus small samples (i.e., samples above vs below each individual teacher’s median sample size for all students). Thus, ratings of the same high-M and low-M students were averaged separately over lessons in which M was informative or not, and over teachers who gathered large or small students of those students. For convenience, separate ANOVAs were conducted for high-A and low-A environments. In both analyses, the informativeness × motivation level interaction was significant, \(F(1, 14) = 8.08, p < .05\) (for \(A = 0.7\)) and \(F(1, 14) = 12.31, p < .01\) (for \(A = 0.3\)), indicating that polarized A judgments of high and low M students are not overgeneralized but restricted to those lessons where M samples actually vary (see Table 5).

The ANOVA of M ratings yields a strong M levels main effect, \(F(1, 14) = 250.27, p < .001\), reflecting the true M differences. Confusion of M and A is evident in a strong main effect for A base rates, \(F(1, 14) = 43.32, p < .001\).

### Table 5

Judgments of Ability as a Function of Motivation Level, Informativeness, and Ability Environment in Experiment 3

<table>
<thead>
<tr>
<th>Ability environment</th>
<th>Informative</th>
<th></th>
<th>Uninformative</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High M</td>
<td>Low M</td>
<td>High M</td>
<td>Low M</td>
</tr>
<tr>
<td>When A = 0.7</td>
<td>0.64</td>
<td>0.58</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>When A = 0.3</td>
<td>0.32</td>
<td>0.48</td>
<td>0.38</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Teachers erroneously inferred from generally high A (0.7) a higher degree of M (0.551) than from low A (0.473). No other effect reaches statistical significance.

**Discussion**

Experiment 3 corroborates the contention that the sampling approach can be extended from objectively observed performance to subjective inferences of ability based on motivational cues. When there are many M cues, A judgments are enhanced or reduced, depending on whether the motivational cue (raising hand) is associated with a high or low ability base rate. This context-sensitive use of M cues in different environments cannot be explained in terms of rigid heuristics but calls for a more flexible, stimulus-sensitive account. If student participation (high M) was frequently met with success, teachers inferred high A from high M. Conversely, if motivation often cooccurred with failure, they inferred reduced A from high M. This holds only for those lessons that actually afford different samples of M-based inferences. Thus, biased judgment outcomes did not appear as a consequence of overload, reduced motivation, or motivated distortions, but as an adaptive way of dealing with actually observed stimulus uncertainty.

**GENERAL DISCUSSION**

Returning to the question asked at the outset, it is indeed possible that a female student whose achievement in maths is low, who participates little, and whose general performance level in other subject matters is also poor, will benefit rather than suffer from teachers' judgment biases. The regressiveness of inductive judgments is of advantage for poor rather than smart students, and a low participation rate implies small samples and thereby keeps the regression effect strong. So remaining mute might be a good strategy. For the same reason, the stereotypical hypothesis that maths is a male domain may not produce a negative bias against that girl, if the hypothesis induces positive testing. Smaller samples about girls than boys in maths may serve to conceal the girl's weakness, leading to less negative judgments of poor girls than poor boys.

Thus, at least under certain boundary conditions, our experiments support some uncommon predictions that set our environmental learning approach apart from previous accounts of biased judgments referring to prior expectancies as the major explanatory construct. Besides the basic accuracy of teacher evaluations emphasized by some authors (e.g., Jussim, 1989, 1991), deviations from the objective data in judgments are almost always interpreted in terms of teachers' prior expectancies and (self-fulfilling) prophecies (Miller & Turnbull, 1986; Rosenthal & Jacobson, 1968). Although different kinds of expectancies can be distinguished, reflecting the overgeneralization of stable traits, group stereotypes, or naive behavioral theories, the common denominator is that expectations are assumed to exist before the assessment of stimulus data.

Our intention was not to question the existence and the significance of even
very strong effects of prior expectations. Hardly anybody would come to doubt that prejudice, stereotypes, or wishful thinking can markedly distort judgments and decisions (Hamilton, Sherman, & Ruvolo, 1990; Kunda, 1990; Pyszczynski & Greenberg, 1987). Rather, the strategy of the present research was to create conditions that highlight alternative sources of bias, arising as a normal by-product of environmental learning. In the absence of any prior expectancies, ordinary learning principles—concerning sample size and differential regression—can create a variety of judgment biases that deserve the same theoretical attention as the commonly noted reliance on expectancies. To highlight the distinctive nature of our alternative approach, we have created conditions in which unbiased environmental-learning processes override three of the most prominent types of prior expectancy.

First, any tendency to overgeneralize global differences between smart and poor students was overridden by the regressiveness of all inductive learning, producing an opposite tendency to overestimate the achievement of poor students and to underestimate the achievement of smart ones (Experiment 1). Any self-fulfilling prophecy or perpetuating influence (Cooper, 1979; Rosenthal & Jacobson, 1968) of classifying students into smart versus poor has to overcome the inertia of the initial regression effect. The regressiveness of inductive judgment affords a pervasive remedy against prior expectancy effects of the first type, which consists in the use of global knowledge as a basis for judgments under uncertainty, yielding a positive bias for smart students and a negative bias for poor students. The rules of inductive learning counteract such a kind of halo effect. It remains a matter of empirical research to figure out which tendency is stronger under natural conditions. However, for theoretical reasons, the impact of regression effects must not be simply ignored.

Systematic biases—defined as deviations of subjective estimates from objective parameters—arise nevertheless because regression is a function of sample size. Technically speaking, the influence of large samples is to dampen a basic regression effect, making actually existing differences between smart and poor students more visible. Of two equally smart students, the one who was represented by a larger sample received regularly better ratings. Conversely, large samples led to more negative evaluations of poor students. It does not matter why teachers draw large versus small samples about individual students. Their sampling may be a function of sitting position, students’ names, their appearance, teacher’s sympathy, or students’ motivation (frequency of raising hands). For some of these reasons the judges might be blamed (e.g., sympathy for particular students) but others are fully “innocent,” representing natural fluctuations in the students’ behavior. In any case, sample size affords a permanent source of judgment bias, reflecting the richness of the learning environment rather than biased cognitive processes.

In Experiment 2, the second type of prior expectancies, based on stereotypes, was overridden by sample-size differences resulting from positive-test strategies in hypothesis testing. Depending on the teacher’s hypothesis-testing goal, sampling focused particularly on gender groups (boys vs girls) in particular disciplines (language vs maths). When the hypothesis to be tested was congruent
with common gender stereotypes, teachers would collect larger samples on boys in maths and on girls in language. As a consequence, smart boys in maths and smart girls in language appeared even smarter, but not poor boys and girls in their respective gender domain. However, when testing counter-stereotypical hypotheses, teachers would exhibit a similarly strong bias in the reverse direction. They would draw large samples about girls in maths and boys in language and consequently judge smart girls in maths and smart boys in language even smarter, but not poor girls and boys. In this manner, our environmental learning approach led to a refined picture of the mechanisms by which stereotypes can affect the inductive search and processing of information. At the same time, our approach sheds new light on the interpretation of hypothesis confirmation biases through positive testing (Crott, Giesel, & Hoffmann, 1998; Klayman & Ha, 1987), independent of an expectancy bias.

Experiment 3 demonstrated that the third type of prior expectancies, reflecting naïve behavioral theories or scripts, also can be overridden by environmental learning. When ability was held constant and students varied only in motivation but teachers were nevertheless forced to judge ability, they tended to infer ability from motivational cues. Just as in the case of objectively addressed student responses, the teachers' subjective inferences from motivational cues (raising hands) were subject to sample-size effects. Ability inferences from motivational cues led to polarized judgments as sample size (i.e., opportunities to make inferences) increased. However, the direction of the bias was not determined by the common script that raising a hand is the first step in an episode ending with a correct response. In a high-ability environment, when raising hands was followed by a high base rate of correct answers, motivation was used as a positive cue, giving higher ability judgments to highly motivated students. In contrast, in a low-ability base rate environment, when raising hands was typically paired with wrong responses, high motivation led to reduced ability judgments. Again, this demonstration highlights the power of environmental base rates to moderate inferences based on scripted knowledge.

To understand the systematic biases in relative judgments of smart versus poor students that result from ecological sampling effects, we need not postulate any deficits of shortcomings in the teachers' cognitive processes. There is nothing irrational in selecting larger samples from students who raise their hands very often, in focusing on those events specified in a hypothesis (Oaksford & Chater, 1994), or in utilizing an associative link to infer ability from motivation. Drawing flat, fully equal samples from each and any target object in the environment would be detrimental to adaptive cognition and behavior. Being fed with, and relying on, samples of differing size is inevitable, because environmental objects always differ in salience, proximity, accessibility, and relevance to the individual's problem-solving goals (Friedrich, 1993; Klayman & Ha, 1987).

Biases may result from such unequal stimulus samples not because judges are misled by cognitive or motivational attractors or distracted by convenient heuristics from systematic processing but, on the contrary, precisely because they are so sensitive to the information inherent in the samples themselves. Normatively appreciated information-processing devices such as a Bayesian
computer algorithm, when fed with the same sample of unequal size, might exhibit the same judgment biases (e.g., polarized judgments with increasing sample size). By emphasizing accurate assessment and sensitivity to environmental variation as origins of judgment bias—as opposed to cognitive and motivational distortions—the sampling approach helps to reconcile the apparent conflict between those theorists who point out the remarkable accuracy of student evaluations and those concerned with systematic biases. The coexistence of both positions was already conceded by other researchers (cf. Gigerenzer & Goldstein, 2000; Jussim et al., 1994). The contribution of the present research is to suggest a concrete set of cognitive–ecological processes by which accurate stimulus processing can result in biased judgments.

If there is one deficit for which teachers themselves can be blamed, it is not the construction and use of heterogeneous samples but their inability to monitor and control the consequences of sampling effects (Koriat, 1993; Nelson, 1996). Granting that the stimulus environment provides us with samples that differ greatly in size—as well as other pragmatic factors, such as source, conditionality, and selectivity (Fiedler, 2000)—human intelligence might have evolved devices for controlling and correcting the impact of sampling effects. There is growing evidence to conclude that human judges often fail to understand sampling constraints and the deductive rules needed to combine samples of different origin (Evans & Dusoir, 1977; Kleiter et al. 1997; Sedlmeier & Gigerenzer, 1997; Tien, 1974; Tversky & Kahneman, 1971). They do not sufficiently notice the effect of sample size or number of learning trials on their memory (Koriat, 1997); they tend to treat conditional implications (if p, then q) as if they were reversible (Evans, 1989; Wason, 1966); their sampling space is often affected by logically irrelevant factors (Gavanski & Hui, 1992); they take samples for granted that exaggerate the base rate of rare but important events in the population (Fiedler, Brinkmann, Betsch, & Wild, 2000); and they may blindly trust in selectively obscured samples (Patrick & Iacono, 1991).

Why don’t judges take the impact that sampling constraints have on their judgments into account? Why have not appropriate metacognitive devices for the correction of biases resulting from unequal or incomparable stimulus samples evolved? Although an informed answer to this intriguing question (see Fiedler, 2000) is not yet available, an answer might be inherent in a seminal article by Einhorn and Hogarth (1978). What our findings demonstrate is that judges deal quite accurately with the observed data in front of them. What we have called a meta-cognitive deficit might originate in an inability to take into account unobserved data that might have been, but were not, sampled. Human intelligence may not have evolved devices for taking such unavailable data into account (cf. Einhorn & Hogarth, 1978).

REFERENCES


Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. Psychological Review, 93, 258–268.


Received July 17, 2001