Judgment and Decision Making, Vol. 14, No. 6, November 2019, pp. 624–643

Learning to reason: The influence of instruction, prompts and scaffolding, metacognitive knowledge, and general intelligence on informal reasoning about everyday social and political issues

David Perkins*

Abstract

Twelve experiments examined ways of improving informal reasoning, as assesed by presenting students with accessible, current, and interesting social and political issues, eliciting reasoning about them, and scoring the reasoning for quality of argument. The experiments addressed: (1) the impact of established instructional programs that emphasized critical thinking (Experiments 1–4); (2) the impact of an investigator-designed high school level minicourse (Experiments 5–7); (3) the responsiveness of subjects to prompts that asked them to develop arguments more fully, and the relation of their responses to general intelligence (Experiments 8–10); (4) checks on the validity of the testing methodology (Experiments 11–12). Two of the established instructional programs had a beneficial effect. The minicourse had a particularly large effect on students' attention to the other side of the case, the most neglected aspect of informal reasoning. The prompting studies showed that subjects could develop their arguments far more than they normally did. Finally, subjects with higher intelligence were actually somewhat more biased in their reasoning. In summary: people can reason much better than they typically do on the sorts of issues posed; people are not performing near the limits of their abilities; strategies and standards of good reasoning can improve reasoning; and education can develop students' reasoning much further than education typically does.

Keywords: myside bias, informal reasoning, training, education

1 Summary

A series of experiments was undertaken to investigate whether informal reasoning could be improved or was largely limited by relatively stable capacities such as general intelligence in the psychometric sense. Informal reasoning refers to reasoning outside of a formal logical or mathematical context, for instance as in building a case in an essay or making a personal decision. The experiments were motivated by prior research that had shown that conventional education at the high school, college, and graduate school levels had very little impact on informal reasoning and that reasoning per-

This article is a minor edit for clarity of the final report, February, 1986, of the project Learning to Reason, supported by the National Institute of Education, Grant No. NIE-G-83-0028, Project No. 030717, grant period September 30, 1983 to January 31, 1986. We thank NIE for its support. We note that opinions expressed here are not necessarily shared by the National Institute of Education or its successor, the Office of Educational Research and Improvement, and do not represent the policy of either entity. Project assistants Barbara Bushey and Michael Farady receive much appreciation; they contributed extensively to the research and also to the writing of some elements the final report. Of course, the principal investigator and primary author, David Perkins, bears full responsibility both for that report and this version.

A number of other individuals and institutions are due thanks for their help in carrying the project to a conclusion. We value the generative relationship with our NIE project officers, first Martin Engel and later Kent Viehoever. Israel Scheffler of the Harvard Graduate School of Education served throughout as an advisor to the project: we are grateful for his guidance. Volunteers extended the compass of the research from time to time. During the summer of 1984, Beverly Bushey and Jana Schnurr conducted studies of the influence of directly requesting people to develop their arguments further (Experiment 9) and of the impact of oral versus written methods of gathering arguments (Experiment 11). During the spring of 1985, Andrea Meier, with help from Keith Carreiro, conducted a study comparing people's reasoning on one of our social issues with their reasoning on vexed and important personal decisions. Several high schools, colleges. and graduate schools cooperated in allowing us to work with their students. We do not list these institutions in order to preserve confidentiality, but we extend our thanks for their help. In the same spirit, we thank the many individual participants in our experiments who together allowed us to assemble the conclusions presented here. Finally, we greatly appreciate the initiative of the editorship of Judgment and Decision Making in making this research report from some years ago available to the professional community today.

Editor's note (J. Baron): This article is being published now, for the first time, because it is historically important, especially in the development of research on "confirmation bias" or "myside bias". It may contain the first use of the latter term. We decided not to bring the literature review up to date, as this would be a formidable task and would remove the paper from its historical context. The original report inspired, and was sometimes cited in, some of the work on "actively open-minded thinking" and "myside bias"; readers should consult that work for more recent reviews. Note also that we kept the original format as a final report rather than trying to convert this to a briefer and more focused journal article. Many issues (such as correlations with intelligence) are addressed several times in different experiments, so some results that may seem to suffer at first from small sample sizes are conceptually replicated in other experiments.

*Harvard Graduate School of Education. Email: David_Perkins@harvard.edu.

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

formance correlated substantially with general intelligence.

The general testing methodology involved posing accessible, current, and interesting social and political issues, eliciting reasoning about them, and scoring the reasoning in several ways for quality of argument. Twelve experiments were carried out, addressing four matters: (1) the impact on informal reasoning performance of established instructional programs that paid special heed to critical thinking (Experiments 1–4); (2) the impact of a high school level minicourse in informal reasoning designed by the investigators (Experiments 5–7); (3) the responsiveness of subjects to content-free prompts designed to press them to develop arguments more fully and the relation of their responses to general intelligence (Experiments 8–10); (4) studies designed to crosscheck the validity of the testing methodology (Experiments 11–12).

The results demonstrated that informal reasoning is subject to substantial improvement. Two of the three established instructional programs had a statistically significant impact which, while not great, was several times that of the miniscule influence of normal education for the same time period. The minicourse designed by the experimenters had a sizable impact on students' attention to the other side of the case, the most neglected aspect of informal reasoning. The studies involving content-free prompts showed that subjects could develop their arguments far more than they normally did without guidance. The studies also disclosed that subjects with more general intelligence actually tended to be more biased in their reasoning than subjects with less general intelligence. Finally, the methodological studies supported the soundness of the methodology.

In summary, the findings demonstrated that people can reason informally much better than they typically do on the sorts of general issues posed. The findings argue strongly against the hypothesis that people are performing near the limits of their mental abilities when they engage such issues and for the hypothesis that strategies and standards of good reasoning can lead people to reason substantially better. It appears that education can develop students' reasoning skills much further than education typically does.

2 Introduction to the Program of Investigation

The development of students' reasoning is a long-time aspiration of education at all levels. Schooling hopes for more than equipping learners with a repertoire of facts about arithmetic, geography, English, and so on. Ideally, students should emerge from a dozen or more years of education better able to consider claims critically and evenhandedly, confront the multiple factors often bearing on important personal and social decisions, and reach resolutions as sound as the available information allows. It might be hoped that the normal course of education fosters such abilities. After all, while a good deal of schooling concerns itself with low-order facts and skills, students do from time to time write essays, discuss political and other issues, hear critical analyses of existing viewpoints and attempt their own critiques, and so on.

Regrettably, a major study conducted a few years ago suggests that conventional education has a very limited impact on the development of informal reasoning ability – the ability to construct arguments outside of the formal contexts of mathematics and logic. Such arguments typically call for several lines of reasoning on both sides of the case to do justice to the multiplicity of factors that impinge on complex everyday matters; the demands of informal reasoning are somewhat distinct from those of formal reasoning because of this (Perkins, 1985; Perkins, Allen & Hafner, 1983). In the research mentioned, the investigators collected informal arguments on accessible, contemporary issues from populations ranging from the first year of high school through the fourth year of graduate school, as well as from individuals who had been out of school of any sort for a number of years (Perkins, 1985; Perkins, Allen & Hafner, 1983). The methodology encouraged subjects to think carefully about the issues and to report their thinking thoroughly, avoiding features that might cue an autocratic or defensive stance; for instance, the word "argument" itself was avoided entirely because of its contentious tone in everyday parlance. Several different sorts of counts as well as ratings of argument quality were employed to assess the complexity, evenhandedness, and general soundness of subjects' arguments.

The analysis disclosed a minimal impact of education on reasoning (Perkins, 1985). Comparisons of first year with fourth year students in high school, college, and graduate school revealed only very slight gains in consequence of three years of education and maturation. For instance, on the average, subjects added only 1/10 of a line of argument per year of education. Also, the overall level of performance was unimpressive. Although the issues were selected for allowing elaborate arguments on both sides based on common knowledge, the subjects' arguments tended toward brevity and neglect of the side of the case opposite that adopted by the subject, this despite marked press in the methodology toward thoroughness. Those subjects who had been out of school for several years appeared not to have learned any "lessons of life" about informal reasoning; broadly speaking, their level of performance resembled that of students with similar degrees of education. Finally, subjects' performance in general correlated substantially with IQ, suggesting that good informal reasoning might simply reflect general intelligence in the psychometric sense. Overall, the results argued that students do not reason very well, that neither education nor "life" helps them much to do better, and that perhaps neither education nor "life" could, since good informal reasoning depends on general intelligence, which is not very subject to change.

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

The last point, however, did not suit the character of difficulties with informal reasoning that appeared in the data. For example, as already noted, subjects commonly neglected arguments on the side of the case opposite their own, "otherside arguments" as we call them in this body of research. But, it would seem, people could learn to cue themselves to pay more heed to otherside arguments. For another example, subjects often overlooked common-knowledge counterexamples to general propositions (Perkins, Allen & Hafner, 1983). Plausibly, if they asked themselves for counterexamples, they might retrieve some. In general, many of the lapses evinced by the subjects were the sorts of lapses that might be reparable by a stronger repertoire of reasoning strategies and the disposition to use them. The possibility that higher-order strategies might improve intellectual performance also gains support from research in contexts of mathematics instruction, reading instruction, and other areas, where successful teaching experiments have occurred (e.g., Bolt, Beranek and Newman, 1983; Palinscar & Brown, 1984; Schoenfeld, 1982; Schoenfeld & Herrmann, 1982; see the reviews in Nickerson, Perkins & Smith, 1985). There is no apparent reason why informal reasoning should not allow similar treatment.

This pattern of results and interpretations led to the key question for the program of research reported here: Is informal reasoning a skill dependent on a heuristic repertoire and dispositions that might be acquired, enabling substantial improvements in performance? Alternatively, is informal reasoning a skill largely reflective of general intelligence and therefore resistant to improvement short of the intensive treatments that sometimes yield changes, typically temporary, in general intelligence measures (Berrueta-Clement, Schweinhart, Barnett, Epstein & Weikart, 1984; Garber & Heber, 1982; Jensen, 1983, 1984; Ramey, MacPhee & Yeates, 1982)?

With support from the National Institute of Education, four kinds of studies were undertaken to examine this issue.

- Our prior research on the impact of education had examined fairly conventional educational settings. In this program of research, we measured with pretests and posttests the impact of instruction in existing settings that gave special attention to informal reasoning, for instance a high school debate class (Experiments 1–4).
- Not confident that skills of informal reasoning were optimally addressed even in such settings, we designed our own "minicourse" in informal reasoning and taught it three times to different groups of high school students, with pretesting and posttesting to gauge its impact (Experiments 5–7).
- Concerned to probe in detail the factors that limit and empower people in informal reasoning, we conducted experiments designed to probe whether subjects were performing near their capacity limits or, with some general strategic guidance, could immediately in the same session perform much better (Experiments 8–10).
- Recognizing that our methodology for investigating informal reasoning would benefit from checks, the investigators conducted two studies to validate aspects of the methodology (Experiments 11–12).

The program of investigation spoke to the general issue raised as follows. If findings disclosed little impact of instruction particularly oriented to informal reasoning on performance, if even the investigators' especially designed instruction had little impact, and if the process-oriented experiments in which subjects were directly pressed to develop their arguments further showed that subjects had little ability to do so, this would argue that informal reasoning is a performance dominated by capacity limits, admitting modest improvement at best through instruction. On the other hand, if results on all three fronts showed substantial responsiveness of subjects, this would argue that people could learn to reason much better than they do and further imply that education should serve them in developing this important skill much better than it does. Of course, intermediate results between these two extremes would call for intermediate interpretations.

We turn now toward describing the particular experiments.

3 Experiments 1–7: The Impact of Instruction on Informal Reasoning

As noted earlier, we both tested the impact of some established instructional programs and designed and tested our own brief intervention. Since the testing methodology for the two sorts of studies was the same, it is convenient to describe them together.

3.1 Method

3.1.1 Subjects

Experiments 1 through 4 involved administering pretests and posttests of informal reasoning to students in established instructional programs. One high school class participated – a debate class in a public school. There were three programs at the university level – one group of college freshman, one group of graduate students in education, and one group of first year law students. The freshmen were enrolled in a college that emphasized critical thinking, the graduate students were enrolled in a class that explored aspects of thinking, but without extensively addressing reasoning as treated here, and the first-year law students were drawn from a prestigious law school. Experiments 5 through 7 involved high school students participating in a short course in

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

TABLE 1: Summary of Groups in Experiments 1–7.

Setting	Kind of Instruction	Duration of Instruction	N
Public high school	Debate	2 semesters	33
College	Critical liberal arts	2 semesters	35
Graduate school	Educational design	1 semester	27
Professional school	Law	2 semesters	41
Public high school	Reasoning class	4 weeks	08
Vocational/tech high school	Reasoning class	4 weeks	13
Parochial high school	Reasoning class	4 weeks	16

reasoning devised by the investigators. These groups were drawn from a vocational-technical, a parochial, and a public school. Each subject group was balanced for sex. The high school and education students participated voluntarily; the rest received a moderate fee for taking the pre- and posttests. Table 1 summarizes characteristics of the seven groups.

3.1.2 Treatments

These experiments took advantage of the "natural" treatments in certain extant educational programs as well as a treatment designed by the investigators. A total of five types of programs were evaluated. Debate, critical liberal arts, graduate students of education taking a course that emphasized aspects of thinking, and law were selected because it seemed likely that each of these instructional programs might engender improvements in generic thinking skills. In all these groups, first year students were utilized in order to give the greatest chance for capturing the effect of the respective programs on informal reasoning. The fifth program was an experimental reasoning course developed by the investigators for high school students. Some comments on each type of program follow.

Critical liberal arts. This program was chosen because it places an emphasis on aggressive problem solving. The program examined encourages individual initiative in both recognizing and working on problems. Students are called upon to tap their own critical resources, which includes finding and utilizing information sources, in order to navigate difficult intellectual terrain. The liberal arts focus suggests that success in the program would depend on developing general thinking strategies, rather than subject-specific expertise. This general focus, combined with the emphasis on individual initiative, seemed as though it might foster a style of examining issues that would transfer to out-of-school contexts of critical thinking.

Education course. This group of students were participating in a course that highlighted various aspects of systematic thinking and learning in the context of educational design. Informal reasoning as such was treated briefly but was not the focus. It was thought that the passing treatment of informal reasoning plus the general emphasis on patterns of thinking might enhance students' informal reasoning as measured here.

Law school. In addition to familiarizing students with actual laws and cases, one of the goals of law training is of course to develop skills of thinking like professional lawyers. These plainly include informal reasoning skills. The professional lawyer benefits from recognizing the complexity in issues, considering both sides – so as to anticipate what lawyers on the other side may argue – and attending carefully to the meaning of words and the weight of evidence. We hypothesized that the first year of law school would develop such skills and that they would carry over to performance on our non-legal issues.

Debate. In debate programs, students examine issues, collect evidence on both sides, and argue from one side or the other. It is necessary to become relatively expert on at least a few major topics. It is also necessary to learn to recognize how arguments fit together, that is, what points are most compelling and to which rebuttals they are vulnerable. There is a close affinity to law programs in that, again, close attendance to language and to evidence is crucial to success. Attention to both sides of the case is crucial, since a debater may be assigned to either side at the time of a formal match and even argue one side and the other during the same weekend. The instruction emphasized attention to both sides of the case. We hypothesized that debate instruction would develop these skills and that they would carry over to other contexts of critical thinking.

Experimental reasoning courses. The investigators designed experimental reasoning courses for three high school groups, involving lessons designed to emphasize those aspects of informal reasoning that our prior research suggested were most important and accessible. The classes met for about one hour, four days a week for four consecutive weeks, for a total of 16 lessons. Each course was team-taught by two research assistants who alternated days teaching and observing classes.

The basic content was the same for all three courses, al-

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

though the investigators varied the emphasis somewhat. The course content highlighted two basic principles and three standards for evaluating reasoning. The first principle encouraged generativity. Students were urged to utilize all of their knowledge about issues discussed in class and to come up with many reasons. The second principle concerned "myside bias," responding to the universal proclivity for considering mostly reasons that support one's own side of the case. Exercises had the students apply themselves to generating reasons contrary to their own position.

The three standards of reasoning emphasized were truth, relevance, and completeness: A good reason is both true and relevant to the conclusion, and collectively the reasons should be complete, taking into account all the true and relevant reasons. Exercises had students critiquing both their own and others' reasoning performances with these standards in mind. The classes were practice-intensive, structured around cycles that included a brief lecture by the instructor, a guided class performance, and individual writing exercises. Each class included pre and post-quizzes that were collected and later examined in order to gauge the efficacy of the lessons.

As noted above, although the essential content was similar in all of the courses, the emphasis varied depending on the level of the students. The instruction given the group at the vocational-technical school emphasized comparing and contrasting reasons in order to minimize bias, and expressing arguments in abbreviated but complete form. The parochial school group worked more explicitly on generativity and on critiquing arguments according to the standards. The public school group worked less on generativity, more on critiquing, and more on strategies for generating contrary points of view. These emphases, however, did not lead to any obvious difference in pattern of gains among the groups.

3.1.3 Test Procedure

After giving written consent to participate in the study, subjects completed a written questionnaire at the beginning and at the end of their respective courses of study. Students in the high school groups filled out their questionnaires in class, under the supervision of their instructors. The college, graduate and law students performed the task on their own. Those who completed their questionnaires outside of class were instructed to write their answers in a quiet setting in which they would not be distracted or interrupted and when they would have ample time to complete the task without hurrying. All subjects were told not to consult sources or people for help or information.

The questionnaire gathered information such as age and sex of the subject and then posed a hypothetical question about a topic of current interest. Subjects were instructed to indicate their snap judgment yes or no on the question, degree of confidence in their snap judgment, degree of interest in the question, and the extent of any prior thought. Next, the subjects were directed to think about the question and write down their thoughts. Subjects were instructed to write down all points thought of, even those that might not count in the end. Blank pages were provided to ensure that the reasoning performance would not be limited by space availability. Four groups – the three groups of high schoolers in the reasoning course and the graduate students in education – completed a double questionnaire treating two issues pre and two post. The other groups completed one issue pre and one post. After finishing the pretest, each subject filled out a short form written IQ test, the Quick Word Test (Borgatta & Corsini, 1964).

The issues used in the research were chosen for being genuinely vexed and timely. Four issues were employed in a counterbalanced design. The issues were selected, after piloting, because they permitted substantial arguments on both sides of the respective cases, proved accessible even to the youngest subjects, and did not depend for their analyses on background knowledge that varied greatly across the subject population. The issues used were the following:

- Would providing more money for public schools significantly improve the quality of teaching and learning?
- Would a nuclear freeze agreement signed between the U.S and the U.S.S.R. significantly reduce the possibility of world war?
- Should all 19 year olds be required to fulfill a one year social service obligation? (This issue was prefaced by a brief description of a non-military peacetime "draft" whereby 19 year olds would be required to work in hospitals or on public construction projects such as road or bridge repair.)
- Would a ban on selling and owning handguns significantly reduce violent crime?

3.1.4 Scoring

The written responses of the subjects were scored on several scales providing measures of the quality of the subjects' arguments. The pretests and posttests were marked with codes and shuffled together so that during scoring one could not tell pretests from posttests. The scoring was performed by two judges working independently; they co-scored a random subsample of the data to permit checking interjudge agreement. After the scoring was completed, each scale was examined for the correlation between the judges' scores. When two issues were completed pre and two post, the correlation between subjects' performances on the first and second issues was examined to test whether the questionnaires and the method of scoring measured a property of the subject or merely a property of individual performances.

It is natural to ask whether the simple counts and quality ratings to be discussed can do justice to the myriad ways in

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

which subjects' arguments might be weak or strong? The answer to this is that the measures used serve quite well for our subject population because, as it has turned out, most arguments people produce are relatively sparse and uncomplicated. Consequently, some simple counts and quality ratings capture reasonably well differences from subject to subject and group to group. If one has in mind the complexity of arguments that occur in professional psychology or philosophy journals from time to time, with their many subarguments and the different sorts of technical objections that might be raised, the present scoring system would indeed be simplistic. But for our subjects' characteristic responses, it serves nicely.

One scale was discarded for poor interjudge agreement. Seven scales remained on which the subsequent analysis focused. The scales were as follows:

- *Sentences*. A count of the number of sentences in an argument provided a simple measure of complexity. Compound sentences were counted as more than one sentence. For example, a sentence "Money would probably make the teachers work harder, but the students' motivation to learn is the key to improvement," would be counted as two sentences, one about teachers and one about students' motivation.
- *"Myside" and "otherside" arguments*. Judges counted arguments in each subject's performance. What we are calling an argument might better be described as a line of argument. This is a distinct way of arguing a point relevant to particular question. For example, the assertion, "A year of mandatory social service by 19 year olds would provide them with valuable work experience and many previously neglected community jobs would get done," would be counted as two lines of argument. "Myside" arguments supported the subject's initial point of view while "otherside" arguments opposed it. It should be recalled that the issues were chosen because they were vexed, that is, arguable on both sides from several perspectives. Consequently, subjects' failure to mention reasons contrary to their own position evinced neglect of or inability to address the other side rather than absence of reasons on the other side. Irrelevant arguments, arguments that did not directly address the question asked, were not counted in the "arguments" category. In summary, the lines of argument measures provided indices of thoroughness in a subject's argument, revealing how many distinct points pro and con a subject addressed.
- *Bothsides*. Myside and otherside arguments were summed to arrive at the "bothsides" arguments measure.
- *Elaborations*. Elaborations – steps within each line of argument – were also counted. For example, "A year of social service would provide 19 year olds with valuable work experience. They would learn job skills and also be exposed to different kinds of careers," would be counted as one line of argument and two elaborations. This measure provided an indication of the level of detail, or depth, of an argument.
- *"Myside" and "otherside" quality ratings*. Each performance was given two ratings by each judge as regards overall quality, one reflecting the treatment of the subject's side, and a second for the treatment of the other side of the case. This holistic rating used a 5 point scale ranging from 0 to 4. On this scale, 0 stood for no response at all (which occasionally happened on the other side of the case); 1 for a reassertion of the claim or its contrary or very simplistic appeal ("Nobody would say that"); 2 for somewhat weak reasons given, for instance personal examples; points of questionable truth or relevance; 3 for some true and relevant support; 4 for most major arguments on the topic given with good elaboration and connection to other issues. The judges could choose intermediate points like 3.4. Using this scale, the judges could incorporate considerations of soundness of argument not captured by the mere counts mentioned earlier.
- *Myside ratio*. In order to assess evenhandedness, a derived score was used: the total of myside arguments divided by the sum of myside plus otherside arguments. A myside ratio of 1 describes a performance comprised entirely of arguments supporting the subject's side of the case, with no opposing points. A myside ratio of .5 describes a performance with an equal number of myside and otherside arguments.

3.2 Results

3.2.1 Validity of the measures

Three kinds of correlations were calculated to check the validity of the measures: between the two judges, between the myside and otherside holistic quality ratings and the myside and otherside argument counts, and, for groups doing two issues pre and post, between scores on the first and second issues. Interjudge agreement correlations were calculated by group to ensure scoring consistency during the study and then were calculated for all groups combined. The overall interjudge correlations ranged from .70 for myside arguments to .93 for otherside arguments (N=141, p<.0005). Correlations between holistic quality ratings and lines of argument were .67 for myside and .89 for otherside (N=444 arguments, p<.0005), revealing a high degree of consistency between the qualitative and quantitative measures. Finally, first issue / second issue correlations proved the questionnaire to be

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

reliable, as all but one were significant at the .005 level or better. The nonsignificant correlation was for rating of the other side argument on the pretest. The other correlations ranged from .35 ($p<.005$), to .61 ($p<.0005$), N=64.

Although interjudge correlations were strong, in pooling data over the two judges, subjects' scores were normalized to erase any systematic scoring differences between them that would not be reflected in the correlation coefficients (for instance, one judge systematically scoring a little higher than the other). Scores were also normalized by issue to eliminate differences due to any differential accessibility of the issues.

3.2.2 Impact of instruction

Detecting the impact of instruction called for comparing pretest and posttest performances. Table 2 summarizes the results, displaying the pretest scores, the gain scores (posttest minus pretest) and the significance of the pre-post differences as calculated with t-tests matched by subject.

Instruction in debate and the critical liberal arts program yielded gains in myside arguments as measured by counts and quality ratings, while not significantly affecting otherside performance, considered here both in terms of number of otherside arguments and in terms of the ratio of myside to otherside arguments. The graduate education students and the law students did not show such gains, the education students displaying a modest gain in bothside arguments while the law students showed no differences at all in these measures and a significant loss on elaborations and sentence count. In sum, these programs yield sporadic results which occur in the area of bolstering subjects' own positions.

The reasoning classes consistently had the effect shown in Table 2 of boosting attention to the other side of the case. Students produced more, but not significantly so, myside arguments, while producing significantly more and more highly rated otherside arguments, resulting in a more even-handed treatment of an issue as reflected in the myside bias measure.

In Table 2, the scores of the three different groups who participated in our reasoning classes (Table 1, last three rows) are combined (Table 2, last column). This reflects the fact that the pattern of results for each of the classes was virtually the same in terms of gain. In order to determine the role that IQ plays in acquiring new reasoning skills, the performances of two groups which differed significantly in IQ as measured by the Quick Word Test were compared with a repeated measures ANOVA. Though the prescores of the groups differed significantly, the gain scores did not.

4 Experiments 8–10: The Role of Capacities and Metacognitive Repertoire in Reasoning

As discussed in the introduction, substantial correlations between general intelligence and informal reasoning scores in prior research along with the relative unresponsiveness of informal reasoning to conventional education suggest that informal reasoning might be limited by matters of intellectual capacity. The findings reported in the previous section already challenge this picture. In the present section, we report three studies that continue to cast doubt on such a capacity view.

4.1 Experiment 8: Scaffolding Informal Reasoning

As already emphasized, quite commonly subjects' arguments on an issue prove to be sparse and strongly biased toward their own side of the case. Such shortfalls might simply reflect a failure of subjects to deploy their reasoning capacities fully and systematically. This question might be tested by providing subjects with general guidance as they reason, in effect "scaffolding" their efforts.

Scaffolding is a term commonly used to characterize how a skilled individual, typically older, can help a less skilled one to manage a performance by supporting the learner's efforts at points where the learner is at a loss, while hanging back wherever the learner proves able (Greenfield, 1984; Rogoff & Gardner, 1984). Scaffolding as an experimental procedure has been used to demonstrate the capabilities of subjects when another provides a higher order structure for the task in question (Heller & Reif, 1984; Perkins & Martin, 1986). We designed an experiment to determine how fully subjects might develop their arguments with the help of scaffolding, after offering initial arguments on their own. Would their initial arguments prove to be close to some sort of capacity ceiling, with scaffolding having little impact, or would they develop their arguments much more extensively in response to scaffolding, suggesting that with a better metacognitive repertoire they could construct much richer arguments on their own? With this question in mind, all scaffolding was generic in character, never providing specific help with the particular issue being discussed.

4.2 Method

4.2.1 Subjects

There were 20 subjects, junior and seniors in high school, balanced for sex. An effort was made to obtain subjects of varying general intelligence by asking for subjects in different sorts of high school classes. Subjects were paid a nominal fee for participating.

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

TABLE 2: Pretest and gain scores for all groups, Experiments 1–7.

	Debate class, N=33	Critical liberal arts, N=35	Educational design, N=27	Law school, N=41	Experimental reasoning, N=37
Measures	pre / gain	pre / gain	pre / gain	pre / gain	pre / gain
Myside arguments	2.7 / 1.0**	2.4 / 0.6*	3.0 / .04	4.1 / 0.1	2.6 / 0.4
Otherside arguments	0.5 / 0.0	0.9 / 0.4	1.1 / 0.5	0.9 / 0.3	1.2 / 1.1***
Bothsides arguments	3.2 / 0.9*	3.4 / 0.9**	4.0 / 0.9*	5.0 / 0.4	3.9 / 1.5***
Myside rating	2.1 / 0.8***	3.3 / 0.5**	3.7 / 0.0	4.4 / 0.0	2.2 / 0.2
Otherside rating	0.5 / −0.1	1.2 / 0.6	1.2 / 0.5	1.1 / 0.0	0.9 / 0.9***
Elaborations	3.3 / 2.6**	4.3 / 0.3	5.4 / −0.2	6.9 / −1.6*	3.5 / 0.2
Sentences	7.5 / 4.2**	14.1 / 0.7	14.4 / −0.5	18.0 / −3.5**	8.7 / 1.6*
Myside bias (ratio)	0.8 / 0.0	0.8 / −0.1	0.7 / 0.0	0.8 / 0.0	0.8 / −0.2***

* p<.05, ** p<.01, *** p<.001, one-tailed.

4.2.2 Procedure

An investigator worked with each subject individually, guiding the subject through a multistep interview. After giving their names, ages, and addresses, and signing consent and payment forms, subjects were asked to list rules, or pieces of advice for thinking well, any things that you should "make sure to do or to watch out for when you have to make an important decision." When subjects indicated that they had finished, they were pressed to try to add five more bits of advice to their lists. This was to encourage the subjects to explore more thoroughly their conscious metacognitive repertoire of reasoning tactics.

Subjects were then presented, both in writing and orally, with one of two issues to think about. The issues were:

- Would providing more money for public schools significantly increase teaching and learning?
- Would a nuclear freeze significantly reduce the probability of world war?

Subjects were asked for a snap judgment yes or no and a rating of their confidence in that judgment. They rated confidence on a 4 point scale, 1 standing for not at all confident and 4 for very confident. They also were asked to estimate the amount of time they had thought about the issue prior to the experiment. Subjects were then asked to list thoughts about the issue, not in order to persuade someone but to show what it is to think well about the issue. Fifteen minutes were allowed for this phase, termed the "initial arguments".

During the remainder of the interview, the investigator elicited expansions, elaborations and refinements of the initial arguments, to determine how far beyond it generic questions could lead the subjects. The scaffolds employed were as follows.

- *Otherside scaffold*. Prior work had shown that reasoners tend to neglect the other side of the case. Consequently, if arguments for only one side of the case had been given, the subject was asked to think of reasons for the other side.
- *Quota scaffold*. Prior work had shown that reasoners tend to underexplore issues. Consequently, subjects were given a simple quota scaffold: "Try to come up with three more reasons for saying 'yes' to the question . . . Now try to think of three more reasons for saying 'no'."
- *Preserving the question*. Our prior work suggested that sometimes subjects lose track of the question. Accordingly, subjects were asked to restate the question, after which they were allowed to look at the original written version and to correct themselves if necessary.
- *Relevance scaffold*. Prior work suggested that subjects sometimes treated as relevant reasons that were not in fact relevant to the exact question. Therefore, the interviewer then asked subjects to go over their reasons and indicate which ones addressed the exact question. In cases where subjects did not identify irrelevant reasons as such, an example was given of how people sometimes make mistakes about relevance and the subject was asked to recheck the reasons. If irrelevant reasons were still unnoticed, another example of irrelevance was given and the subject was asked to recheck the reasons again. At this point, subjects were invited to write down any new reasons that had occurred to them, although in fact hardly anyone had more reasons to offer at this point.
- *Organizing and prioritizing scaffold*. Prior work had shown that it is often hard to tell which reasons reasoners consider their main reasons for a particular con-

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

clusion and sometimes even whether an argument was "pro" or "con". Therefore, the interviewer asked the subjects to organize and prioritize their arguments. Arguments were labeled as being "yes" or "no" in relation to the question, and then evaluated in terms of importance. Subjects indicated which reasons they considered "main" and which were "points of support". Subjects then indicated which reasons were the first and second most powerful ones for the "yes" and the "no" sides.

- *Critiquing scaffold*. Subjects were then guided through a thorough critique of each of the top two reasons for both sides. First, subjects were asked how confident they were of a particular reason. They were then scaffolded in attempting to disconfirm the reason from two different perspectives and given the opportunity to revise their original confidence rating. If no revision was made, the interviewer asked why not.

Finally, the interviewer reviewed with subjects their original snap judgments and asked if the subjects wanted to change them. The final judgment was recorded and the same procedure was repeated with the original confidence rating. If no changes were made, the interviewer asked why not. The Quick Word Test was then administered.

4.2.3 Scoring

It is worth noting here that the issues picked for the experiment were rich, vexed, and accessible; they could be argued from several pro and con common sense perspectives dependent only on everyday knowledge. Therefore, few arguments overall or neglect of one side of the case indicated a shortfall in a subjects' performance, not a ceiling effect due to the issue. In summary, the argument counts provided a measure of the breadth of performance – how thoroughly had a subject explored the various possible lines of argument pro and con?

The written responses of the subjects were scored by two judges independently. When the scoring was completed, the scores of the two judges were tested for correlation. Several measures were discarded because of poor interjudge correlation. The remaining scales were:

- *Total advice*. A count of the total number of bits of advice about reasoning that a subject offered.
- *Argument counts*. The judges counted myside and otherside arguments. "Myside" arguments were those that supported the subjects' original snap judgments while "otherside" arguments were those that opposed. Myside and otherside arguments were summed to give a "bothsides" measure. Separate counts were made for the initial arguments, the otherside scaffold, and the quota scaffold. "Argument" here is used to denote what might be called a "line of argument". For instance, regarding the issue of school funding and quality of education, the response, "More money could be used to hire more teachers and improve the facilities," was scored as two arguments because two ways of addressing the issue are evident: hiring more teachers and improving facilities.
- *Elaborations count*. Details and minor points of support for arguments were scored as "elaborations". A response such as, "Schools could use more money to have more classes, like art classes and classes for special students," was credited with one argument – more classes – and two elaborations – art classes and special classes. The elaboration count provided an index of the level of detail of a subjects' performance.
- *Accuracy of restatement*. Judges scored subjects' restatement of the question for accuracy on a four point scale, 0–3; 0 represented a restatement very far from the original, 1 a restatement that substituted a different cause or effect for that related in the original (all the issues concerned causes and effects), 2 a restatement with a minor discrepancy, and 3 a restatement with no difference in meaning from the original. This measure gave an indication of subjects' competence in keeping the exact question in mind.
- *Relevance*. The judges reviewed subjects' classifications of reasons as relevant or irrelevant and decided whether their classifications were appropriate.
- *Prioritizing arguments*. The judges considered each argument the subject had singled out as strongest or next strongest, myside and otherside. Using a three point scale, the judges rated whether it was, in their view, (0) indeed first or second strongest as the subject said, (1) one of the strongest two, but the opposite one, or (2) not among the strongest two.

4.3 Results

4.3.1 Matters of validity

Sufficient time to think. All but one subject reported having adequate time to think about the issue during the initial argument. The average time taken was about 10 minutes (9.6). The subject who said he needed more time turned in a superior performance, time limitation notwithstanding.

Reliability of scoring. To assess the reliability of the measures, interjudge correlations were calculated. Two kinds of measures had to be dropped because of nonsignificant interjudge correlations. In rating the arguments subjects put forth as second strongest, the judges did not achieve adequate interjudge reliability on the rating scale. Ratings of subjects' reasons in the critiquing section, where judges were supposed to assign a percentage indicating pre-critique confidence warranted by the reason under consideration, also

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

TABLE 3: Initial performances and gains from scaffolding (Exp. 8).

Measures	Initial performance	Scaffolded performance	Gain	Effect size
Myside arguments	3.5	3.8	109%	2.3
Myside elaborations	4.6	3.5	76%	0.9
Otherside arguments	0.7	4.9	700%	4.4
Otherside elaborations	0.3	4.0	1333%	6.1

had to be dropped because of nonsignificant correlations. Among the remaining measures, the correlations ranged from .39 (N=20, p<.05) for the count of new otherside elaborations on the quota scaffold to .92 (N=20, p<.001) for initial otherside arguments. Generally, the interjudge correlations were about .7.

4.3.2 Initial arguments

Subjects initially produced 3.45 myside arguments and 4.6 myside elaborations. The other side of the case was represented by .73 arguments and .28 elaborations. The ratio of myside arguments to total arguments was .83. This could not be considered a strong performance in light of the complexity of the issues, especially with reference to the other side of the case.

4.3.3 Do generative capacities limit performance?

The initial arguments advanced by subjects proved limited in accord with prior research conducted as part of this line of investigation. The various measures obtained allowed investigating the causes of this shortfall. First of all, we consider factors that might have influenced the general cognitive fluency and, more broadly, efficiency of the subjects.

Closeness to ceiling performance. In general, subjects' response to the scaffolding showed that their initial arguments fell well short of capacity. Scaffolding produced large gains for both arguments and elaborations. The impact of scaffolding was investigated by summing additional arguments and elaborations that occurred during the otherside and quota scaffolds. Table 3 summarizes the results.

Of particular interest is the very large gain in otherside arguments, an increase of 700% as a result of scaffolding, with an effect size (gain considered in ratio to the standard deviation of the original performance) of 4.4. Otherside elaborations increased by 1333% with an effect size of 6.1. Myside arguments showed a more modest gain, but doubled nonetheless, while myside elaborations increased by 76% for an effect size of 0.9. These results show that subjects could have developed their arguments substantially more than they did; they were not operating close to any "generativity ceiling."

It is useful to put the effect of scaffolding on the balance of myside and otherside arguments in percentage terms. Barely 16% of the initial arguments were for the other side, while 58% of the scaffolded arguments were, bringing the balance of the overall performance to 45% otherside arguments, a final percentage that did not differ significantly from 50%. In other words, scaffolding had the effect of correcting bias, as measured by sheer number of arguments. However, its effect on subjects' opinions was much more modest, as will emerge later.

Initial fluency in relation to response to scaffolding. One might suggest that, even if subjects' initial argument counts did not reflect ceiling performance, they reflected capacities in some manner. Such a view would predict a positive correlation between initial arguments and arguments in response to the quota scaffold, a highly reliable prompt for eliciting more arguments. Correlation coefficients between the initial arguments and the quota question, for the measures myside arguments, otherside arguments, and myside elaborations, were nowhere near significance. A significant correlation appeared for myside elaborations, .59, N=20, p<.01. This result certainly gives little support to the notion that initial responses reflect capacity strongly.

Relation of performance to general intelligence. The obvious capacity measure available for the subjects was their scores on the Quick Word Test, a short form IQ instrument. One might forecast positive correlations between Quick Word scores and both unscaffolded and scaffolded performance. Instead, a more complex picture emerged. The data disclosed a correlation between Quick Word scores and initial myside arguments, r=.53 (N=20, p<.02 two-tail test), while showing virtually no correlation (r=−.18, n.s.) for initial otherside arguments. The same trend held, weakly, for the scaffolded measures, Quick Word scores correlating significantly with myside arguments (r=.64, N=20, p<.01, two tail) and nonsignificantly for otherside arguments (r=.40, n.s., two tail). These figures argue that people with high capacity in the psychometric sense of intelligence testing invest their intelligence conservatively to support their own positions. This inclination appears, although to a less extreme degree, even when scaffolding encourages more exploratory thinking.

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

4.3.4 Do discriminative capacities limit performance?

Productivity measures like argument counts aside, one might propose that subjects' arguments suffered from incapacities to maintain focus on the issue, discriminate relevant from irrelevant reasons, or discriminate stronger from weaker reasons. The data gathered spoke to these matters as well.

Restating the question. The mean rating for subjects' restatement of the question was 2.1, just over the "minor mistake" point on the 0–3 point scale. This indicated that, for the most part, subjects did not have difficulty retaining what the question was.

Correlations between Quick Word scores and ability to restate the question disclosed a possible capacity element in maintaining a representation of the issue. Ability to restate the question correlated positively with Quick Word scores, r=.49, p<.05 (one-tail test, N=16). Considering that Quick Word scores were found to correlate negatively, although nonsignificantly, with initial otherside arguments, one would predict that restating the question might correlate negatively with initial otherside arguments. In fact, such a relationship was found, r=−.43, p<.05 (one-tail test, N=16). Consequently, maintaining the issue has some association with more biased reasoning. No correlation was found between restating the question and initial myside arguments (r=.005).

One would predict that keeping the question in mind would be important for working with the scaffolds. A medium-low, positive correlation (r=.34, n.s.) was found for myside arguments and a significant positive correlation was found for otherside arguments, r=.48, p<.05 (one-tail test, N=16), confirming the prediction.

Discriminating main and support points. Both subjects and judges sorted main from support points for initial and scaffolded arguments. The subjects agreed with the judges 62% of the time. Considering that this discrimination certainly has room for subjective differences, it does not appear that this discrimination posed serious problems for the subjects on the average. Nonetheless, some capacity contribution from general intelligence appeared: percentage agreement between subjects and judges, taking the judges as a standard, correlated with Quick Word scores (r=.44, p<.05 one-tail, N=20).

Discriminating relevance. Both subjects and judges assessed whether each main point bore on the issue as stated. Subjects agreed with the judges 78% of the time. Again, it does not appear that this discrimination presented the subjects with difficulty. Again, however, some capacity contribution from general intelligence appeared: percentage agreement between subjects and. judges correlated with Quick Word scores (r=.42, p<.05 one-tail, N=20). .

Discriminating stronger from weaker arguments. It will be recalled that the judges did not achieve adequate inter-judge agreement on their ratings of the reasons subjects advanced as their second strongest. Consequently, the analysis focused on the judges' ratings of the reasons subjects advanced as their strongest. On the 0-2 point scale, for myside arguments, the mean scale value was .95 and for otherside arguments .55. In other words, on the average for myside arguments, the judges thought that the argument singled out by the subject as strongest was second strongest; for otherside arguments, the judges agreed fully with the subject somewhat more often. So it appears that discriminating stronger from weaker reasons was not a major problem for the subjects.

Turning to correlations with Quick Word scores, a peculiar pattern emerged. Good selection of myside arguments was associated with less intelligence (r=.52, p<.02 two-tail, N=20), while good selection of otherside arguments was associated with more intelligence (r=−.53, p<.02 two-tail, N=20). A plausible interpretation is that intelligence does improve selectivity in general, but that more intelligent subjects also generate considerably more myside, but not otherside, arguments. Therefore, selecting among a larger pool of arguments, they were more likely to disagree with the judges' choices. Accordingly, the myside correlation can be considered an artifact.

To summarize results from this section, discriminative capacities did not appear to hamper performance very much. To be sure, usually there was some correlation between the adequacy of subjects' discriminations, using the judges as a standard, and Quick Word scores. This demonstrates a capacity element in discriminations. However, the average level of accord between subjects and judges was generally quite high, arguing that, despite some relation to general intelligence, discriminative capacities did not function as a sharply limiting factor.

4.3.5 Does the lack of substantive gains in arguments or revisions of stance limit performance?

The results examined so far argue that subjects had the capacity to construct substantially more elaborate models of situations than they did in their initial arguments and that their further reasoning was not substantially limited by difficulties of discrimination. It might be, though, that their scaffolded reasoning was an empty exercise, not yielding important new reasons nor leading them to change their initial views or alter their confidence. We examine these issues now.

Importance of scaffolded reasons. Given that scaffolding yielded large percentage gains in argument counts, did the scaffolding provoke subjects to think of significant arguments or did it just encourage verbiage? To investigate this matter, we used subjects' selection of important reasons as a gauge relevant to subjects of the significance of the arguments provoked by scaffolding: We asked how many subjects chose their first and second most powerful myside and otherside arguments from among those generated by the quota scaffold. There were relatively few instances where subjects

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

chose a myside argument from the pool of arguments generated by scaffolding; on the average, .38 out of the 2 top myside arguments were selected from the scaffolded arguments. Apparently people have little difficulty marshalling what are subjectively considered to be strong myside arguments. The case was markedly different for otherside arguments; subjects chose 1.55 of their top two otherside arguments from among those generated by scaffolding.

One could argue that a quota scaffold might be worthwhile for a person relatively bereft of ideas, but that a productive thinker, being able to generate a varied pool of ideas on his or her own, would benefit little from scaffolding. Following such reasoning, one would predict a significant negative correlation between the number of arguments produced initially and the number of first and second arguments selected from the pool of scaffolded arguments. Correlation coefficients were calculated to test this prediction. A low, nonsignificant, negative correlation (−.185) obtained for myside arguments, while a somewhat higher, nonsignificant, positive correlation (.317) obtained for the otherside. The tests did not support the idea that quota scaffolds are of less benefit to productive thinkers than to not-so-productive thinkers. Furthermore, as mentioned earlier, the number of arguments generated by the simple quota scaffold, a highly reliable prompt for eliciting more arguments, did not correlate significantly with the initial argument counts. No correlation was found for myside, otherside, or bothsides measures.

Influence of scaffolded reasoning on position and confidence. Perhaps the new and relatively strong otherside arguments elicited by scaffolding carried little weight in subjects' minds. In fact, 15% of the subjects (3 out of 20) reversed their original positions at the end of the sessions. Recall that subjects rated their confidence at the beginning and at the end of the session, using a 1 to 4 scale, 1 for not at all confident and 4 for very confident. The average initial confidence rating was 3.2, or somewhat confident. The data showed a change in the amount of confidence, represented by the absolute value of the initial confidence rating subtracted from the final confidence rating, of 1.1 during the protocol. Taking direction of change into account, there was little average change: 0.6 in the direction of diminishing confidence in the original position, including the subjects who changed positions in the figures. In fact, of those who did not change their position, just 5 changed confidence: 4 became more confident and 1 less.

One would expect that the more confident people are, the less likely they are to change position at all. Indeed, initial confidence correlated negatively with inclination to change mind, r=−.39, p<.05 (one-tail test, N=19). Along the same lines, one would predict that the less confident people are, the more inclined they would be to generate opposing schemas to try to sort the matter out. In fact, the data showed such a trend; confidence correlated negatively with the percentage of otherside arguments in the initial arguments, although not significantly, r=−.36, n.s. The more confident subjects were less productive overall, initial confidence correlating negatively with total initial bothsides arguments, r=−.55, p<.01 (two-tailed test, N=20). The question arises as to the degree to which initial confidence impedes one from generating in response to scaffolding. A correlation coefficient calculated to test this matter was low and negative (r=−.19, n.s.) suggesting that high confidence is not a significant barrier to extending one's argument via scaffolding.

However, the correlation of intelligence with defensive thinking is corroborated by correlation coefficients calculated between Quick Word scores and the mind change measures. Quick Word scores correlated negatively with the propensity to change one's mind in either direction, r=−.55, p<.02 (two-tail test, N=19). The correlation between Quick Word scores and the mind change variable that included direction showed a significant, positive correlation of .69, p<.01 (two-tail test, N=19), suggesting a tendency for more thinking to produce greater confidence in the more intelligent, despite the generation of a greater proportion of otherside arguments. Considered together, the two tests suggest that the students with higher mental capacity in the sense of general intelligence are less inclined to change their minds at all, and that when they do, they tend to become more confident in their positions.

4.3.6 How did subjects' initial knowledge and metacognitive repertoire influence performance?

Prior thought. As mentioned earlier, issues were chosen for their currency and vexedness. How familiar were the subjects with the issues? The average length of prior thought given by subjects to the respective issues reasoned about was 30 minutes, according to their own statements. Perhaps prior thinking about the issues had yielded a richer array of arguments to mention and/or an entrenched stance. The latter might help to explain the relative reluctance of subjects to change their confidence. Correlation coefficients calculated to test these possibilities were low and nonsignificant. Length of prior thought given to the issue did not correlate strongly with confidence, mind change, or production of arguments.

Explicit metacognitive repertoire. Scaffolding amounts to providing a subject with metacognitive guidance. But, of course, subjects entered the experiment with some explicit metacognitive knowledge about how to reason well. Did those better equipped with metacognitive rules in fact perform better initially, and might they therefore benefit less from scaffolding, because it presses them to do what they already do for themselves?

Subjects who produced more advice for thinking well did not prove to be more productive on the initial arguments. The correlations between advice and myside and otherside initial arguments were nonsignificant, although the pattern was in-

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

teresting: a positive correlation between advice and myside scores (.37) and a negative correlation between advice and otherside scores (-.38). In keeping with this, amount of advice proved to be negatively correlated with the percentage of initial otherside arguments, r=−.44, p<.05 (two-tail tailed test, N=20). Thus, at least according to the measures of this study, the capacity to produce rules for thinking is not a predictor of good performance in the sense of balanced reasoning.

With this finding in mind, one may wonder whether the advice count was counting reasonable versus empty advice. In general, the advice tended to be sensible if not deep. Common injunctions were to consider all viewpoints, get the facts, talk to someone with experience or expertise, be true to your self but considerate toward others, take your time, find a quiet place, be logical. Five out of twenty subjects explicitly suggested considering pros and cons. In summary, the advice count did seem to reflect repertoires of reasonable advice.

More fluent advice givers were better advice takers. There were significant correlations between the amount of thinking advice given and myside and otherside arguments on the scaffolded performance. For myside arguments r=.48, p<.05 (two-tail test, N=20) while for the otherside r=.52,; p<.02 (two-tail test, N=20).

On the other hand, a nonsignificant negative correlation emerged between giving generic advice and change in confidence in either direction (that is, absolute value of difference between final and initial confidence; r=−.37, n.s., N=19). The correlation between amount of advice and change of position (final confidence minus initial confidence) was .66, p<.01 (two-tail test, N=19) indicating that when the prolific advice givers did change their positions, even though they were generative with the otherside scaffolds, they were, in the end, inclined to become more confident.

One possible interpretation of this pattern of results is that the more generally intelligent subjects in the conventional psychometric sense had acquired or could retrieve more general advice, and, independently, tended more to bias. Their advice did not do them much good. Another factor may be that the subjects did not regard their advice as bearing on a question of belief but only of decision. We asked for advice about decision making, but the issues were posed as matters of truth or falsity. To be sure, to decide on the truth or falsity of a proposition is, as a point of logic, to make a decision, but the subjects may not have seen it as such.

4.4 Experiment 9: Impact of Simple Requests on Reasoning

One possible reading of the results for the scaffolding experiment just reported might hold that the rather elaborate scaffolding went much further than necessary. Subjects would respond equally well to a fairly straightforward request to develop their arguments more fully on both sides of the case, without quotas, reviews of the exact issue, and so on. If this were so, the sort of metacognitive knowledge needed to provoke a substantially more developed reasoning performance would be much simpler than that suggested by the scaffolding process. An experiment in which subjects were pretested, explicitly requested to provide more arguments on both sides, and then posttested directly examined this possibility.

4.5 Method

4.5.1 Subjects

There were 20 high school subjects, ranging from 15 to 18 years old, balanced for sex and with some effort to ensure a spread of general intelligence by drawing from more and less academically able populations.

4.5.2 Procedure

Written protocols were administered in classrooms to two or more subjects at once. After signing consent and payment forms, subjects were briefly informed of what the whole task entailed, namely, taking a pretest, a short period of instruction, a posttest, and a vocabulary test. The pretest was then administered.

The pretest collected subjects' names and ages, then presented either the funding for public schools issue or the nuclear freeze and world war issue, as discussed earlier. The pretest then asked for the subject's snap judgment, indication of confidence in the snap judgment, interest in the question, and amount of prior thought given to the question. The pretest then asked subjects to write out their thoughts about the issue as thoroughly as possible. After expressing their arguments, the subjects were asked to note their current position on the issue and give another confidence rating.

The completed pretests were collected and the posttests were passed out. Each subject received a posttest with a different question from the question he or she had addressed on the pretest. Subjects were directed to look at their protocols while the experimenter gave instructions, emphasizing and elaborating on the instructions in the protocol. Subjects were asked outright to give as many reasons as possible, even insignificant ones. They were encouraged to give reasons on both sides of the case. These same points of emphasis and elaboration were printed on the protocol next to the space provided for the various tasks. In all other ways, the posttest was the same as the pretest.

When the posttest was completed, the protocols were collected and the subjects completed the Quick Word Test.

4.5.3 Scoring

The written protocols were scored by two judges working independently. When the scoring was complete, correla-

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

TABLE 4: Gains in demand study (Expt. 9) compared to gains in scaffolding study (Exp. 8).

	Myside arguments		Otherside arguments	
Measures	Demand	Scaffold	Demand	Scaffold
Original arguments	3.1	3.5	0.8	0.7
Gain	−0.6	3.8	1.2	4.9
% Change	−19%	109%	150%	700%
Effect size	0.3	2.3	1.3	4.4

tions were calculated to measure interjudge agreement. The measures scored included myside and otherside arguments.

4.6 Results

Sufficient time to think. Eight subjects who took the pretest by special arrangement outside of class time were under no pressure to finish by any particular time. Some of twelve subjects who took the protocol during class time may have been rushed to finish the posttest.

Reliability of scoring. Interjudge reliability was calculated. For myside arguments, the correlation was .78, p<.005, N=14. The correlation for otherside arguments was .76, also significant at p<.005.

Effect of explicit demand on performance. When subjects were simply asked for more arguments and for otherside arguments explicitly, performances did improve, but not in a global fashion. Otherside arguments increased significantly from 0.8 to 2.0 (p<.005. two-tail matched t-test). However, myside arguments actually decreased, from 3.1 arguments to 2.5 (n.s.). Better balance between myside and otherside consequently in part reflected the decrease in myside arguments. The significant increase in otherside arguments shows that the improvement in otherside arguments in consequence of scaffolding reported in the previous study in part reflected an elementary press for more otherside arguments.

Keeping in mind that some of the subjects may have rushed a bit on the posttest, still the scaffolding study appeared to have a much stronger impact on subjects' arguments. Table 4 compares key results from the two. Although both the scaffolding study subjects and the demand study subjects started out with about the same number of otherside arguments, the scaffolded subjects show much larger gains. The results suggest that an elementary press for "more" of the sort examined in the present study does not really substitute for the more elaborate scaffolding of the previous study in eliciting a substantially more developed argument about the issue.

4.7 Experiment 10: Influence of General Intelligence on Otherside Arguments

The scaffolding experiment disclosed an intriguing relationship between general intelligence and reasoning performance: Individuals with greater general intelligence as measured by the Quick Word test developed myside arguments more fully but not otherside arguments. In effect, they invested their intelligence in buttressing their own case rather than in exploring the entire issue more fully and evenhandedly. The investigators realized that a reanalysis of data collected during the studies of the impact of instruction on informal reasoning reported earlier could probe this relationship further.

4.8 Method

4.8.1 Subjects

The reanalysis involved a subset of the subjects discussed under Experiments 1–7 above. The subjects included 99 college students from the three programs described earlier, as summarized in rows 2–4 of Table 1 – freshmen in a small private liberal arts college that stressed critical thinking skills, graduate students enrolled in an interactive technology program at a well-regarded school of education, and first year law students at a prestigious law school. Also, the reanalysis included 36 high school students from a debate program and reasoning course as described earlier, as summarized in rows 1 and 5 of Table 1. The reanalysis dropped four subjects from the college sample and five from the high school sample for reasons of missing data or gender balance.

4.8.2 Procedure

The subjects, at the beginning and end of courses of instruction, responded to pretests and posttests as described under Experiments 1–7, but only pretest results were used for this reanalysis, to support comparisons with the results from Experiment 8 on scaffolding, which did not involve an instructional intervention. The critical liberal arts undergraduates, the law students, and the high school debate students each reasoned about one issue current at the time;

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

the education graduate students and the high school students reasoned about two. A counterbalanced design was employed to ensure that the order in which the questions were answered did not affect the results.

4.8.3 Scoring

The protocols were scored for myside and otherside arguments, elaborations, and quality ratings by two judges, as described earlier. Correlations for assessing interjudge agreement were calculated. The data were normalized to eliminate any systematic differences between the judges and the subjects' responsiveness to the different issues.

4.9 Results

Sufficient time to think. Subjects were allowed to do the protocols at their leisure. Time constraints were not a problem.

Reliability of scoring. Interjudge correlations ranged from .90, p<.001, for otherside arguments to .76 for rating of myside performance, p<.001, N=61.

Correlations with general intelligence. For the college sample, Quick Word scores correlated positively with myside arguments (r=.37, p<.001 N=99, two-tail), but virtually not at all with otherside arguments, –.08. Judges' ratings of performances showed the same relationship – myside correlated with Quick Word scores significantly at .39, p<.001, while otherside showed virtually no correlation, r=.10, n.s. For the high school sample, correlations did not reach significance between Quick Word scores and either myside or otherside arguments.

In summary, the results for the college students showed the same pattern of myside positive correlation but no otherside correlation with general intelligence that emerged in the scaffolding study. The correlation coefficient was not as large as in that study, however. For the high school samples, this relation did not obtain; we have no ready explanation why it should appear in two cases and not in this third.

5 Experiments 11–12: Methodological Studies

5.1 Experiment 11: Written versus Oral Reporting of Reasoning

Our principal method of investigation asked subjects to produce written arguments. It is natural to wonder whether oral arguments might not allow subjects to explore the ramifications of an issue more fluently, particularly with younger subjects who might not write with ease. On the other hand, one might suppose instead that the explicit formulation of ideas in black and white would actually assist subjects to marshal their thinking. With these alternative viewpoints in mind, we conducted an experiment directly addressing the comparative responsiveness of subjects under conditions of written and oral reporting of arguments.

5.2 Method

5.2.1 Subjects

There were 30 subjects balanced for sex: 16 high school students and 14 first year college undergraduates. With the high school subjects, care was taken to have a normally wide range of intellects. Subjects were paid a nominal participation fee.

5.2.2 Procedure

Protocols were administered to subjects individually. After signing a consent form, subjects responded orally or in writing to either the nuclear freeze or the school funding questions:

- If the United States and the Soviet Union signed a nuclear freeze, would the possibility of world war be reduced?
- Would increased funding to public schools significantly increase teaching and learning?

Subjects were asked to list reasons for answering the respective questions yes or no, being careful to list all reasons that came to mind. After completing the first issue, each subject was asked to perform the task again, but with whichever question he or she had not already addressed and in whichever form, written or oral, he or she had not used. A counterbalanced design ensured that the order of issue and form of response would not influence the results. Oral responses were tape-recorded.

5.2.3 Scoring

The oral protocols were scored directly from tapes. All of the protocols of the high school group were scored by two judges working independently. A subsample of the college group protocols was scored by two judges, also working independently. Interjudge correlations were calculated after scoring was completed.

5.3 Results

The analysis focused on the four measures: myside arguments, otherside arguments, bothsides elaborations and sentences.

Reliability of scoring. For the high school group, interjudge correlations ranged from .70 (N=30, p<.0005, one-tail test) for myside arguments, to .97 (N=30, p<.0005, one-tail test) for sentences. The average correlation was .84. For the college group, correlations ranged from .49 (N=7, n.s.)

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

for myside arguments, to .93 (N=7, p<.005, one-tail test) for otherside arguments. The average correlation was .75.

Correlation of oral and written performances. Correlation coefficients calculated to test the relationship of oral to written performances showed significance for all measures of the high school sample. The correlations ranged from .48 for myside arguments to .82 for otherside arguments. For the college sample, only sentences disclosed a significant correlation (.56).

Group differences between oral and written performances. Apart from the matter of correlations, one can ask whether oral or written performances in general yielded higher scores. In general, there was no difference between the two kinds of performances of the two groups. The high school group produced significantly more sentences in the oral medium, but this is not surprising considering it is easier for most people to speak a sentence than to write one. The college group produced more sentences also, but at a nonsignificant level.

Though it might seem that the greater number of sentences in the oral condition would include more substantive points, this did not occur. On the contrary, the high school subjects produced more, though not significantly more, myside arguments, otherside arguments, and elaborations in the written condition. The data from the college sample produced a similar picture, though the undergraduates did produce slightly more otherside arguments orally, but at a nonsignificant level.

In summary, the experiment disclosed no dramatic advantage for either oral or written reporting. The correlations in the high school sample also argued that both methods were measuring the same thing, although the lack of significant correlations in the college sample is surprising in this regard. One may perhaps not take alarm at that result considering that correlations using the same methodology across two issues typically have not been that high, of the order of .3 or .4. All in all, the results justify the practice in this series of studies of collecting subjects' responses in writing, since no decisive advantage for oral responses appears, even for the youngest subjects.

5.4 Experiment 12: Reasoning on Sociopolitical versus Personal Issues

Another natural methodological concern with this line of inquiry addresses the investment of subjects in the issues considered. To be sure, issues with some currency were selected and in general subjects indicated that they found the issues fairly interesting. However, evidently a personal decision crucial to an individual's future would provoke much more care. An experiment was designed that sought to sample subjects' reasoning on one of our characteristic issues along with reasoning on an important personal decision. Correlation coefficients between the subjects' scores on both tasks would allow us to assess whether performance on our issues was at least somewhat representative of the subjects' reasoning in more personal circumstances.

5.5 Method

5.5.1 Subjects

The subjects were 39 adults, balanced for sex, ranging in age from to 19 to 57 years old. The average age was 34. Educational background ranged from high school education to recipients of doctoral degrees. The average number of years of education was 16. Subjects were recruited by posters soliciting people in the midst of a major decision concerning such issues as employment, health, education, family/marital status, and so on. A modest fee was paid for participation.

5.5.2 Procedure

The subjects were mailed questionnaires that collected the usual background information and then asked subjects to start with either their personal decision or a vexed sociopolitical issue, namely:

- Would a nuclear freeze between the United States and the Soviet Union significantly reduce the likelihood of world war?

The order was counterbalanced across subjects to compensate for possible order effects. After addressing one issue or the other, subjects rendered an initial judgment.

The participants rated their confidence in their choices for the nuclear freeze and personal decision questions on a percentage scale both before and after writing out their reasoning. The amount of time subjects had spent considering the respective issues in the past was collected.

5.5.3 Scoring

Scoring was performed by three judges working independently. All of the performances relating to the nuclear freeze were scored independently by two judges. A random subsample of performances relating to the real life issues was also scored independently by two judges. Interjudge correlations were calculated after scoring was completed.

5.6 Results

Reliability of scoring. Interjudge correlation coefficients for the nuclear freeze performances ranged from .53 on otherside arguments (p<.0005, N=39) to .99 on sentences (p<.0005, N=39). The average correlation was .8. The average correlation for the personal issue performances was .7 while the range was .56 (N=14, p<.025) to .80.

Degree of development of arguments. Subjects produced more developed arguments for their personal issue than for

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

the nuclear freeze question. They offered an average of 4.7 myside arguments for their personal issue in contrast with an average of 2.5 myside arguments for the nuclear freeze question. A matched t-test showed a significant difference between the two figures (p<.001, two-tail). A similar relationship obtained for the otherside arguments, where 4.0 arguments were produced for the personal issue while 1.4 arguments were produced for the nuclear freeze question. Again, a matched t-test showed a significant difference between the figures (p<.001, two-tail).

Time spend developing arguments. Subjects claimed to have spent an average of 125 hours thinking about their personal decisions prior to the study. For the nuclear freeze issue, they claimed to have spent 10.5 hours prior to the study. This figure is substantially higher than the amount of prior thought on the issues reported by subjects in our other experiments. Despite the time, their arguments on the nuclear freeze issue were not dramatically more developed than those from subjects in our other experiments.

Balance of arguments. Subjects were asked to choose a vexed personal issue to reason about. Therefore one would expect that personal issue performances might be more balanced in terms of pros and cons than sociopolitical issue performances. Such a relationship did in fact manifest. The myside ratio (myside arguments divided by myside plus otherside arguments) for the personal issue was .56, which is not significantly different from .50, which would describe a perfectly balanced performance. The myside ratio for the nuclear freeze performances was .71, which is significantly different from .50. The two myside ratios are significantly different from each other (p<.003, two-tail t-test). Note that this is not evidence that reasoning on personal issues is generally more balanced than reasoning on less personal issues: subjects were selected for their involvement in a personal decision they were having trouble with. It may well be that many people make important life decisions in one-sided ways, but, of course, they do not perceive themselves as having trouble with these decisions.

Correlation between performance on the personal and nuclear freeze issues. Differences in degree of development of the argument notwithstanding, does performance on the sociopolitical issue tap the same underlying ability as performance on the personal issue? To answer this question, one needs some sense of the typical relation between performance on two sociopolitical issues – a test-retest correlation in effect. Prior research showed correlations of about .36 (N=64, p<.01, one tail) between myside and between otherside counts on two different sociopolitical issues, such as the nuclear freeze question and the aforementioned school funding question. About the same correlation held between subjects' personal issue and nuclear freeze performances in this experiment. For myside, r = .38 (N=39, p<.01, one tail) while for otherside the figure was .34 (N=39, p<.01, one tail).

In summary, the results suggest that people develop their arguments on vexed personal issues considerably more fully than on the sorts of issues used in our research; however, performances on either tap the same underlying competence, the differences simply reflecting time on task.

6 General Discussion

We have described a number of experiments that bear on the issue formulated at the outset: To what extent is good informal reasoning a matter of general intelligence in the psychometric sense or other capacities versus a metacognitive repertoire that guides the reasoner effectively? Here we review the principal findings and explore the implications for further research and educational practice.

6.1 Validity of the methodology

Interjudge reliability. Both within particular studies examining the main issue of this program of investigation and in studies specifically for the purpose, the general validity of the methodology in use was examined. The methodology involved counts and ratings of various sorts that had to be done by judges, with more than one judge scoring "blind" at least a random subsample of the data in any particular experiment. In general, interjudge correlations were quite high. From time to time, a scoring dimension was eliminated because adequate interjudge agreement was not obtained.

Sufficient time. Another continuing concern was whether subjects had sufficient time to reason about the issues given them. After all, one can ponder many issues for weeks before being forced to a decision by, for instance, the advent of election day. It was usually but not always possible to arrange matters so that the session during which a subject reasoned did not have significant time limits. In prior similar research, we collected subjects' own ratings of whether they thought the time was sufficient, and usually they did (Perkins, 1985). The general point is that, without some kind of scaffolding, subjects run out of ideas on an issue surprisingly soon – typically within five minutes or so. Accordingly, one does not need to provide a great deal of time for subjects to ponder during a single session. If time limited performance in our subjects, it probably did so because subjects carried out their reasoning in a single session, not because they did not have all the time they could use in that session. The remarks below on significant personal decisions bear on this point as well. On the other hand, many subjects had thought about the issues before participating in our experiment, so one would hope that their prior reflections as well as their thinking during the experiment informed the arguments they offered.

Issues one cares about. A natural reservation about the methodology is that subjects might not feel very invested

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

in the issues they were asked to reason about. Of course, we tried to pick timely issues that many people do express considerable concern with and subjects generally did indicate moderate interest in the issues. Nonetheless, it is unlikely that the subjects cared about the issues employed as they would care about personal decisions that deeply affected their lives. To explore this matter, the "real life decision making" study was conducted, in which subjects both performed one of our normal reasoning tasks and reported in similar fashion on their thinking about an important life decision that was current for them.

As one might expect, the results disclosed that subjects' arguments concerning their real-life problem were more elaborate, about twice as developed in fact. This presumably reflects not only their investment in the issue but the fact that these subjects reported having spent a great number of hours pondering their personal decisions. The correlations between their performance on the social issues used in the usual experimentation and their performance on their personal decisions were positive and statistically significant, indicating that the two methods were to some degree measuring the same thing. The correlations were not high; on the other hand, they were no lower than correlations between performance on two of our regular issues for those experiments where subjects did two issues on pretest and two on posttest. Consequently, our usual measures do seem to be indicative of people's reasoning abilities.

6.2 Informal Reasoning as a Matter of Capacity

Impact of instruction. The program of research investigated the extent to which different degrees of skill in informal reasoning reflected capacity limits of general intelligence, or perhaps of some other sort. Our own teaching experiments as well as the pretesting and posttesting of established instructional programs with a special emphasis on critical thinking demonstrated that instruction can lead to students reasoning significantly better according to various of the measures employed. The rates of gain for established instructional programs were not dramatic, but they were several times greater than the very small rates of gain reported by Perkins (1985) for conventional education. They concerned arguments on one's own side of the case, even in the debate class where both the conduct of debate and the class itself emphasized examining both sides. The gains for our own brief instructional interventions appeared entirely in improvement in attention to the other side of the case; they were quite substantial gains, especially considering the brevity of the instruction.

These results demonstrate that the normal trajectory of development reported by Perkins (1985) does not reflect a capacity ceiling. People can improve their informal reasoning much more quickly with appropriate instruction. Such a finding accords with positive results in certain other efforts to teach thinking strategies (e.g. Bolt, Beranek, and Newman, 1983; Nickerson, Perkins & Smith, 1985; Schoenfeld, 1982; Schoenfeld & Herrmann, 1982; Palinscar & Brown, 1984). The results also suggest that attention to the other side of the case is a particular trouble spot calling for very direct attention.

Impact of scaffolding. The most dramatic improvements in performance came not from instructional settings but from settings in which the subjects were led by a series of general questions, not specific to the particular issue, to develop their arguments further. Such scaffolding helped subjects to expand the breadth of their arguments several fold and construct arguments that more evenhandedly addressed both sides of the case. Since the scaffolding was entirely generic in nature, presumably people could learn to scaffold themselves in such a manner. In effect, the scaffolding was a metacognitive strategy of a sort that people might learn. The scaffolding results suggest that people often reason informally well below their capacity. Neither the established instructional programs nor our own teaching experiments achieved nearly as much gain. These results accord with other experiments in which scaffolding a performance has considerably enhanced it (Heller & Reif, 1984; Perkins & Martin, 1986).

Role of general intelligence. Turning to the question of general intelligence specifically, the research disclosed a vexed relation between general intelligence in the psychometric sense and good reasoning. As one might expect, individuals with more general intelligence performed somewhat better in various discriminations associated with reasoning, although all subjects performed quite well. Also, individuals with more general intelligence developed their myside arguments more fully. However, they invested no more effort than less generally intelligent participants in otherside arguments. In effect, their arguments were more developed but also more lopsided. This is a nice empirical demonstration of the point that general intelligence and rationality are not the same thing: Rationality calls for evenhandedness, which general intelligence does not necessarily promote. Presumably, the ideal rational person both has high general intelligence to assist in the development of arguments and a metacognitive repertoire of skills and dispositions that promotes thoroughness and evenhandedness.

Role of other capacities. The scaffolding experiment investigated the role of certain other capacities – capacity to maintain a focus on the issue in question, discriminate more from less important arguments, and so on. Actually, it is not clear whether these ought to be called capacities. In any case, that question is moot since subjects' reasoning performance did not turn out to be severely limited by shortfalls in these areas. In consequence the principal factors constraining subjects' normal reasoning performance appear to be matters of underexploration of issues and myside bias in ex-

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

amining issues, rather than limitations of general intelligence or particular judgmental capacities.

6.3 Informal Reasoning as a Matter of Metacognition

All the instruction examined experimentally was instruction that either directly or indirectly would be expected to enhance subjects' metacognitive repertoire for reasoning. Consequently, it is reasonable to attribute the gains reported to such a cause, although other causes might be proposed as well. The scaffolding experiment, as already emphasized, in effect provided subjects with metacognitive guidance. The dramatic development of subjects' arguments in response to scaffolding accordingly defines a level of aspiration for what a fairly refined set of metacognitive strategies might accomplish.

In contrast with these outside interventions, high school subjects' own metacognitive repertoires bore a vexed relation to their reasoning performance: Subjects who could say more about how to reason tended to be more one-sided in their reasoning but not significantly more prolific. These subjects also tended to be more intelligent, so this trend may relate to the trend toward bias mentioned earlier for such subjects. It also may be that the subjects did not connect their ideas about decision making to our issues, which were posed as matters of deciding the truth of a claim. In any case, the results offer no encouragement for the idea that even brighter students' metacognitive repertoires equip them for good informal reasoning.

6.4 Implications for Research and Education

Inevitably, so complex a subject as informal reasoning and its difficulties cannot be exhausted by any program of investigation. Two issues in particular emerge from the present inquiry as especially calling for further research. First of all, our research showed that, not surprisingly, people handle important life decisions better than the sorts of issues used in our studies, notwithstanding the general timeliness and interest of those issues. The research reported here still seems to us entirely relevant, since as citizens, in many managerial roles, and in many other ways people need to take seriously and make wisely a number of decisions in which they are not deeply personally invested. At the same time the findings demonstrate the need to look more carefully at people making important life decisions. It is by no means clear that they are performing well, although they seem to be performing better. For instance, scaffolding people's reasoning in important decision making contexts might substantially extend their exploration of causes and consequences and the soundness of resulting decisions. Also, the methodology used provides no information about whether people are as prone to bias in important personal decisions as they prove to be with our issues. This, too, is an important question.

A second area for investigation concerns subjects' readiness to change their minds. As discussed earlier, while scaffolding leads to subjects substantially extending their arguments and bringing them into better balance, it has some, but much less, impact on their final positions. We need to understand better the factors that figure in those final positions and their degree of tractability to change. This calls for more focused research on the aspects of informal reasoning that lead to actual resolution, in contrast with those factors that figure in constructing a better balanced and elaborated argument prior to resolution.

These questions not withstanding, the present results have immediate implications for the practice of education. The findings reported here argue that when people reason informally they are not typically performing close to their capacity limits, that they can on cue or through instruction perform substantially better, that their normal metacognitive repertoires may not empower them much in that direction, and that they might benefit substantially from direct instruction in good informal reasoning, instruction that conventional education appears not to deliver. All this encourages more explicit attention to the development of informal reasoning abilities on the part of education. Appropriate instruction, in the form of separate courses or integrated with the subject matters, should be able to help learners to attain levels of performance substantially greater than those they characteristically display.

References

Berrueta-Clement, J. R., Schweinhart, L. J., Barnett, W. S., Epstein, A. S., & Weikart, D. P. (1984). Preschool's long-term impact: Summary of the evidence. Chapter VI in *Changed lives: The effects of the Perry preschool program on youths through age 19* (pp. 94–105). Ypsilanti, Michigan: The High/Scope Press.

Bolt, Beranek, and Newman. (1983). *Final report, Project Intelligence: The development of procedures to enhance thinking skills*. Cambridge, Massachusetts: Author.

Borgatta, E. F., & Corsini, R. J. (1964). *Quick Word Test*. New York: Harcourt, Brace & World.

Garber, H., & Heber, R. (1982). Modification of predicted cognitive development in high-risk children through early intervention. In Detterman, D. K., & Sternberg, R. J. (Eds.), *How and how much can intelligence be increased?* (pp. 131–127). Norwood, New Jersey: Ablex.

Greenfield, P. M. (1984). A theory of the teacher in the learning activities of everyday life. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 117–138). Cambridge, Massachusetts: Harvard University Press.

https://doi.org/10.1017/S1930297500005350 Published online by Cambridge University Press

Heller, J. I., & Reif, F. (1984). Prescribing effective human problem-solving processes: Problem description in physics. *Cognition and Instruction, 1*(2), 177–216.

Jensen, A. R. (1983). The nonmanipulable and effectively manipulable variables of education. *Education and Society 1*(1), 51–62.

Jensen, A. R. (1984). Test validity: g versus the specificity doctrine. *Journal of Social and Biological Structures, 7*, 93–118.

Nickerson, R., Perkins, D. N., & Smith, E. (1985). *The teaching of thinking*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Palinscar, A. S. & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*, 117–175.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology, 77*(5), 562–571

Perkins, D. N. (1985). Reasoning as imagination. *Interchange, 16*(1), 14–26.

Perkins, D. N., Allen, R., & Hafner, J. (1983). Difficulties in everyday reasoning. In W. Maxwell (Ed.), *Thinking: The frontier expands* (pp. 177–189). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Perkins, D. N., & Martin, F. (1986). Fragile knowledge and neglected strategies in novice programmers. In E. Soloway & S. Iyengar (Eds.), *Empirical studies of programmers* (pp. 213–229). Norwood, New Jersey: Ablex.

Ramey, C. T., MacPhee, D., & Yeates, K. 0. (1982). Preventing developmental retardation: A general systems model. In D. K. Detterman,.& R. J. Sternberg (Eds.), *How and how much can intelligence be increased?* (pp. 343–401). Norwood, New Jersey: Ablex.

Rogoff, B., & Gardner, W. (1984). Adult guidance of cognitive development. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 95–116). Cambridge, Massachusetts: Harvard University Press.

Schoenfeld, A. H. (1982). Measures of problem-solving performance and of problem-solving instruction. *Journal for Research in Mathematics Education, 13*(1), 31–49.

Schoenfeld, A. H. & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 484–494.