EVIDENCE EVALUATION AND USE IN UNDERGRADUATES' EVERYDAY REASONING

TONY ANDERSON, REBECCA SODEN AND SIMON HUNTER

SYNOPSIS

This article examines the use of and reaction to evidence in argumentative reasoning in undergraduates. The study involved 30 higher education students. Participants were pre- and post-tested using a modified version of Perkins' procedure, in which they separately set out their preferred (myside) case separately from their dispreferred (otherside) case on topics of general interest like the effect of TV violence on behaviour. The peer interaction task required them to identify strengths and weaknesses of items of evidence of different types (anecdotes, generalised claims and research evidence). The results showed that participants could make some appropriate criticisms of items of evidence, and were most positive about research-based evidence. However, generalisations were overwhelmingly the most frequent form of evidence used by the participants, even at post-test, following exposure to a variety of evidence forms on the topics.

INTRODUCTION

Increasingly, the ability to think critically is an example of a generic or transferable skill that is a desired outcome of post-school education (Harvey, Moon, Geall, and Bower 1997; National Committee of Enquiry into Higher Education, 1997). Plainly, critical thinking is implicated in core and key skills such as communication and problem solving. Whilst the notion of critical thinking is broad-ranging and encompasses a variety of thinking skills (for useful discussions see eg. McPeck, 1981; Ennis, 1987), we shall focus upon a more specific aspect of critical thinking highlighted by Kuhn (1991), i.e. the ability to conduct reasoned argument. This specifically includes the abilities to a) differentiate opinions from evidence, b) to support opinions with non-spurious evidence, c) to propose opinions alternative to one's own and to know what evidence would support these, d) to provide evidence that simultaneously supports one's own opinions while rebutting alternatives, and e) to take an epistemological stance which involves weighing the pros and cons of what is currently known. We will therefore use the terms 'critical thinking' and 'argumentative reasoning' interchangeably.

This type of thinking has been the focus of much research in recent years, both from the point of view of documenting its strengths and weaknesses in the adult population (eg. Perkins, Allen and Hafner, 1983; Kuhn, 1991) and also with a view to exploring possible ways of improving matters. The view that the ability to deploy cogent critical reasoning in the sense alluded to above is the exception rather than the rule in the adult population has been given support by the research of Perkins and his colleagues on the one hand and Deanna Kuhn and her colleagues on the other (eg. Perkins et al., 1983, Kuhn, 1991).

Perkins and his colleagues (eg. Perkins et al., 1983) asked participants to take a position on issues such as whether passing a law to require drinks manufacturers to provide a 5-cent return on empty bottles and cans would reduce litter. The participants' position statements were analysed in terms of their presentation of lines of argument on each side of the issue, and this revealed a strong 'myside bias' in terms of the numbers of arguments advanced in support of the subject's preferred case as compared to its opposite. In these tasks, the reasoner creates his or her own premisses and evaluates them in terms of their plausibility, a process that according to Perkins et al. involves building, testing and modifying mental models until they are robust against possible objections. In characterising the model-building process, Perkins et al. contrast what they call a 'makes-sense epistemology' (the tendency to minimise cognitive load by contenting oneself with the simple criterion of whether a given proposition makes intuitive sense and rings true) with a 'critical epistemology' (where subjects understand the pitfalls of justification and are more likely to ask when and why their model might fail, and build a model that is more robust against criticism as a result). Perkins claimed to demonstrate that participants typically fail to explore the full ramifications of their reasoning. For example, scrutiny of participants' protocols by a reasoning expert found them to be readily objected against; it was not difficult to envisage alternative outcomes issuing from the same premisses, or to pinpoint influential factors that the participant had completely overlooked.

The theme of commonsense reasoned argument as typically involving a 'metacognitive shortfall', and in particular as being characterised by a poor grasp of justification, is echoed in the work of Kuhn (eg., 1991). Kuhn asked participants to explain the causes of three social phenomena (school failure, unemployment, and recidivists' return to crime, in decreasing order of expected familiarity). She sought to obtain individual subjects' theories, their justifications for their theory, any alternative theories they could envisage, any justifications they could envisage for those theories, and how they would rebut evidence in favour of theories they did not personally subscribe to. She found a willingness to assert causal explanations, but poor differentiation of theory from supporting evidence, frequent generation of 'pseudoevidence' (i.e. richly elaborated plausible scenarios that restate the theory rather than provide genuine, covariation evidence for it), and frequently poor levels of ability to envisage alternative theories or evidence in their favour. Domain expertise made no difference to the quality of argument (in the sense that parole officers did not reason any more cogently about recidivism than did other participants), but a college-level education was associated with better performance (in terms of generating evidence, envisioning multiple lines of causation, and having a more sophisticated evaluative epistemology) than no further education, and expertise in philosophy proved most helpful in facilitating performance. Even in college-educated students, however, there was room for improvement. Such findings have implications for employability and citizenship: most employees are expected to provide evidence about why they prefer one course of action over another.

Thus far, the results cited would appear to yield a somewhat bleak picture of the ordinary person's argumentative reasoning skills. However, a number of possible palliatives have been suggested in the research literature. For example, one of Perkins's studies (see Perkins, 1989) involved 'scaffolding' the reasoners on a one-to-one basis, requesting them to rank order reasons in terms of quality, think of obvious counterarguments to specific propositions, and in general reason more thoroughly and deeply, particularly on the non-preferred side of the argument (the 'otherside'). Again, Perkins (1989) found that training interventions involving the explicit teaching of thorough situation modelling do indeed effect improvement, although improvement is often on a modest scale.

Two bodies of literature address questions about the nature of good thinking and how it might be fostered in courses on how to think well: the literature on critical thinking (eg. Perkins and Grotzer, 1997, Sternberg, 1987) and the literature on teaching in Higher Education (eg. Laurillard, 1993; Lonka and Ahola, 1995). Practices claimed to be optimal for developing the sorts of thinking expected of graduates have been well described (Entwistle, 1994; Gibbs, 1992; 1993; Ramsden, 1992; Sternberg, 1987; Wisker & Brown 1996). However, the mechanisms by which such practices exert effects on critical thinking are as yet not well understood.

A third type of palliative that has been used in previous research is the use of peer interaction techniques to overcome mental set or 'cognitive hysteresis' (Miyake, 1987) and these interventions have also enjoyed some degree of success. Kuhn, Shaw and Felton (1997), used peer interaction specifically to enhance argumentative reasoning, with each participant engaging in a succession of one-to-one discussions with each of five other participants, the topic at issue being capital punishment. Following peer discussion, participants were more likely to display more sophisticated forms of reasoning, such as multiple-sided reasoning. Likewise, our own previous work (Anderson, Howe, Soden, Halliday and Low, 2001) demonstrated that a combination of teaching about reasoned argument, modelling its processes, and engaging in peer critiquing exercises over a 10-week intervention period significantly improved further education students' use of evidence in their written project work compared to a previous control cohort. The Anderson et al. study demonstrated that, whilst control participants presented their projects with very little in the way of justificatory evidence, students who had undergone the teaching intervention used significantly greater quantities of evidence-based justification, albeit typically of a weak nature (i.e. using anecdotes or generalised claims rather than citing research-based evidence).

The present study complements the earlier study by exploring evidence-based justification in (predominantly) first-year university undergraduates, as opposed to the further education college students tested in the previous research. The study attempted to address the following issues. What forms of evidence do beginning undergraduates typically use to justify their beliefs? Given a peer-based exercise in which various items of evidence of different types (anecdotes, generalised claims, and research-based evidence) are presented for critique, what do the students perceive the strengths and weaknesses of these different types of evidence to be? Does working with a peer on such a task alter the participants' subsequent use of evidence?

METHOD

Participants

The participants were 30 volunteer undergraduate psychology students, mainly in the first three months of their first year of study but with a small number (n = 5) in the first three months of their second year. The inclusion of a small number of second year students reflects some difficulties in participant recruitment. However, Perkins (1985) notes that each additional year of education within a cohort adds little to their scores on informal reasoning; therefore, the reasoning of the second year students is unlikely to be greatly more sophisticated than that of the first years, and all participants can be treated as a homogeneous group. Each participant was paid £5 for taking part.

The sample was predominantly female, reflecting the gender composition of the population. Their modal age was 18-19 years. Given that they were to participate in a peer interaction task, they were recruited in pairs such that the members of a pair knew and were friendly with each other, to ensure easy and natural communication. The pair matching was self-selected. This principle is used in most peer observation schemes in teacher education courses (such as the one for our own university staff). There are pragmatic grounds for self-selection: imposing a condition that might provoke psychological discomfort from revealing weaknesses in one's thinking to a person not of one's own choosing might dissuade students from participating in this type of research. Students knew what would be required and presumably chose partners with whom they could work without undue anxiety.

Participants were also tested in pairs for pre- and post-testing, although responses were made individually and conferring was not allowed.

Materials

Pre- and post-tests were identical to each other, and requested that subjects consider a topical issue, decide upon their view, and subsequently set out a case supporting their view, citing as much in the way of evidence and reasons for holding their view as possible. Subsequently they were asked to set out a separate case arguing for the opposite view, again citing as much evidence and arguments as they could. The pre- and post-test responses were written, as were the instructions. Had it been possible to recruit a larger sample, an alternating order of conditions (myside, otherside) would have been used.

The peer interaction task involved the presentation to the pairs of participants of sample arguments in favour of a particular case, presented by a (mythical) person, on each side of the topics encountered in the pre-tests. Therefore, these items of evidence were relevant to the subsequent post-test. Each argument contained three items of evidence: one anecdote (i.e. referring to one or a small number of specific cases or instances), one generalisation, and one item of (genuine, not fabricated) research-based evidence. For example, evidence given against the death penalty included the facts that: hanging can be inhumane, exemplified by Ruth Ellis's drawn-out death occurring as a consequence of her neck failing to break immediately (anecdote); the death penalty fails to deter those murders which are spontaneous rather than pre-planned (generalisation); and a review of cases demonstrated that 350 defendants have been wrongly convicted of murder in America in this century alone (research evidence). Similar items of evidence were collected for each of the three topics used. These items were taken from actual published research (in the case of items of research evidence) and from responses given by an independent panel of students tested prior to the study (in the case of anecdotes and generalisations).

Design

The study used a repeated-measures, pre-test/treatment/post-test design.

Procedure

Participants initially were pre-tested at least 24 hours before the main task. In the pre-test they set out separate cases on each side of two issues, drawn from whether capital punishment would reduce the incidence of violent crime, whether viewing TV violence influences real-life violence, and whether cannabis should be legalised. The three topics were selected from a range of social topics of general interest, by pilot testing to ascertain the degree to which participants were able to make cases on these issues. Those particular topics were found to be ones which participants could readily generate arguments on both sides of the issue, and all were crime-related and therefore thematically similar. They were also all topics to which Psychology (the academic discipline studied by our participants) was directly relevant. Indeed, the content of the study would not have been out of place as a tutorial topic for the students. Pre-testing involved the individuals working separately to set out their chosen preferred and dispreferred cases. Perkins' original task required the generation of a single integrated argument that considered both sides of the case. However, it could reasonably be objected that rhetorical considerations would lead participants to represent their preferred case more strongly. After all, they are attempting to set out a persuasive argument, and setting out only a weak otherside case might occur for purely rhetorical reasons, rather than due to a failure of critical thinking. Hence our requirement to set out separate myside and otherside cases.

Example instructions for the pre- and post-tests for the TV violence topic would be as follows:

> ‘Consider the following contentious issue. Does the depiction of violence within television programmes and films/videos affect the level of violence that occurs in real life? Some people would assert that yes, violent behaviour portrayed in these media influences behaviour, effectively increasing the likelihood of violence in real life. Others, however, would deny that violence on TV/films and video influences people’s behaviour in the real world.
>
> ‘We would like you to think for a few minutes about this issue, decide which of these two points of view you agree with, and write a paragraph or two setting out your view and supporting it with as many reasons and as much evidence as you can think of. There is paper available for you to take notes whilst you consider the issue, and to plan your argument, if you wish. Please set out the argument for your position in such a way as to convince another person that you are correct, that is, giving as coherent and convincing a case for your chosen position on the issue as you can. Include as many arguments and items of factual evidence that you can.’

After having set out the ‘myside’ case in response to the above instructions, the instructions to set out the ‘otherside’ were given as follows:

> ‘Thank you for setting out your chosen point of view on the ‘media violence’ issue. We would now like you to think for a few minutes about the **opposite** argument on this issue. That is, we want you to think about the case for the point of view opposite to the one you have just set out. Again, we would like you to ponder the issue, and write a paragraph or two setting out this view and supporting it with as many reasons and as much evidence as you can think of. There is paper available for you to take notes whilst you consider this side of the issue, and to plan your argument, if you wish. Please set out the argument for this position in such a way as to convince another person that you are correct, that is, giving as coherent and convincing a case for this position on the issue as you can. Include as many arguments and items of factual evidence that you can.’

In the peer interaction task, participants were instructed to work together within their pairs, jointly considering and discussing each item of evidence in turn, rating it on a 5-point scale from totally unconvincing to absolutely compelling, and listing up to three strengths and up to three weaknesses of each.

Data coding

The participants’ written pre- and post-test responses were coded blind with respect to their pre- and post-test status. The number of lines of argument given were coded, in addition to the numbers of: a) unjustified statements made (including rhetorical questions designed to make a point, eg. ‘Would it not be more beneficial to rehabilitate murderers?’; ‘Everyone should have the chance to reflect on their own actions’), b) anecdotes cited (reference to specific cases, eg. the Guildford four, Jamie Bulger, Ruth Ellis); c) generalisations made (i.e. any general claim, eg. ‘Violence on television does increase aggression levels’, ‘A lot of people who see violence on TV do go out and copy what they saw’); and d) items of research evidence cited (defined as any reference to formal research-based or statistical evidence – eg. mention of the well-known experiment by Bandura and his colleagues (Bandura, Ross and Ross, 1963), or descriptive reference to well-known psychology experiments). A sample of the written work was coded by a second coder, and the overall inter-rater reliability coefficient was 0.91.

The participants’ dialogues that took place as they undertook the peer interaction task were coded using a scheme based on that originally used by Perkins, Allen and

Hafner, 1983; see Table 1 below for a listing of category names and definitions). The categories used to criticise evidence included several that emerged in the experts' objections to participants' arguments noted in Perkins et al. (1983), for example, 'alternative cause', 'neglected influential factor', 'scalar Insufficiency', 'Irrelevant argument' and 'counter-example'. Others were derived empirically by examination of a sample of the data. Unlike Perkins et al.'s original study, we also asked participants to list strengths as well as weaknesses of arguments. In consequence, some of the coding categories concern positive reactions to an item of evidence (eg. 'agreement by virtue of evidence form', 'resonates with experience'). Again, a sample was coded by a second independent coder, and the resulting overall dialogue coding reliability was 93%.

Table 1. Dialogue coding categories, with examples.

Minimal Agreement *'That's true'; 'That's a reasonable proposal'; restatements of argument*
Resonates with experience *'I've seen that happen myself'; 'I can think of dozens of examples'*
Alternative Cause '*How does she know it was the cannabis that led to the addiction to harder drugs?'*
Reference to rhetoric (positive and negative) *'She seems a bit unsure'; 'That's a bit weak'; 'He knows what he's talking about'*
Reference to evidence form (positive) *'That's something that actually happened'* *'It's research, so it should be reliable'*
Reference to evidence form (negative) *'It's not quantitative'; 'Personal experiences are not valid as evidence'*
Generalisability *'Not everyone is like that'; 'That may not be true in all cases'*
Neglected Influential Factor ' *'Legalising cannabis would take away the excitement of doing something illegal'*
Scalar insufficiency *'That's only true for very violent films'*
Irrelevant argument *'It doesn't matter how painful Ruth Ellis's death was – the argument is about whether you should have capital punishment, not about particular methods'*
Counter-example *'But kids can make the distinction between hard and soft drugs'*

RESULTS

Pre- and post-tests

Table 2 shows the number of lines of evidence given on each side of the case in both pre- and post-tests.

Table 2. Mean number of lines of argument given at pre-test and post-test for the myside and otherside cases. Standard deviations are in parentheses.

Pre-tests:	**Post-tests:**
Myside vs Otherside (n = 30)	Myside vs Otherside (n = 15)
Myside 3.5 (0.87)	Myside 3.5 (0.93)
Otherside 2.9 (0.77)	Otherside 3.0 (0.88)
$p<0.003$	NS

Like Perkins' participants, ours are producing just over three lines of argument for their preferred case at both pre-test and post-test. In the case of our participants, otherside arguments average out at around 3, in contrast to Perkins' figure of 1.3. Despite the small mean difference, there are significantly more myside lines of argument at pre-test ($F1, 29 = 10.67$, $p < 0.003$); it should be noted that whilst this difference is statistically significant, it is unlikely to be educationally significant. At post-test, there are very similar mean numbers of lines of argument on each side. The myside-otherside difference in lines of argument is not, however, significant at post-test ($F1, 14 = 1.97$, $p= 0.18$). There is therefore no strong evidence of pre-post test gain. Note the sample attrition between pre- and post-tests: the study required participants to attend two sessions, and some participants failed to attend for the second session, despite their payment being contingent on attendance at both sessions. Examination of pre-test scores suggests that there was little difference between those who completed both parts of the task and those who did not; attrition appears not to have distorted the data.

Table 3 shows the mean numbers of items of evidence used in the different categories at pre and post-test, for both myside and otherside arguments. The patterns are broadly the same in all cases, with generalisations being significantly more prevalent than all other forms. Comparing myside and otherside arguments, at pre-test there appear to be more unjustified statements and more anecdotes in myside than in otherside arguments: these differences are in fact significant ($t = 2.6$, $p < .01$ and $t = 2.8$, $p<0.008$ respectively). Similar patterns appear at post-test, but the differences are non-significant (statements: $t = 1.26$, $p = 0.22$; anecdotes: $t = 0.59$, $p = 0.57$). Perhaps surprisingly, there is no evidence of change in the distribution of evidence types between pre- and post-test: despite exposure to a range of items of evidence during the peer interaction task, the balance of evidence types supplied remains similar across pre- and post-test, with generalisations being the most prevalent of all evidence types in all cases (pre-test, myside $F3, 87 = 78.8$, $p<0.0001$; pre-test, otherside, $F3, 87 = 109.7$, $p<0.0001$; post-test, myside, $F3, 42 = 55.4$, $p< 0.001$; post-test, otherside, $F3, 42 = 53.4$, $p < 0.001$).

Table 3. Evidence use in written pre and post tests.

1 Myside	**Pre**	**Post**	**2 Otherside**	**Pre**	**Post**
Statements	1.4	1.2	Statements	0.87	0.9
Anecdote	1.0	0.4	Anecdote	0.42	0.5
Generalisation	5.5	4.8	Generalisation	4.9	4.5
Research	0.3	0.2	Research	0.13	0.23
P<	.0001	.0001	P<	.0001	.0001

Joint ratings of evidence forms

Table 4 shows the ratings of strength of the different types of evidence, and the mean numbers of cited strengths and weaknesses, given by the students.

Table 4. Ratings of argument types by the pairs, and numbers of strengths and weaknesses nominated

	Rating	**Strengths**	**Weaknesses**
Anecdote	2.66	1.44	2.06
Generln.	2.7	1.46	1.92
Research	3.3	1.69	1.79
	<.01	NS	NS

The ratings demonstrate that research evidence is regarded as significantly better than the other two types of evidence ($F2, 22 = 5.5$, $p<0.01$), and follow-up Scheffe test demonstrate that research evidence is rated significantly higher in strength than the other two types of evidence, which are non-significantly different from each other. The numbers of strengths and weaknesses given for the different types of evidence show non-significant trends such that research has the largest number of strengths and the fewest weaknesses. There are significantly more weaknesses than strengths in the case of anecdotes (Wilcoxon $Z = 2.825$, $p< 0.005$), but the differences in numbers of strengths as opposed to weaknesses are non-significant for the other two types of evidence. Research evidence is thus seen as being the best quality evidence and is clearly set apart from the other two evidence types in the strength ratings.

Reactions to evidence forms: the dialogue analysis

Table 5 shows the mean total volume of dialogue, the mean volume of dialogue agreeing with evidence items, and the mean volume of dialogue disagreeing with evidence items, by type of evidence discussed.

Generalisations generate most dialogue overall ($F2, 20 = 24.9$ $p < 0.01$). There is no significant variation in the volume of dialogue in which the different types of evidence are agreed with ($F2, 20 = 2.6$ $p = 0.10$), but there is in dialogue where the evidence is disagreed with ($F2, 20 = 6.2$ $p < 0.008$) – and again, it is generalisations that get most of the attention. The precise form that this disagreement takes will now be examined further.

Tables 6 and 7 show the frequencies of specific forms of agreement and disagreement respectively, by evidence type.

Perhaps the most striking point of all in the two tables is the low values of the mean volume of dialogue in the various categories. That is, the total dialogue (expressed as numbers of speaker turns) is small in most cases, clearly indicating

Table 5. Analysis of peer dialogues

Mean total volume of dialogue, mean volume of dialogue agreeing with evidence item, and mean volume of dialogue disagreeing with evidence item, by type of evidence discussed.

	Total	**Agreement**	**Disagreement**
Anecdote	34.4 (9.7)	12.8 (4.9)	21.6 (6.1)
Generalization	48.7 (18.4)	19.5 (11.3)	29.2 (10.1)
Research	38.1 (14.0)	18.9 (8.4)	19.2 (8.0)
P <	.02	NS	.01

Table 6. Frequency of specific forms of agreement

	Anec.	**Generln**	**Resrch**
Minimal Agreement*	6.2 (5.6)	9.1 (6.3)	4.3 (2.8)
Resonates w.experience	0.18 (0.4)	0.63 (0.9)	0.18 (0.4)
Elaborate	2.4 (2.7)	7.5 (7.7)	5.4 (7.4)
Cite form**	3.4 (2.4)	1.4 (3.9)	7.6 (4.2)

* p<.01
** p<.0001

that the dialogues are not lengthy (and therefore, not extensive or deep). A three-utterance dialogue could take the form of articulation of objection – expression of agreement from interlocutor – acknowledgement, and this minimal style of dialogue was not uncommon. Protracted discussion sequences were rare. Some of the very low values (e.g. for scalar insufficiency or neglected factor objections in table 7) strongly imply not only that such objections are not thoroughly discussed, they are also infrequent in an absolute sense. Some categories of objection are, therefore, rare in occurrence, and those that are less rare are not often the subject of protracted discussion.

Table 6 shows that generalisations stand out as having the greatest volume of dialogue of the minimal agreement type ($F2, 20 = 3.9, p < 0.04$, although anecdotes are non-significantly different from generalisations in this respect. Research, on the other hand, stands out as being associated with dialogue of the type, 'agreement by virtue of form' ($F2, 20 = 14.4, p < 0.001$).

Table 7 shows that research evidence attracts a greater volume of dialogue in the scalar insufficiency category ($F2, 20 = 7.5, p < 0.004$), and significantly less dialogue in the 'objections by virtue of form' category, than do the other two forms of evidence. On the other hand, anecdotes are given a greater volume of discussion in terms of

Table 7. Frequency of specific forms of disagreement

	Anec.	Generln	Resrch
Alternative Cause	1.5 (1.6)	4.1 (1.5)	3.1 (1.5)
Alternative effect	1.0 (1.3)	1.6 (0.6)	1.1 (2.4)
Generlisbty	4.3 (4.3)	3.7 (1.8)	2.6 (2.2)
Neglected Factor	0.8 (1.3)	0.4 (0.8)	1.2 (2.7)
Irrelevant Argument*	5.5 (6.7)	2.4 (4.0)	0.8 (1.6)
Scalar Insufficiency	0.09 (0.3)	0.18 (0.6)	1.6 (1.9)
Counter-Example	1.4 (1.7)	3.5 (3.4)	1.2 (1.6)
Rhetoric	1.8 (2.4)	3.7 (5.9)	3.4 (3.0)
Form**	3.3 (3.6)	4.4 (3.6)	0.4 (0.9)

* $p<.03$; @ $p< .004$; ** $p<.02$

their being irrelevant ($F2, 20 = 4.03$, $p < 0.03$). No form of evidence receives more elaboration discussion than the others ($F2, 20 = 2.16$, $p = 0.14$). However, when elaborations are made, they tend to take the form of generalisations, whether the type of evidence being elaborated upon is anecdotal ($F2, 20 = 3.7$, $p<0.04$), is itself a generalisation ($F2, 20 = 8.8$, $p<0.002$), or is an item of research evidence ($F2, 20 = 4.5$, $p<0.02$). Overall, generalisations are the most common form of evidence cited, are most often used to elaborate encountered evidence, and although they generate a greater volume of sceptical dialogue, they aren't particularly associated with any specific type of objection; this is perhaps surprising, in that they might seem a priori vulnerable to criticism for neglect of critical distinctions.

DISCUSSION

Turning first to the results from the written pre- and post-tests, perhaps the first point to emphasise is that, despite the setting out of separate myside and otherside cases, we still found significantly greater numbers of lines of argument on the participants' preferred case. Myside bias is not, therefore, merely a matter of rhetoric.

In terms of the three types of evidence examined, generalisations predominate regardless of which side of the case is being presented and whether this is at pre- or post- test. This finding, together with the significantly greater number of unjustified statements and anecdotes in myside as compared to otherside arguments at pretest, suggests that this group of undergraduates' default way of providing evidence is to make generalised claims, and to embellish these with anecdotes and unjustified statements in the case of myside arguments.

A sceptic might argue that, since generalisations summarise a body of evidence whilst filtering out details, they therefore might have the virtue of appearing considered and authoritative, and it is not uncommon therefore to find even research scientists using this type of evidence when making claims (although we might expect many of the generalisations from this group to take a qualified form). Hence, it is unsurprising that undergraduates should use this type of evidence by default. It should however be stressed that a) the topics were such that research evidence already known to the participants would have been relevant, b) our criteria for classifying an item of evidence as research-based were anything but stringent or demanding, and c) despite exposure to items of research evidence in the group task, the volume of such evidence cited did not change at post-test. The students appear to value research evidence, but don't often use it, even by the weak criteria that we used to score it. Participants instead use generalisations extensively (both in their written cases and in the elaborations that they gave in reaction to supplied evidence) and aren't quite sure how to criticise them (the dialogues show that whilst anecdotes were criticised in terms of irrelevance, and research could be dismissed on the grounds of scalar insufficiency, there was no category of disagreement that was most strongly associated with generalisations). This suggests that a good starting point for any putative critical thinking intervention would involve the analysis and critique of generalised claims.

The students' focus on generalisations contrasts with Kuhn's (1991) finding that anecdotes are a popular form of evidence to cite in support of opinions. Intuitively, anecdotes might seem to have a vivid, compelling quality, since they concern concrete actual instances. Nevertheless, they are open to criticism in terms of their interpretation, and even accepting a particular interpretation, their representativeness. In practice, our participants' reactions to anecdotes varied depending on whether the anecdote in question was consonant with their own view. Reactions to anecdotes that were consonant with participants' preferred cases were often along the lines of asserting them as undeniable facts ('that's something that actually happened'), whereas anecdotes that ran counter to participants' theories were dismissed as irrelevant or a matter of personal opinion.

These results suggest that first and second year undergraduates have a long way to go towards being able to use the mode of enquiry associated with their discipline to come to an independent judgement about evidence. While the content of lectures, tutorials and prescribed reading includes illustration of the discipline's epistemology it seems that this is not enough for students to acquire this aspect of disciplinary competence: they seem to need explicit and extended instruction if they are to progress in the critical use of evidence.

What they seemed to have learned is that evidence citing is an important part of psychology. However, their grasp of the fact that citation in itself is insufficient to support judgements they reach appears to be somewhat weaker. In particular, they seem not to have grasped the prerequisites of evidence weighing, such as the ability to identify limitations in the research they cite or to challenge assumptions behind studies. A speculation is that, unless the relevant behaviours are well practised, they are difficult to bring into working memory when confronted with a reasoning task, particularly when they have to compete with less effective reasoning behaviours such as those described by Kuhn (1991) which, given their ubiquity, one might assume are very well practised.

The pre- and post-test results also show that there was little or no detectable change as a result of the peer interaction intervention. The quantity and quality of peer dialogue was perhaps lower than might be expected. The tasks and materials employed in the study were intended to be ecologically valid in the sense that they should be fairly typical of the types of questions used to generate small-group discussion in psychology tutorials. The fact that the dialogues generated in response to these tasks were somewhat sparse accords with tutors' everyday experiences with students at this stage in their undergraduate careers. Both lack of motivation and a lack of practice in participating in such discussions could contribute to this problem. More research is needed to tease out the relative contributions of these factors, and to establish the conditions under which such tutorial discussions are effective.

Alternatively, perhaps the intervention was simply too short in duration to yield a detectable impact; Kuhn et al. (1997) successfully used peer interaction to improve critical thinking, as did Anderson et al. (in press). However, both of these studies employed interventions that extended over several sessions, and the Anderson et al. study additionally employed explicit instruction and modelling techniques in addition to peer interaction. Future studies should, at the very least, employ extended interventions, and the issue of whether structured peer interaction alone can effect an improvement is one which is worthy of systematic study.

Overall, it is clear that this sample of undergraduates working on these topics showed imbalances in their use of evidence that contrast with those observed in previous research conducted largely with non-students, and imbalances in the distribution of reactions to examples of different types of evidence. Such observations, should replication prove them to be representative, suggest areas where practising thinking skills could profitably be targeted.

There are also some more general methodological lessons to be extracted from the study. Perhaps this type of task tends to underestimate participants' powers of argumentative reasoning. An assumption that appears to be prevalent in the literature is that the use of a fairly narrow range of topics of general social relevance provides a fair measure of participants' reasoning skills. If, however, such skills are closely tied to a domain's epistemological knowledge as some (e.g. McPeck, 1981; Bonnett, 1995; Chi, Glaser and Farr 1998; see also Gardner and Johnson, 1996) would assert, then it is possible that the same individual would give radically different performances on different topics, reasoning especially well on a topic about which he/she is thoroughly knowledgeable, yet performing much more poorly on the type of general socially relevant topic used in previous research by Perkins and Kuhn (and as used here). Yet the possibility that critical thinking is closely tied to domain knowledge is acknowledged minimally if at all in conceptions of generic and transferable skills. Empirical research suggests that people think differently depending on the context for which thinking has a point. (Lave, 1988b; Lave and Wenger, 1991; Engestrom, 1996). Drew (1998), in questioning the notion of 'key' or generic skills, reviews evidence suggesting that transfer from classrooms to workplace is scant. Hyland and Johnston (1998) call this 'the mythology of transferability'.

The results from the present study have implications for both the school and higher education curricula. Given that the undergraduates we tested were among the best qualified school leavers, perhaps it is surprising that their use of evidence was poorer than one might expect. One possible explanation is that they had indeed learned to use evidence in school subjects, but that this competence simply did not transfer to the topics in the study. This would be consistent with a substantial body of research in psychology (eg., Detterman and Sternberg, 1993, Gick and Holyoak, 1980). Butler (1998) reviews research that suggests ways of dealing with this in school (see also Perkins, 1983). One implication for the curriculum is that much more explicit attention needs to be given to practising this type of thinking skill.

REFERENCES

Anderson, T., Howe, C., Soden, R., Halliday, J. and Low, J. (2001). Peer interaction and the learning of critical thinking skills in further education students. *Instructional Science, 29, 1-32.*

Bandura, A., Ross, D. and Ross, S. A. (1963). Imitation of film-mediated aggressive models. *Journal of Abnormal and Social Psychology*, 63, 575-582.

Bonnett, M., (1995), Teaching thinking and the sanctity of content, *Journal of Philosophy of Education*, 29, 3, 295-309.

Butler (1998) 'promoting self-regulation in the context of academic tasks: the strategic content learning approach'. Paper presented at the meeting of the American Psychological Association, San Francisco, 18 August, 1998.

Chi, M., Glaser, R. and Farr, M., (eds), (1988*). The Nature of Expertise*. Hillsdale, NJ: Erlbaum.

Detterman, D. K. and Sternberg, R. J. (Eds.) (1993*). Transfer on trial: Intelligence, cognition and instruction*. Norwood, NJ: Ablex.

Drew, S. (1998)*, Key skills in higher education: background and rationale*, Staff and Educational Development Association No. 6.

Engestrom, Y. (1996), Developmental studies of work as a testbench of activity theory: the case of primary care medical practice. In S. Chaiklin and J. Lave, *Understanding practice: perspectives on activity and context,* (Cambridge University Press).

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron and R. J. Sternberg (eds.*), Teaching thinking skills: Theory and practice*. New York: W. H. Freeman and Co.

Entwistle, N. (1994), Recent research on student learning and the learning environment. Paper presented to the International Symposium on Independent Study and Flexible Learning, Cambridge, 6 September 1994.

Gardner, P. & Johnson, S. (1996) Thinking Critically about Critical Thinking: an unskilled inquiry into Quinn & McPeck*, Journal of Philosophy of Education*, Vol. 30, No. 3, pp. 441-456.

Gibbs, G. (1992). *Improving the quality of students' learning*. Technical and Educational Services: Bristol.

Gick, M. L. and Holyoak, K.J. (1980). Analogical problem solving. *Cognitive Psychology*, 12, pp.306-355.

Harvey, L., Moon, S., Geall, V. and Bower, R. (1997). *Graduates' work: organisational change and students' attributes*. Birmingham Centre for Research into Quality, University of Central England.

Hyland, T and Johnson, S (1998). Of cabbages and key skills: exploding the mythology of core transferable skills in post-school education. *Journal of Further and higher Education,* 22, 163-172.

Kuhn, D. (1991). *The skills of argument*. Cambridge: Cambridge University Press.

Kuhn, D. Shaw, V. and Felton, M. (1997) Effects of dyadic ointeraction argumentative reasoning. *Cognition and Instruction*, 15(3), 287-316.

Langer, E. (1997), *The Power of Mindfull Learning,* Addison-Wesley, New York.

Lave, J. (1988). *Cognition in practice: mind, mathematics and culture in everyday life*. Cambridge: Cambridge University Press.

Laurillard, D. (1993) *Rethinking University Teaching*, Routledge: London.

Lonka, K. & Ahola, K. (1995) Activating instruction: How to foster study and thinking skills in higher education. *European Journal of Psychology of Education*, Vol. 10, No. 4, pp. 351-368.

McPeck, J. E. (1981). *Critical thinking and education.* Oxford: Martin Robertson.

Miyake, N. (1987). Constructive interaction and the iterative process of understanding. *Cognitive Science*, 10, 151-177.

National Committee of Enquiry into Higher Education (1997). Higher Education in the Leaning Society (The Dearing Report) London, HMSO.

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

Perkins, D. N. (1989) Reasoning as it is and as it could be: an empirical perspective. In D. M. Topping, D. C. Crowell and V. N. Kobayashi (eds.). *Thinking across cultures: The third international conference on thinking*. Hillsdale, N.J.: Lawrence Erlbaum, pp.175-94.

Perkins, D. N. (1993) Teaching for understanding. *American Educator*, Vol. 17, No. 3, pp. 828-35.

Perkins, D. N., Allen, R. and Hafner, J. (1983) Difficulties in everyday reasoning. In W. Maxwell, (ed.*), Thinking: The expanding frontier*. Philadelphia, PA; The Franklin Institute, pp.177-189.

Perkins, D. N., and Grotzer, T.A. (1997) Teaching Intelligence, *American Psychologist*, Vol. 52, No. 10, pp.1125-1133.

Ramsden, P. (1992), *Learning to Teach in Higher Education*, Routledge, London.

Ramsden, P. (1994), 'Using Research on Student Learning to Enhance Educational Quality', Griffith Institute for Higher Education, Occasional Papers, Publication No. 2.

Sternberg, R.J. (1987) 'Teaching critical thinking: eight ways to fail before you begin', *Phi Delta Kappa,* Vol. 68, pp.456-459.

Wisker, G. & Brown, S. (Eds.) (1996) *Enabling Student Learning*. Kogan Page, London in association with the Staff and Educational Development Association.