Supporting conceptual change in school science: A possible role for tacit understanding
Christine Howe, Amy Devine, Joana Taylor Tavares

To cite this version:

HAL Id: hal-00719063
https://hal.archives-ouvertes.fr/hal-00719063
Submitted on 19 Jul 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Supporting conceptual change in school science: A possible role for tacit understanding

Journal: International Journal of Science Education

Manuscript ID: TSED-2010-0457-A.R1

Manuscript Type: Research Paper

Keywords: conceptual change, science education, experimental study

Keywords (user): conceptual change, science education, object fall
Supporting conceptual change in school science: A possible role for tacit understanding

Abstract

When students reason during school science, they often refer to conceptions that are derived from out-of-school experiences and are poor proxies for science orthodoxy. However, for some areas of science, these conceptions represent only a proportion of students’ full conceptual knowledge, for tacit understanding exists that is superior to the understanding displayed when reasoning. Noting that tacit understanding is engaged when events are judged as natural or non-natural, the paper is concerned with software that: a) requires the direction and speed of falling objects to be predicted, i.e. a typical science reasoning task that engages conceptual knowledge; b) presents simulations of predicted motion in the expectation that its naturalness or non-naturalness would be recognised. An evaluation study is reported where children aged 8 to 12 years worked with the software in contexts that typify computer use in classrooms, i.e. individually under adult guidance (n = 44 children) or in pairs with a classmate (n = 48 children). They were observed while they did this. Reasoning about object fall was assessed via change from individual pre-tests administered prior to software usage to individual post-tests administered a few weeks afterwards. The children who worked with the software showed greater pre- to post-test gains in conceptual understanding than control children (n = 47 children), who lacked software experience. The gains were especially marked for the children who worked in pairs. The approach taken is contrasted with traditional approaches to conceptual change in school science, e.g. strategies that rely upon classroom experiments.
Introduction

It is now beyond doubt that when students reason during school science, they often refer to conceptions that are derived from out-of-school experiences and are poor proxies for science orthodoxy. For instance, reference to unorthodox conceptions is documented in a significant proportion of the c.8000 entries that comprise Duit’s (2007) bibliography. Scholars vary over the terms that they use to characterise these conceptions: misconceptions, prior conceptions, alternative conceptions, alternative frameworks, everyday theories and intuitive notions are a sample of the numerous phrases that have traditionally been used (see, e.g., Driver & Erickson, 1983). Scholars also differ over whether they regard the conceptions as organised into tightly structured theories (McCloskey, 1983; Vosniadou, 1994), loose conglomerates (diSessa, 1993; Wagner, 2006), or a mixture of both (Howe, 1998; Oliva, 1999).

However, reference to unorthodox conceptions during reasoning in science is an established fact, and nowadays is widely recognised as a challenge that educators must address. Science teaching should, in other words, support processes of change from the use of unorthodox conceptions during reasoning to the use of received ones. Taking the need for change as established, the primary aim behind this paper is to contribute a new perspective upon how change can be supported.

As Scott, Asoko and Leach (2007) point out, contemporary approaches to conceptual change are heavily influenced by the work of Posner, Strike, Hewson and Gertzog (1982). According to Posner et al., effecting change depends upon fulfilment of four conditions: a) there must be dissatisfaction with existing conceptions; b) new conceptions must be minimally understood; c) new conceptions must appear initially plausible; d) new conceptions should suggest the possibility of a fruitful research programme. The fourth condition is probably of marginal significance in school
contexts, where students seldom regard themselves as on the threshold of research
programmes. However, it is hard to imagine compliance with the first three conditions
not being necessary, and so there have been numerous attempts to incorporate them
into teaching programmes. Typically the programmes require students to conduct
experiments to test their initial conceptions (Harlen, 1999; Hodson, 1993; Scott et al.,
2007). It is anticipated that results will, on the one hand, trigger dissatisfaction with
existing conceptions through their failure to provide support, and on the other hand,
render the target science both comprehensible and plausible through compatible data.
In reality, research evaluating the approach has obtained patchy results, with many
studies reporting minimal impact upon conceptual understanding (Chinn & Brewer,
1993; Duit, 1999; Duit, Treagust, & Widodo, 2008; Harlen, 1999; Limón, 2001).

Discussions of why evaluations have proved disappointing typically revolve
around three types of explanation: a) experiments often fail to generate the requisite
data; b) Posner et al.’s three conditions may be necessary but they are insufficient; c)
cognitive limitations eclipse the significance of experimental evidence. Experimental
failure is graphically and comprehensively documented in Claxton (1991), with
research indicating that inadequate control of variables is often responsible (Howe,
Tolmie, & Sofroniou, 1999; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Schauble,
1990; Scottish Executive Education Department, 2005). As regards possible
insufficiencies in Posner et al.’s conditions, social-constructivist critiques (see Duit,
1999; Duit et al., 2008; Limón, 2001) show how cultural (primarily discursive)
practices shape the significance accorded to experiments. A specific illustration
appears in Howe, Rodgers, and Tolmie (1990): when 8- to 12-year-old children
conducted experiments in small groups, any value from evidence regarding which
objects floated or sank depended on the discussions that the evidence generated. In
particular, when groups debated a range of ideas about how the evidence should be interpreted, group members showed improved understanding between individual pre-tests prior to group work and individual post-tests a few weeks later. When groups held similar ideas and debate was thereby precluded, conceptual understanding was identical at pre- and post-test, despite access to equivalent experimental evidence.

As regards cognitive limitations, difficulties have been identified at all ages with adopting the meta-cognitive perspective required to link conceptions with experimental data (Driver, Leach, Millar, & Scott, 1996; Kuhn, 2001). Conceptions or data can be brought into focus but seldom both together. Meta-cognitive difficulties of this kind also help to explain the ‘confirmation biases’ that have been reported, i.e. the apparent discounting of experimental results that counter preconceptions (e.g. Dunbar & Klahr, 1989; Koslowski, 1996; Mynatt, Doherty, & Tweney, 1978). This said, discounting also occurs when the conceptual implications of discordant results are recognised, suggesting that confirmation biases need to be acknowledged as a limitation in their own right. For instance, many of the 9- to 12-year olds reported in Howe and Tolmie (2003) expressed surprise at different rates of cooling from black and white containers because they ‘knew’ colour makes no difference. They usually attributed their results to defective thermometers (see Chinn & Brewer, 1993, for equivalent examples with other topics). Moreover, neuroscience research demonstrates how data that confirm preconceptions and data that disconfirm both result in brain activation (Fugelsang & Dunbar, 2005). The key difference is that only confirmatory data activate brain areas associated with learning; disconfirmatory data activate areas associated with error detection.

Nevertheless, while the problems must be recognised, it is unclear whether they should be interpreted as challenging Posner et al.’s (1982) conditions or
challenging the use of experiments to fulfil those conditions. The latter is plausible for at least two reasons. First, when experiments are used in accordance with Posner et al.’s conditions, their results have to be treated *simultaneously* as counter-evidence on initial conceptions and indicative evidence for target alternatives. Yet the cognitive literature contains numerous illustrations of difficulties with considering single events from multiple perspectives (e.g. Flavell, Miller, & Miller, 1993). Karmiloff-Smith (1992) reiterates the point in the specific context of science. Second, if experiments are sufficiently robust to be consistent with curriculum targets, they will, of necessity, be decontextualised from the lifetime of social and physical experiences that support initial conceptions. They may therefore appear isolated and meaningless rather than momentous. So the question is raised of whether an alternative to experiments can be identified to fulfil Posner et al.’s conditions, and if it can be whether it has positive implications for conceptual change. The remainder of this paper attempts to demonstrate that, for some areas of science at least, the answer may be affirmative.

Specifically, a non-experimental strategy for promoting conceptual change is described that concurs with Posner et al. but: a) separates the challenging of existing conceptions from the highlighting of target alternatives; b) calls on experiences that are just as familiar and personally meaningful as the ones that support initial conceptions. An encouraging evaluation study is then reported.

An alternative strategy for promoting conceptual change

*The potential relevance of tacit understanding*

The strategy stemmed from evidence that, for some areas of science, the conceptions that students call upon when reasoning represent only a proportion of their full
conceptual knowledge. Reasoning, no matter whether it involves planning actions, predicting events or interpreting outcomes, requires explicit engagement with conceptual knowledge, i.e. events must be related to underlying conceptions and relations must be reflected on and used to draw inferences. Thus, reasoning tasks are inherently more demanding than tasks, such as much routine action but also judgments of familiarity, that require only tacit engagement, i.e. tasks that necessitate relating events to conceptual knowledge, but do not demand reflection or inference. Substantial differences have been demonstrated between the conceptual knowledge displayed when tacit engagement suffices and the conceptual knowledge displayed during reasoning. Unsurprisingly, given the lesser task demands, these differences usually suggest that the former is superior. For instance, Reed, McLeod and Dienes (2010) found that skilled ball players, who reliably make the eye gaze adjustments that allow balls to be caught, show no comprehension of the role of eye gaze when explaining how catching is achieved. Displaying scenarios where rolling cylinders of varying mass collide with a stationary cylinder, Kotovsky and Baillargeon (1998) found that 6-month-old infants show surprise when post-collision speed fails (non-naturally) to be proportionate to the rolling cylinders’ mass but not when it is proportionate. Yet Inhelder and Piaget (1958 – later Howe, 1998) found that throughout the primary school age range, children have difficulties reasoning about the distances that objects of varying mass roll along tracks.

It was the ability to differentiate between natural and non-natural events, illustrated in the work of Kotovsky and Baillargeon, which suggested the strategy to be developed below. In principle, the ability seemed to offer opportunities for separating evidence that challenges initial conceptions from evidence that supports curriculum targets and for employing evidence that is familiar and meaningful. As
regards separation, it seemed possible that, with topics where the ability to
differentiate exists, events could be displayed that reflect students’ erroneous
reasoning in the expectation that the non-naturalness will be recognised, potentially
creating dissatisfaction with existing conceptions but not in itself introducing
alternatives. As an independent yet coherent step, authentic events could be presented
subsequently, anticipating that their naturalness too will be recognised. Through this
second step (and in a fashion that mirrors the role intended for experimental results in
traditional teaching programmes), target conceptions should be highlighted. Because
the recognition of naturalness and non-naturalness calls upon knowledge which
students already possess (and may, in some cases, have possessed since infancy), it
also employs evidence that is fully grounded in past experiences. Such evidence will
be at least as familiar and personally meaningful as the data that support unorthodox
conceptions in reasoning. Indeed, it fulfils one of Chinn and Brewer’s (1993) main
criteria for credible evidence, ‘data that the individual already believes’ (pp.25).

In short, the suggested strategy involves engaging tacit understanding through
tasks that require discrimination between natural and non-natural events in the hope of
boosting the conceptual knowledge that is displayed when reasoning. To provide a
concrete prototype (and as a prelude to evaluation), the strategy has now been
embodied in computer software, in particular in a programme that starts by requiring
events to be predicted, i.e. a reasoning task. It then presents simulations of predicted
events, followed with simulations of how events proceed in reality and invitations to
make comparisons. Simulations were essential given the need to present non-natural
displays when predictions are incorrect, and they have the advantage of sidestepping
the ‘failed experiment’ problem outlined above. While questions have been raised
about whether students treat simulations as genuine representations (Hennessy &
O’Shea, 1993), they have been used successfully in work with ‘alternative realities’ (e.g. Hennessy, Twigger, Driver, O’Shea, O’Malley, Byard, Draper, Hartley, Mohamed, & Scanlon, 1995), albeit work that juxtaposes natural and non-natural events in complex reasoning tasks rather than uses them merely to stimulate recognition of naturalness.

**Object fall software**

The topic selected for software implementation was patterns of motion when objects fall after being dropped from some height. Student reasoning about both the direction and speed of motion has been extensively explored, with significant errors reported. With direction, the focus has been upon the paths along which objects travel when they fall after moving horizontally, e.g. when balls roll over cliffs or litter is dropped from moving vehicles (e.g. Anderson, Tolmie, Howe, Mayes, & Mackenzie, 1992; Eckstein & Kozhevnikov, 1997; Eckstein & Shemesh, 1989; Hood, 1995; Kaiser, Profitt, & McCloskey 1985; Krist, 2000; Marioni, 1989; McCloskey 1983; Whitaker, 1983). Regardless of student age (from preschool- to undergraduate-level), the main finding is that objects are variously predicted to fall vertically, travel backwards, fall diagonally forwards, or continue horizontally in space before falling vertically.

However, they are seldom predicted to trace the parabolic paths in a forward direction that they actually follow. With speed, the emphasis has been upon fall from rest, with two main messages that again apply across a wide age range (e.g. Baker, Murray, & Hood, 2009; Cahyadi & Butler, 2004; Champagne, Klopfer, & Anderson, 1980; Chinn & Malhotra, 2002; Gunstone & White, 1981; Hast & Howe, 2009; Howe, 1998; Nachtigall, 1982; van Hise, 1988). The first is that when objects vary only in
mass heavy items are typically predicted to fall faster than light items rather than
tavel at speeds that, even taking air resistance into account, are usually almost
identical. Another is that, regardless of mass, objects are expected to reach maximum
velocity quickly, and then fall with constant or decelerating velocity.

At the same time, there is evidence that students are capable of differentiating
between natural and non-natural displays of object fall in precisely the contexts that
are associated with faulty reasoning. For instance, Kaiser, Profitt, Whelan and Hecht
(1992) found that when undergraduates viewed computer-simulated kegs falling from
aeroplanes, they consistently judged forward parabolas as natural and other
trajectories as non-natural. Kim and Spelke (1999) obtained equivalent results with
preschool children from two years of age, using judgment tasks and tasks that
involved monitoring of eye gaze as indices of surprise. When Shanon (1976)
presented videotapes of balls falling with constant or accelerating velocity to an
undergraduate sample, he found that constant velocity was consistently judged as non-
natural, while acceleration was regarded as natural. Moreover, using a series of
computer-simulations, the present authors [Reference to be included after review]
found primary school children capable of differentiating natural direction and speed
from non-natural direction and speed during fall: a) from rest; b) after horizontal
motion; c) through air alone; d) through air and then water. The children were much
less successful at predicting direction and speed with comparable scenarios.

With evidence across a wide age range for tacit understanding that surpasses
the understanding displayed in reasoning, object fall was regarded as an appropriate
(and educationally relevant) topic. Accordingly, the software, which was programmed
using Macromedia Director, used scenarios that showed a girl in a hot air balloon,
holding and then dropping a ball, these scenarios having been employed in the
authors’ preceding research. Sixteen scenarios were used, organised into two blocks of eight. One block addressed the direction of fall, and the other addressed changes in speed\(^1\). Within each block, the scenarios depicted all possible combinations of whether: a) the balloon was stationary or moving at the time the ball was released; b) the ball fell through air alone or through air and then into a swimming pool; c) a green or silver ball was used, with the green ball shown via a real equivalent to be very light and the silver ball (of identical size) shown to be extremely heavy. Each time the task was presented, the order of blocks and order of scenarios within blocks were randomly varied via the computer programme. At the start of each scenario, an instruction to notice which ball was being used appeared beside a close-up of the girl and ball (see Figure 1a). Clicking ‘Ready’ caused the picture to zoom out so the whole scenario was visible. An instruction to notice if the balloon moves appeared (see Figure 1b), with clicking ‘Go’ activating the scenario, i.e. with stationary scenarios the ball was released and with moving scenarios the balloon moved part way across the screen prior to release. [Insert Figure 1]

At the moment of release, the action froze, and predictions of subsequent motion were invited. With the direction scenarios, the prediction phase began with the appearance of three small white circles under the ball: a) directly below; b) behind; c) in front (see Figure 1c). This was accompanied with an instruction to select the point that the ball would travel through. Once a circle was selected, this turned red and the other circles disappeared. At the same time three further white circles appeared below the selected circle: a) if the selected circle was directly under the ball, the new circles were directly below, behind, and in front; b) if the selected circle was behind the ball, the new circles were directly below this circle, parabolically behind and diagonally behind; c) if the selected circle was in front of the ball, the new circles were directly
below, parabolically in front and diagonally in front (see Figure 1d). The instruction to select a point also re-appeared. Once a second circle was chosen this too turned red, the other circles disappeared, and three further circles appeared below the second circle in the same relative positions, accompanied with the instruction to select. While the first and second circles were always located in air, the third set varied with medium, i.e. just above the grass or under the water’s surface (just below with the green ball depicting where it would naturally float; on the pool’s floor with the silver ball reflecting the fact it would sink).

With the speed scenarios, the prediction phase involved answering questions, with response options (the words that are capitalised below) presented under each question and selected by clicking with the computer mouse. Order of options was varied randomly each time the task was presented. With scenarios where the ball fell only through air, questioning began with: ‘As the ball falls through the air will the speed CHANGE or stay the SAME?’ If this question was answered correctly via CHANGE, a second question followed: ‘As the ball falls will its speed get SLOWER or FASTER?’ There was no second question if the initial question was answered incorrectly via SAME. With scenarios where the ball fell through air and water, questioning began with: ‘Will the ball fall at the SAME speed through the air and the water or at DIFFERENT speeds?’ The follow-up question after correct selection of DIFFERENT was: ‘Will the ball travel slower/faster through the AIR or the WATER?’ with use of ‘slower’ or ‘faster’ determined randomly for each presentation. Once more, there was no second question after incorrect selection of SAME.

With both the direction and speed scenarios, correct predictions triggered the message ‘Well done! You are correct’ inside a gold star to the right of the screen. This was accompanied with ‘Click upon GO to see the ball fall’, with clicking activating a
simulation of the ball falling naturally. Incorrect predictions triggered ‘Click on GO to see what you thought would happen’, with clicking this time activating a simulation of the ball falling non-naturally in the predicted fashion. Once the latter simulation was complete, ‘Did it look correct?’ appeared to the right of the screen, with YES and NO as response options. Selection of YES was not expected to occur very often, but if it was chosen ‘Sorry the answer is WRONG’ appeared on the right of the screen accompanied with ‘Click on NEXT to see what really happens’. Selection of NO triggered ‘Yes, the motion of the ball was incorrect’, together with the invitation to see what really happens. In any event, indicating that the incorrect motion was in fact incorrect triggered ‘What is the difference between the incorrect and correct motion?’ and then ‘Why are they different?’ The option was available of replaying the incorrect and/or correct motion as many times as was deemed necessary to answer the questions. The computer programme was designed to record details of button presses at all stages, including predictions.

Method

Overview

With the strategy implemented in software, the primary issue was whether the software had positive implications for conceptual change, and this was the issue that the evaluation study was intended to resolve. It was recognised that a positive outcome would raise further questions, e.g. whether the software was effective because it implemented the strategy, but it seemed advisable to address one issue at a time. Accordingly, after securing university ethical approval, a three-stage study was conducted with a sample of 8- to 12-year-old children. This age group was selected
because: a) the topic area is consistent with curriculum requirements for the 8- to 12-year range (DfEE/QCA, 1999); b) the authors’ previous research [Reference to be included] indicates that the materials are comprehensible for children in this age range, and interest can be maintained across 16 scenarios.

During the first stage of the study, the sample was pre-tested to establish baseline reasoning about object fall. For the second stage, about two-thirds of the pre-tested sample were assigned to work with the software, with the remaining third intended as an untutored control. Since the evaluation was designed merely to ascertain whether the software supported conceptual growth, comparison with an untutored control group was appropriate, rather than, for instance, comparison with a group that studied object fall in an alternative fashion. Half of the children selected to work with the software (i.e. one third of the pre-tested sample) did this individually, with an adult on hand to highlight relevant features and offer guidance upon request. The other half worked with a classmate, with the adult offering guidance upon request but otherwise being deliberately passive. As noted, social-constructivist analyses have drawn attention to the cultural context in which procedures derived from Posner et al. (1982) are situated, emphasising discourse practices. While exploring cultural influences was not a primary aim, it was recognised that the software could not be presented in a contextually ‘neutral’ fashion. Accordingly, an attempt was made to mimic two contexts, which not only occur frequently during computer-based sessions in classrooms (Crook, 1994) but are also known to be associated with markedly different patterns of discourse (Anderson, Tolmie, McAteer, & Demissie, 1993). The final stage of the study involved a post-test, which was administered to the children who worked with the software and to the control group.
Sample

The sample was recruited from Years 4 to 6 of a state-sector primary school located in a predominantly lower middle-class area of Cambridge, UK. With every parent consenting to participation, all children in the relevant age range were potentially included. However, absence from school meant that six children completed only one of the pre- and post-tests. With these children excluded, the sample comprised 139 children (77 girls). Forty-eight children were from Year 4, 47 were from Year 5, and 44 were from Year 6, with a sample mean age of 9.73 years ($SD = 0.95$).

Pre- and post-tests

The pre- and post-tests each comprised 16 items, which were identical across the two tests. Each item was associated with one of three topics: a) a ball falling from a hot air balloon (as with the software), $n = 5$ items; b) a box falling from a helicopter, $n = 6$ items; c) a box falling from a train on a high bridge, $n = 5$ items. The items were selected from a pool of 32 items that had previously been piloted with a similar aged (but different) sample of children. Rasch analysis of pilot predictions (Bond & Fox, 2007) was used to identify items of varying difficulty, whose ‘fit values’ were nevertheless consistent with a uni-dimensional scale. Sufficient items fulfilled these criteria to permit the selection of eight items that addressed the direction of fall and eight items that addressed speed change, while also ensuring that each set of eight covered all possible combinations of fall from rest vs. after motion, fall onto grass vs. into water, and fall of a heavy vs. light object. To facilitate presentation, the items were organised into two blocks, with one block addressing direction and the other block addressing speed. There was a fixed order of blocks during presentation and a
fixed order of combinations within blocks, although both orders were initially
determined at random. For presentation purposes, each test item comprised a coloured
illustration of the scenario accompanied with text.

With each of the eight direction items, the text included instructions to
indicate where the object lands by marking the illustration with a cross and then to
draw the path that the object follows as it falls. With the four speed items that
addressed fall onto grass, the text was ‘As the ball/box falls through the air, will the
speed of the ball/box (varied as appropriate for the scenario)?’ with ‘Get faster’, ‘Get
slower’ and ‘Stay the same’ listed below as possible answers. With the four items that
addressed fall into water, the text was ‘Will the ball/box fall at the same speed
through the air and the water or at different speeds?’ again accompanied with three
possible answers, ‘Faster in the air’, ‘Faster in the water’ and ‘Same speed in air and
water’. In all eight cases, the possible answers to the speed items were presented in
randomly varying orders as multiple-choice options, with response boxes to tick.

Four direction items and four speed items also invited explanations of predicted
paths/speeds, offering six multiple-choice options (again in randomly varying orders)
as possible factors: ‘A force pulling the ball/box downwards’, ‘The balloon/helicopter
/train’s direction before the ball/box started falling’, ‘The air or water pushing the
ball/box upwards’, ‘There is no wind’ (absence of wind was highlighted in the task
instructions – see below), ‘The air or water pulling the ball/box downwards’, ‘The
ball/box’s weight’. With the direction items, the first two of the above factors were
regarded as relevant, and the remaining four as irrelevant (but indicated in the
background literature as often used in reasoning). With the speed items, the first and
third factors were regarded as relevant, and the remaining four as irrelevant.
Procedure

In accordance with the school’s preference, the children completed the pre-test in year groups, i.e. two classes together, sometimes in the library and sometimes in the room normally used by one of the classes. Each child was issued with an answer booklet that contained the test items in sequence (i.e. coloured illustrations and text), and invited to insert their name, school, class and gender on the front cover, together with the date. Thereafter, a researcher took the group through the items, displaying them in sequence on a large screen using PowerPoint, ensuring that the children were looking at the corresponding item in their booklets, talking them through what the task involved, and inviting them to enter responses in their booklets. The researcher also highlighted features that were hard to discern from the booklets alone, i.e. whether the carrier was stationary or moving prior to the ball/box’s release, whether the ball/box was light or very heavy, and the intended absence of wind. The class teachers were always present during the pre-test to assist with procedural questions and to ensure that there was no copying of neighbours (but not to help with item solution). The test took approximately 25 minutes to complete.

Roughly one third of each class was assigned at random to each of the second-stage conditions, i.e. individually but with adult guidance (designated, for brevity, the ‘single’ condition), with a classmate but with minimal adult guidance (designated the ‘pair’ condition), and untutored (designated the ‘control’ condition). This resulted in 44 children being placed in the single condition, 48 being placed in the pair condition and 47 being placed in the control group, with no significant differences between the conditions over gender composition ($X^2(2) = 1.44, ns$) or school year ($X^2(4) = 0.17, ns$). The children assigned to the pair and single conditions came to a private room at their school, where the researcher who had administered the pre-test welcomed them,
provided an overview of the procedure, invited participation and, with consent to continue (given in all cases), obtained basic demographic information. Thereafter she directed the children to a Dell Latitude D820 laptop computer, and showed them how to respond using the computer mouse.

With the children who were working in pairs, the researcher then withdrew to another part of the room, only intervening on the rare occasions that she was asked for procedural advice. On the other hand, she stayed with the children who were in the single condition, again offering procedural assistance upon request but this time also highlighting key features. For instance, she asked the children to describe and explain the difference between the incorrect and correct motion after these had been viewed, scaffolding the use of appropriate explanatory factors. Regardless of condition, the researcher used a grid to record whether each software scenario was associated with: a) discussion of the prediction; b) discussion of the simulated motion; c) analysis of differences between correct and predicted motion; d) interventions from the researcher. Brief notes were made of discourse content. Upon completion of the computer-presented task (which took between 20 and 30 minutes), all children wrote individual answers to two questions: a) when an object falls what makes a difference to the path that it travels through as it falls; b) what makes a difference to the speed of an object when it falls? Finally, between two and four weeks after the single and pair children worked with the software, the post-test was administered, following procedures that were identical to the pre-test.

**Scoring**

Five measures of conceptual knowledge were obtained from the pre- and post-tests:
• Correct predictions - cross (CP/C) = Number of direction items where the cross was positioned at the correct landing point (Maximum score possible = 8.00, with eight direction items)

• Correct predictions - path (CP/P) = Number of direction items where the path from the balloon to the cross was drawn correctly (Maximum score possible = 8.00, with eight direction items)

• Correct predictions - speed (CP/S) = Number of speed items where both prediction questions were answered correctly (Maximum score possible = 8.00, with eight speed items)

• Correct explanations (CE) = Number of correct factors selected to explain predictions across the direction and speed items (Maximum score possible = 16.00, with explanations solicited for eight items and two correct options provided with each item)

• Incorrect explanations (IE) = Number of incorrect factors selected to explain predictions across the direction and speed items (Maximum score possible = 32.00, with explanations solicited for eight items and four incorrect options provided with each item).

Results

As noted, the pre- and post-test items were selected from a larger pool, which had been tested with a pilot sample. Rasch analysis of the predictions made during piloting was used to create a uni-dimensional scale of varying item difficulty. As a result, the CP/C, CP/P and CP/S measures can be assumed to be reliable. On the other hand, explanations were not subject to Rasch analysis during scale development (even...
though explanation items were also used during piloting), and therefore the reliability of the CE and IE measures could not be assumed a priori. As a result, Cronbach’s alpha was calculated for the CE and IE measures as a post hoc test of reliability. Acceptable outcomes were achieved (for CE, $\alpha = 0.70$; for IE, $\alpha = 0.76$).

Accordingly, analyses were conducted on the total scores that the children obtained on each of the five measures, without any need to examine specific items separately.

The key question was whether working with the software proved beneficial. To address the question, mixed-model two-way ANOVAs were conducted on the CP/C, CP/P, CP/S, CE and IE scores obtained at pre- and post-test, with test (pre, post) as a within-participants factor and condition (single, pair, control) as a between-participants factor. With the prediction measures (CP/C, CP/P, CP/S), significant main effects of test were obtained in all three cases (for CP/C, $F(2, 136) = 52.58, p < 0.001$; for CP/P, $F(2, 136) = 34.62, p < 0.001$; for CP/S, $F(2, 136) = 7.40, p < 0.01$). With all three measures, performance was significantly higher at post-test than at pre-test. The main effects of condition were non-significant, but with all three measures there were significant interactions between test and condition (for CP/C, $F(2, 136) = 8.09, p < 0.001$; for CP/P, $F(2, 136) = 5.18 p < 0.01$; for CP/S, $F(2, 136) = 5.16, p < 0.01$). One-way ANOVAs were conducted to clarify these interactions. As can be seen in Table 1, there were no significant differences between the conditions at pre-test on any of the three prediction measures. However, by post-test, all differences were statistically significant, with the control group invariably performing below the children who worked with the software. In fact, the mean post-test scores obtained by the control group were little different from their mean pre-test scores, with mean CP/S scores actually declining slightly across the two tests. In other words, it was only the children who worked with the software who progressed from pre- to post-test. This
said, while progress was detected in both the single condition and the pair, it was particularly marked with the pairs. [Insert Table 1]

As regards the factors that influence direction and speed, conceptual growth would be signaled by pre- to post-test increases in CE scores and pre- to post-test decreases in IE scores. With this in mind, Table 1 suggests modest benefits from the software for CE scores in the pair condition and for IE scores in the single condition. The control group, by contrast, obtained lower CE scores at post-test than pre-test and higher IE scores. Nevertheless, all changes are modest, and the two-way ANOVAs produced no significant main effects of test or condition and no significant interactions between test and condition. Follow-up analyses are not therefore appropriate. Overall then, the software had unmistakable benefits for the conceptual dimensions that were its actual focus (patterns of direction and speed), particularly when the children worked in pairs with a classmate. However, its implications for the underlying determinants of motion were, at best, embryonic.

Thinking about why it was the CP/C, CP/P and CP/S measures that were associated with the strongest gains, it may be significant that, as Table 1 makes clear, pre-test scores on all three measures were considerably below the maximum possible. Thus, making appropriate predictions was a significant challenge, and this may have focused the children’s attention on the prediction element, i.e. inputting the anticipated motion and receiving feedback from the simulations. They may have paid less attention to the stages at which explanations were highlighted, i.e. interpreting why the predicted and correct motion differed and answering the two written questions at the end. Indeed, if the children’s focus can be assumed to have been upon predictions, the differences between the single and pair conditions may also become interpretable. Table 2 shows the mean number of scenarios on which the researcher
recorded the four discourse categories, i.e. total number of scenarios across each condition where the category was recorded divided by 44 (single condition) or 24 (pair condition). The most striking difference relates to discussion of the predictions, which was never recorded in the single condition but was the most frequently used discourse category in the pair condition. With 63% of the scenarios around which prediction discussion occurred, the focus was on predictions alone, i.e. asserting possibilities and confirming or challenging these. With the remaining 37%, there was also an attempt to justify positions. In either case, such discussion would have been highly salient, if formulating predictions was the children’s major focus. Therefore, the fact that it occurred only with the pairs could be significant. [Insert Table 2]

**Discussion and implications**

The design of the object fall software was informed by three considerations. The first was that Posner et al.’s (1982) conditions for effecting conceptual change have normally been interpreted as requiring experimental appraisal of initial conceptions. It has been assumed that results will simultaneously challenge these conceptions and render the target science as comprehensible and plausible. The second consideration related to some potentially misleading implications of using experimental appraisal, and therefore of its patchy success in supporting conceptual growth. Specifically, questions were raised about an approach that necessitates viewing evidence simultaneously from two perspectives, and uses data that is inevitably dissociated from personal experience. The third consideration was the possibility of engaging students’ tacit understanding both to separate the creation of dissatisfaction from the introduction of alternatives, and to utilise experiences that are familiar and personally meaningful. However, while the three considerations informed the software’s design,
at no point were they transformed into hypotheses that the evaluation study aspired to test. Rather, the evaluation study was concerned with the effectiveness of materials that embody the third consideration, and therefore relate only indirectly to the first and second. With this recognised, the following discusses what can be inferred from the results regarding the software’s effectiveness, then considers the potential relevance of tacit understanding for science education in general, and finally makes brief comments about the first and second considerations.

As regards the software’s effectiveness, the children who worked with it made significantly greater progress than control children in the conceptual understanding underpinning predictions. This was true for both direction of fall and changes in speed. Less progress was however detected for explanatory factors. With reference to pre-test scores (i.e. as shown in Table 1), it was suggested that even the formulation of predictions would have proved challenging for the participating children, and therefore this may have pre-occupied them to the exclusion of background factors. Certainly, if the children were pre-occupied with predictions, it was more likely due to the formidable conceptual content than, say, to practical difficulties with operating the software. As Table 2 shows, ‘researcher intervention’ occurred relatively infrequently, implying that the children quickly mastered what they were expected to do and found the software easy to use. This of course augurs well for eventual application in classrooms. However, even if conceptual challenge did result in a focus upon predictions, it is unclear whether this is sufficient to account for the limited progress with explanatory factors. To explore further, it would be desirable to work with samples whose pre-test scores indicate stronger initial knowledge, perhaps because they had benefited from using the software on previous occasions. Indeed, the implications of repeated software usage would be of general interest.
While the software was generally effective, its benefits were most marked for the children who worked in pairs with a classmate. Progress was detected amongst the children who worked individually under researcher guidance, but to a lesser degree. Assuming pre-occupation with predictions, it was proposed earlier that the pairs’ superiority might be related to the fact that discussion of predictions only occurred between classmates. If discussion was significant, it could have been for at least two reasons. On the one hand, the significance could have stemmed from the exchanges (roughly one-third of the total prediction discussions) where the children offered justifications for their predictions. Justificatory dialogue has been shown on numerous occasions to be beneficial for conceptual understanding (see Howe, 2010, for a review of many studies, including ones relevant for science). On the other hand, the dialogue’s content may not have been critical in this particular context. Its very occurrence might have been sufficient to slow the children down and make them think carefully before making predictions. Actually, this second interpretation is probably the most plausible: the justifications usually referred to explanatory factors and therefore had they been crucial for boosting predictions, growth in explanatory understanding might also have been expected (when, as noted, this was negligible). In any event, assuming discussion of predictions was significant for some reason, it becomes important to reflect on its non-occurrence in the single condition. It was certainly not precluded in principle, given the researcher’s availability for guidance. Indeed, the researcher was approached at other points in the task (see Table 2). Perhaps children regard adult help during initial problem solving as illegitimate.

The suggested interplay between task difficulty, discourse patterns and expectations about adult involvement imply a need for caution before extrapolating beyond the object fall context. It remains to be seen whether equivalent software
would prove beneficial with topics that students find easier or harder, and whether at any level of difficulty (or with any student sample) working with classmates is more helpful than working individually with adult support. All such issues are worth exploring, for there are likely to be many areas relevant to science where students display tacit understanding that surpasses the conceptual frameworks they employ when reasoning. Earlier, examples were given relating to horizontal motion as well as vertical, and although currently un-researched, it seems likely that tacit understanding exists in domains beyond force and motion. Heat transfer and light propagation are obvious candidates. Descriptive studies of students’ conceptions typically overlook the tacit dimension, sometimes even implying that the constructs displayed when reasoning define conceptual knowledge. Students are said to hold misconceptions, when a more accurate claim would be that misconceptions are used with particular reasoning tasks. Thus charting tacit understanding requires theoretical re-positioning as well as additional research. On the other hand, once such understanding is identified, it would be straightforward to design software that emulates the object fall structure, and via this obtain a comprehensive picture of how task difficulty, discourse patterns and adult involvement interact to determine learning outcomes.

Positive results using such software would confirm and extend the findings from the present study, that reasoning tasks, which engage tacit understanding, can boost the conceptual level of reasoning itself. However, based on comparisons with untutored control groups, results of this kind would not show whether the boosts were actual consequences of tacit engagement, just as the present study does not show this either. Although the issue was, as noted, regarded as premature for the study, there is no reason why it should not be addressed in the future. The best strategy might be to compare across a series of partial versions of the present software, e.g. software that
was equivalent except for requiring recognition of the incorrectness of non-natural
events, and equivalent except for requiring recognition of the correctness of natural
events. Indeed, such ‘decomposition’ would have to be taken a step further if the
interest was not simply in whether tacit understanding contributes but also in how it
contributes. As noted, the engagement of tacit understanding has the potential to
separate the creation of dissatisfaction from the introduction of alternatives, and to
utilise experiences that are familiar and meaningful. Researchers might wish to assess
their relative contribution to positive outcomes. While such issues can be flagged here
only as matters for future research, there is significance already in the fact that tacit
understanding has this twin potential, and that software which accesses this
understanding can support conceptual growth. At minimum, it shows that viable
alternatives to traditional experimental appraisal do exist, and it is therefore premature
to draw pessimistic inferences about conceptual change from studies that examine
experimental appraisal to the exclusion of these alternatives. In other words, the study
highlights the fact that conclusions to date have been grounded in research that adopts
an overly restricted framework, which can and should be addressed.

While the object fall software avoids experimental appraisal as traditionally
understood, it remains consistent with Posner et al.’s (1982) conditions. It did, after
all, attempt to unsettle existing conceptions and to highlight the comprehensibility and
plausibility of target alternatives. Its challenge is not therefore to the broad paradigm
that Posner et al. stimulated, but to the equation of this paradigm with experimental
appraisal and the conclusions about efficacy that equation has occasioned. Indeed, the
software is not merely consistent with teaching within Posner et al.’s framework; it
could perhaps also be employed to support such teaching in classrooms. While
cautions must be exercised before extrapolating from a one-off, researcher-delivered,
out-of-class investigation, it is encouraging as regards routine practice that the software was easy to use as well as effective. Furthermore, comparison of individual usage under adult guidance and usage with a classmate means that a major source of discourse variation during authentic computer-based teaching has been addressed. At the very least, it would be interesting to examine the consequences of using the software and/or equivalent software with other topics within normal teaching. The range of potential topics is not unlimited; it would be surprising, for instance, if children have much tacit understanding of the earth in space or Mendelian inheritance. However, as noted already, a reasonable range is probable, meaning that in the fullness of time software that engages tacit understanding might have a significant role to play. For the present, all that can be said is that conceptual change needs to be supported within school science, experimental appraisal of existing conceptions is not an unqualified success, and tacit understanding has the potential to take things forward.

Footnote

1 The software can be downloaded for viewing from [website to be included after review]. Please click on ‘2B1’ for the direction of fall block, and ‘2B2’ for the speed change block.

Acknowledgements

The authors wish to thank the Economic and Social Research Council of Great Britain for funding the research, Pepi Sarvary for constructing the software, and the participating children, their teachers and head teachers for enthusiastic engagement.
References


Table 1. Mean Scores at Pre-test and Post-test (SD in brackets)

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th></th>
<th>Post-test</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single</td>
<td>Pair</td>
<td>Control</td>
<td>F (2, 136)</td>
<td>Single</td>
<td>Pair</td>
<td>Control</td>
<td>F (2, 136)</td>
</tr>
<tr>
<td>CP/C</td>
<td>3.98 (1.27)</td>
<td>4.25 (1.30)</td>
<td>4.42 (1.68)</td>
<td>1.13, ns</td>
<td>5.61&lt;sub&gt;ab&lt;/sub&gt;</td>
<td>6.04&lt;sub&gt;b&lt;/sub&gt;</td>
<td>4.70&lt;sub&gt;a&lt;/sub&gt;</td>
<td>5.45, p &lt; .01</td>
</tr>
<tr>
<td>CP/P</td>
<td>3.30 (1.21)</td>
<td>3.35 (1.34)</td>
<td>3.49 (1.04)</td>
<td>0.31, ns</td>
<td>4.52&lt;sub&gt;b&lt;/sub&gt;</td>
<td>4.64&lt;sub&gt;b&lt;/sub&gt;</td>
<td>3.70&lt;sub&gt;a&lt;/sub&gt;</td>
<td>4.03, p &lt; .05</td>
</tr>
<tr>
<td>CP/S</td>
<td>5.55 (1.90)</td>
<td>5.42 (1.92)</td>
<td>5.47 (1.95)</td>
<td>0.05, ns</td>
<td>6.14&lt;sub&gt;ab&lt;/sub&gt;</td>
<td>6.54&lt;sub&gt;b&lt;/sub&gt;</td>
<td>5.21&lt;sub&gt;a&lt;/sub&gt;</td>
<td>5.91, p &lt; .01</td>
</tr>
<tr>
<td>CE</td>
<td>6.61 (3.21)</td>
<td>6.79 (2.67)</td>
<td>7.23 (2.94)</td>
<td>N/A</td>
<td>6.39</td>
<td>7.31</td>
<td>7.11</td>
<td>N/A</td>
</tr>
<tr>
<td>IE</td>
<td>14.57 (5.27)</td>
<td>14.35 (5.03)</td>
<td>14.11 (4.15)</td>
<td>N/A</td>
<td>14.20</td>
<td>14.38</td>
<td>14.57</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Note

When post-test values for CP/C, CP/P and CP/S have different subscripts, values are significantly different (Bonferroni, p < 0.01)
Table 2. Mean Number of Scenarios Associated with Each Discourse Category (Maximum possible = 16)

<table>
<thead>
<tr>
<th>Discourse Category</th>
<th>Single (N = 44 children)</th>
<th>Pair (N = 24 pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discuss prediction</td>
<td>0</td>
<td>3.00</td>
</tr>
<tr>
<td>Discuss simulation</td>
<td>0.25</td>
<td>0.33</td>
</tr>
<tr>
<td>Compare correct and incorrect</td>
<td>0.16</td>
<td>0.33</td>
</tr>
<tr>
<td>Researcher intervention</td>
<td>1.20</td>
<td>1.38</td>
</tr>
</tbody>
</table>
Figure 1. Sample screen displays

a

b.
c.

![Image of a hot air balloon](image1.jpg)

Select the next point you think the ball will travel through.

---

d.

![Image of a different hot air balloon](image2.jpg)

Select the next point you think the ball will travel through.