Memory & Cognition
1994, 22 (4), 494-502

Relative consistency and subjects' "theories" in domains such as naive physics: Common research difficulties illustrated by Cooke and Breedin

MICHAEL RANNEY
University of California, Berkeley, California

While augmenting the literature with data that further exhibit context-specific responding to qualitative motion problems, Cooke and Breedin (1994) exhibit common theoretical and methodological difficulties that undermine their conclusions. Herein, these flaws are explicated and contrasted with features of studies that avoid the pitfalls of (1) theoretical vagueness, (2) overly coarse data aggregation, (3) nondiagnostic, errorful assessment items, and (4) imprecise measures of the variety of (mis/)conceptions (e.g., of "impetus," or inertia). The difficulties call into question Cooke and Breedin's claims that impetus ideas play minor roles in performance and that "naive theories" of motion are largely constructed on line. Because such confusion often arises from the polysemy of "theory," some empirical criteria for "theoryness" are discussed, including subjects' conceptual, temporal, and coherence-based consistencies (regarding researchers' models and isomorphs). While naive physics may be idiosyncratic, baroque, context-driven, and apparently inconsistent, it might (additionally) be based upon fairly a priori, systematic, and temporally stable information.

The cognition of physics, especially of motion, is an increasingly active research field, partly because physics problems are difficult yet formal enough that "answers" can seem more at hand than for other domains (e.g., "What is the trajectory?" vs. "How much of behavior is hereditary?"). There are also well-known findings that some apparent patterns of errors are evidenced by subjects, historical figures, and even ourselves. But what *are* these patterns—especially the one called *impetus?* Are they (1) naive physics "theories" (e.g., McCloskey, 1983), (2) physics "misconceptions" (e.g., McCloskey, Washburn, & Felch, 1983, contra Smith, diSessa, & Roschelle, 1993; cf. Masson, Hill, Conner, & Guindon, 1988), or (3) the results of a fragmented understanding of physical phenomena (e.g., diSessa, 1983, 1988; Ranney, 1987/1988, in press; Ranney & Thagard, 1988)? Such questions, unfortunately, dance about polysemous, ill-specified, and/or politically laden words such as "misconception" and "theory." (Hence, one should assume that, herein, such words are always implicitly quoted.)

Many have suggested criteria for "theoryness," but consistency, coherence, and a relative absence of contradictions seem most appropriate here (Ranney, 1987/1988, 1994, in press). Cooke and Breedin (1994; hereafter referred to as C&B) offer data regarding consistency, but their conclusions are undermined by their coarse methods and levels of analysis. Still, their work adds to the literature that supports "post-McCloskeyan" or "anti-Theory theory" views of dramatic contextual, featural, and situational influences on subjects' physics responses (e.g., diSessa, 1983, 1988, in press; Halloun & Hestenes, 1985; Hojnacki, 1988; Kaiser, Jonides, & Alexander, 1986; Ranney, 1987/1988, in press; Ridgeway, 1992; Schank & Ranney, 1992). Because my own empirical and theoretical work supports this view of context-specific responding among naive/ novice subjects, I welcome articles that have what I call the "loose reasoning" perspective (e.g., Ranney, in press). Still, difficulties in C&B's experiments (beyond those they note) result in an article that only ambiguously supports this (and their "on-the-fly") view. A representative example of such difficulties is the vaguely uniform way in which "impetus theory" and/or "naive theories" are operationalized—in contrast to the many variants discussed in the literature (e.g., by Clement, 1983, and others mentioned below). Coupled with some problematic experimental methods and analyses of limited

Preparation of this article was supported by the National Academy of Education, the Spencer Foundation, and the University of California's Committee on Research. Special thanks are offered to Nancy Cooke and Sarah Breedin for eliciting these thoughts with their stimulating work and to Margaret Jean Intons-Peterson for offering a forum for these thoughts. This article benefited from comments by, and past conversations with, Lauren Resnick, Susan Hojnacki, Nancy Nersessian, Andy diSessa, Paul Thagard, Dale Klopfer, Patti Schank, Micki Chi, Seth Chaiklin, Nancy Cooke, Peggy Intons-Peterson, Stellan Ohlsson, Leo Klopfer, Ehud Bar-On, Jim Greeno, Jim Voss, Bill Prinzmetal, Barbara White, John Clement, Jim Minstrell, Ann Brown, John Frederiksen, Mary Kaiser, Michael McCloskey, Alphonso Caramazza, George Montoya, Chris Hoadley, Bernadette Guimberteau, Michelle Million, the Reasoning Group, and other colleagues and students. Correspondence should be addressed to Michael Ranney, 4533 Tolman Hall, EMST, University of California, Berkeley, CA 94720.

—*Accepted by previous editor, Margaret Jean Intons-Peterson*

sensitivity, C&B's theoretical uniformity yields response consistency data that are not diagnostic regarding the question of "on-line" theory construction. In essence, those who maintain that laypeople have relatively stable theories of motion will probably not be compelled by C&B's analyses.

In the present article, the limitations of C&B's method, analyses, and theoretical view are detailed and contrasted with more diagnostic research on the stability or lability of naive theories of motion; even so, C&B must be applauded for their empirical effort and the scholarly dialogue that their work extends. Thus, the thoughts below largely represent constructive devil's advocacy, a discussion of common pitfalls in assessing/describing lay theories (cf. Donley & Ashcraft, 1992), and a more general framework for considering "consistency" and "theory."

CONSISTENCY AND THEORETICAL MULTIPLICITY IN NAIVE PHYSICS

A priori, we might expect that people (and C&B's subjects) will not be consistent: (1) our processing and short-term memory capacities are limited, (2) our behavior is clearly context- and load-dependent, and (3) we regularly meet changing contexts and high cognitive loads. Without extremely advanced models of both the creature and the environment, who would expect such a subject to seem highly consistent? Furthermore, even if we were capable of complete coherence among the variety of our (often contradictory) beliefs and possible behaviors, attaining global coherence is usually not worth the effort (Ranney, in press). Hence, loose reasoning and on-line theorizing are assured in complex and/or unfamiliar domains such as physics (Ranney, Schank, Mosmann, & Montoya, 1993; Schank & Ranney, 1991, 1992).

Several researchers have studied subjects' response consistency in the naive physics domain (e.g., Halloun & Hestenes, 1985; Hojnacki, 1988; Ranney, 1987/1988, 1988, 1994). One's idea of consistency is always attended by some kind of theoretical approach, and most motion constructs involve (1) the Newtonian view of vector addition and inertia and/or (2) a variety of "impetus" notions. The "received view" of Newtonian physics is evident to most researchers (but see below). Still, debates follow virtually all attempts to define impetus theories. Halloun and Hestenes (1985) offer a useful description of various kinds of impetus beliefs found among responses to physics items. McCloskey and his colleagues have also done this; however, they aggregated considerably divergent impetus-like phenomena to offer the sense of a "theory"—and with inappropriate aggregations (even with sensitive measures), subjects will always fail tests of consistency. Although some have seemed to observe consistent naive theories (e.g., Caramazza, McCloskey, & Green, 1981), these were generally spurious results, often based on the exclusion of outlying or anomalous data (A. Caramazza, personal communication, April 25, 1988), other aggregation problems, too few observations, or rather insensitive measures (see below, and Ranney, 1987/1988).

C&B aggregate subjects' errors so coarsely (as evidence of "impetus") that the possibility of their exhibiting consistency seems questionable: They combine a curvilinear-impetus response for the tube problem with a "straight-down" response for the cliff problem, but these responses do not indicate the same "impetus" notion. It is as if the tube's ball remembered its prior (constrained, curving) motion, while the cliff's ball *neglects* its prior (horizontal) motion. Similarly, their judges might have lumped a diagonal path for the cliff task with either the straight-down or the curvilinear trajectory—but it should depend on one's *reasons* for its diagonality (see below).

Many have proposed impetus-like naive theories (some linked, for example, to Aristotle, Buridan, the Medievalists, and Galileo; see, e.g., Clement, 1983; McCloskey, 1983; Nersessian & Resnick, 1989; Shannon, 1976), and they differ in how they categorize subjects' responses. Since there are several impetus theories, "naive theory" is a misnomer (Ranney, 1987/1988); one can even select among subtypes of constructs such as dissipation, internal force, curvilinear impetus, "overcoming,"[1] and so on, to yield medleys of theories. C&B, though, aggregate responses so much that "impetus" comes to approximate "common errors," and "impetus theory" becomes a fairly undifferentiated mix of divergent misconceptions or malcombined primitives (diSessa, 1983, 1993; Ranney, 1987/1988). Indeed, by defining impetus too broadly and homogeneously (see below), C&B undermine their conclusion that impetus ideas are "unrelated to the accuracy of the associated trajectory judgments"—especially since they report that (1) the correlation between trajectory accuracy and (purportedly) impetus-related true/false answers from their physics test is negative ($-.48$) and (2) the odds of a subject being wrong, given that he/she generated an "impetus" explanation, are much greater than the odds of being wrong if he/she did not generate one (about 49% vs. $<24\%$, in their Experiment 2).

C&B seem to view their massive error aggregation as a "conservative" way to disconfirm the "hypothesis" of consistency (cf. below). They imply that, as the set of impetus errors approximates all errors, they are tallied as more abundant—hence, conservatively boosting the apparent consistency of impetus theories. But this argument need not hold; it depends on error base rates, C&B's "consistency" criteria, and their sensitivity for detecting impetus errors—each of which is challenged in the next section. For now, suffice it to say that these nondiagnostic aspects of C&B's article reflect the theoretical ease with which their method could be used to yield predictions of either the presence or the absence of consistent naive theories of motion.

METHODOLOGICAL PITFALLS

This critique of C&B's methods largely stems from the prior comments on their lack of a *well-specified* naive ("impetus") physics theory. This is not an insurmountable problem in itself. (No such well-specified theory has yet been proposed, to my knowledge.) But C&B pair this

theoretical void with a method that very coarsely "pigeonholes" subjects' responding. Many of us debate about when to use "objective" measures (e.g., categorization) versus more qualitative measures (e.g., interviews and verbal protocols), but the choice is as influenced, unfortunately, by the pragmatics of one's research resources (and Ortgeist/Zeitgeist) as by particular research questions (cf. Donley & Ashcraft, 1992; Ericsson & Simon, 1984/1993). Highly rigorous (and even quantifiable) "qualitative" methods take more resources per datum than do simpler subject- or experimenter-choice or categorization techniques (see below).

Pigeonholing (vs. Cataloging) Natural Responses With Selections and Categorization

An indicative flaw in C&B's study is the forced pigeonholing of responses into a few types that represent neither a single naive "theory" nor a common or stable variety of theories. Both their multiple-choice selection alternatives and (thus) their way of categorizing production responses show this. For instance, as evidenced by C&B's error classifications, the straight-down "misconception" is clearly distinct from the curvilinear impetus "misconception," yet these are aggregated later in their search for naive theories. (Indeed, Experiment 1's pendulum-orientation problem involves yet another sort of impetus.) So, some of C&B's problems are rather isomorphic, but the response typologies hardly seem isomorphic. C&B might have, alternatively, looked at the produced paths in a more detailed and varied way, without adding post hoc forced-choice classifications by coders. The present Appendix, from Ranney (1987/1988), shows examples from a more sensitive cataloging of the wide variety of trajectories people produce for the sorts of problems C&B used. While this approach would further reduce the comparability of C&B's production and selection data, it preserves the data's richness for theories that might describe them. Forcing production data into the small set of categories represented by C&B's selection alternatives hinders a rigorous search for theoretical/conceptual consistency. Hence, the production versus selection contrast is dubious for its lack of more response- and theory-sensitive post hoc analyses, in spite of the judges' high correspondence. This is even more true in Experiment 2, with its binary "holes," challenging its status as a replication/extension. (The binary sorting may also explain some reversals of effects between Experiments 1 and 2.)

Thus, C&B's impetus "principles" are not sufficiently articulated or related to their methods. Theoretically driven relations should exist between postulated principles and one's methodology, especially showing how responses map to misconceptions or pieces of impetus. (Otherwise, why not just use maximally divergent foils?) These relations need not be one to one (which may be hard to do), but they can be done with configurations of responses over tasks or types of data—especially using statistical techniques (e.g., multidimensional scaling; Ranney, 1988) and other methods (e.g., converging; Hojnacki, 1988, and Ranney, 1987/1988, 1994).

In discussing a theory's nature, we usually illustrate it with specific responses and explanations (cf. the above tube and cliff responses). However, one may draw a diagonal path (for instance) for degenerate configurations of reasons/"theories" (e.g., viewing gravity and lateral motion as both accelerative *or* both uniform; Ranney, 1987/1988). Furthermore, having fewer categories can actually reduce apparent consistency by ignoring subtleties among drawn paths. After losing subtleties by pigeonholing, the data may no longer be properly aggregated. Consider a common cliff-standard response that moves horizontally, then curves down, then moves straight down ("H,C,S" in the present Appendix). In Experiment 1, C&B might have coded it as physically similar to any of the four categories—from correct to wrong—depending on how the subject and judges (regardless of judge agreement) interpret the path's features (e.g., shape, landing spot, corners, fit, etc.). The features hold the keys to subjects' potential theories or consistencies. Cutting the alternatives to two (C&B's Experiment 2) may lower the odds of finding consistency, as more subtly different paths are theoretically miscategorized/pigeonholed. (Since the sortings are based on physical—not conceptual—similarity, the paths are seemingly categorized largely on accuracy.) This helps explain why C&B show a drop of "consistently impetus-explaining" subjects from Experiment 1 (5%) to Experiment 2 ($<1\%$) as the categories were halved. It also seems that, by reducing (and confounding) both the choices per problem and the set of problems between Experiments 1 and 2, yet keeping the same (explanation) consistency criteria, one might expect less consistency (as was observed), even though C&B suggest that reducing the choices should increase observed consistency. These points emphasize why C&B should characterize their subjects' drawn trajectories in a more ecological fashion and not force them into categories that disserve their rich features. Although judges' categorizations yielded a small fraction of misfits, this more likely reflects our ability to do similarity matches than it reflects support for the "representativeness of response categories."

Difficulties With the Method of Using Written Self-Reports for Classifying Errors

Other difficulties stem from a paradox in C&B's results. Most of their (even experienced) subjects have some impetus-like ideas, as shown by true/false answers to their physics test. (Note that C&B's "test" and "problems" refer to different item sets.) Yet impetus *explanations* were generally much more rare, suggesting that the written self-report method is inferior to a fairly nonintrusive structured verbal-protocol session—with respect to eliciting explanations that are relevant to misconceptions—in contrast to C&B's (and Donley & Ashcraft's 1992) suggestions about the impracticality and problems attributed to such interviews (cf. below; Hojnacki, 1988; Ranney, 1987/ 1988, 1994). It is difficult to truly conservatively err toward impetus explanations if the (graphical or written) data lack illustrative richness. For instance, one can undercode impetus explanations (e.g., coding them as

"omissions" or "descriptives") just due to their brevity. In a methodological "compromise," C&B tried to elicit longer explanations in Experiment 2, but (understandably) without success. A related account of the paradox is that some "correct" explanations (or "Newtonian" individuals) actually use terms such as *momentum* as if they were impetus, but this is not picked up via written explanations (and thus undercounted) since one cannot query writings. For instance, I find that subjects may say or write "straight path" when meaning "a fluid curve—no major corners."

A different account suggests that the wording of C&B's test questions might have spuriously inflated the tally of "impetus" ideas, as shown by some of the experienced subjects' remarks; the questions rely on nomenclatural trickiness about impetus, momentum/speed, and other quasi-subtleties. Physicists I have discussed these with note similar wording difficulties—especially those relating to impetus: Question 9 actually seems generally true (not false), given that (certainly angular) momentum, a vector, is *always* directional; Question 7 is also true (again, not false) unless one knows that impetus is a "bad" synonym for momentum; Question 10 is reasonable, but it and some others rely on subtleties of inference peculiar to those familiar with "physics-misconception-speak." People perform better on this test if they think that "impetus" is (1) not momentum and (2) a word to avoid.

A final account of the paradox is that C&B systematically mistally some explanation errors as nonimpetus responses: Their "adds velocity" error example (see their Table 3), which naturally dips in frequency when the pendulum tasks are dropped for Experiment 2, is an untallied kind of impetus response. Similarly, some "surface" errors (e.g., on tube items) may be seen as "impetus" errors with richer explanations. (In contrast, Table 3's "impetus" example is due to a poor problem wording that implies pendular "motion" at point B; see below.) These (sometimes converse) dissociations between explanation accuracy and selection(/pigeonholed production) accuracy further suggest that C&B's data are ambiguous, yielding only tenuous conclusions.

In hindsight, some past work likely used dubious items in trying to assess impetus (e.g., test items from McCloskey, 1983), but C&B may too quickly reject the apparent ubiquity of impetus ideas due to this. They also seem to inappropriately relate such ambiguous true/false questions to the plausible alternative of eliciting oral-interview/verbal-protocol responses (and their own issues; see below). This is not mere methodological preference; it seems untenable to advocate collecting less data due to their potential ambiguity, and this should not pose a problem if C&B desire to be rigorously and conservatively biased toward finding "impetus consistency." Many of us have argued and shown that as subjects explicate their ideas more—especially with converging measures, such as graphical depictions—their responses and beliefs are disambiguated, overcoming default descriptions and conversational maxims (e.g., Gutwill, Frederiksen, & Ranney, in press; Ranney, 1987/1988; Schank & Ranney, 1992).

The Importance of Proper Problems and Alternatives in Assessing Naive Theories

Many have used tasks that turn out to lack some desired or purported characteristics—yet another pitfall for well-controlled naive physics research. Often, "isomorphic" problems (or response typologies) are not truly isomorphic on the dimensions of interest, and accompanying text may not match the physics indicated via diagrams. Some tasks use language that leaves out needed caveats (e.g., about friction, masslessness; cf. Anderson, Tolmie, Howe, Mayes, & Mackenzie, 1992) or that unintentionally biases solutions. But it is most troubling when correct alternatives are inadvertently absent from multiple-choice problems. Such troubles are compounded when the foils provided are muddled or not theoretically well motivated, as the garnered data become even more noisy. Since C&B's article illustrates some of these troubles in each of their four problems, I will mention a few as methodological caveats for future research.

The "correct" alternatives to two of C&B's three (e.g., standard) cliff items are incorrect, as the (50-mph!) ball moves *vertically* before it strikes the ground (after a nonparabolic, rather circular, trajectory; "C,S" in the present Appendix; see C&B's Appendix B). This cannot occur if air resistance is ignored, as instructed. The "correct" responses actually indicate a dissipation—an impetus notion in which lateral velocity is lost for no force-based reason. (Note that some subjects offer incomplete "asymptotic impetus" variants; e.g., Ranney, 1987/1988.) That expert judges can agree on the *best of four incorrect drawings* (or their ordinality) does not allay the fact that these "correct" cliff alternatives actually represent a highly robust misconception.

Furthermore, in line with earlier remarks on variants, the order of C&B's foils varies so much that the "immediate-straight-down" response for the cliff problem's three versions is termed "wrong" (standard), "slightly wrong" (perceptual set), and even "almost correct" (orientation). This order does not follow a theoretical principle, since the same sorts of fellow foils are used in each version. This is a local manifestation of a more global question: The order of these foils was determined empirically; but given a conflict between empirical inconsistency (whether from "experts," pilot subjects, or past studies) and a consistent, plausible, theory's approach, it is often best to use the theory's principles to develop tasks and foils. Again, a main problem in C&B's analyses is that response categories were empirically *but not theoretically* motivated, since only a small subset of produced responses was used for categorization. If one must "pigeonhole," I advocate developing categories that represent multiple, plausible, "theoretical positions"—far more than two or four, and ideally enough to cover virtually the full set of distinct production responses.[2] Finally, some of C&B's

flawed problems have previously been used by other researchers; again, dubious and/or ambiguous problems from past work (e.g., the tube problem's wording vs. its drawing) should not be retained due to history alone.

The pendulum problem's alternatives, and the orientation-problem's wording, are even more theoretically and otherwise muddled. C&B realized some of this, dropping the problem from Experiment 2. Schank and Ranney (1992; see also Ranney et al., 1993)—and likely, C&B's own production data—offer better pendulum-orientation alternatives, though they are hardly representative of the full set of elicited path forms (see the present Appendix). While partially empirically derived, our tasks are more free of the troubles found in C&B's items. For instance, C&B's orientation problem (from their Appendix A) states: "While the ball is in motion, the string is cut at point B." If the ball were in motion, the "correct" answer is wrong, as it suggests that B is the swing's endpoint/apex since its vertical path indicates a zero release velocity (Ranney & Thagard, 1988; also see Ranney, 1987/1988, in press; Schank & Ranney, 1992). So, their problem statement (with "motion") conflicts with their "correct" answer (from instantaneous stasis).

C&B's tube and rocket problems also present difficulties.[3] More critically, we must ask, "What naive/impetus notions should/does a problem assess?" Furthermore, how do its alternatives relate to "impetus" ideas assessed in other problems (e.g., curvilinear impetus, internal force, dissipation; Ranney, 1987)—which should be evident in displays like C&B's Table 3? Also in contrast with the limited diagnosticity of a few selection/categorization foils, abstract zero-gravity problems (like C&B's rocket) can be used to more sensitively assess individuals' evolving understandings of impetus or inertia. For example, one can identify dissipation and internal force ideas via particular responses to such far-transfer tasks (Ranney, 1987/1988, 1988).

In short, assessments of the consistency of subjects' naive theories are limited by the diagnosticity of one's measures and methods. Given difficulties with C&B's (or other scholars') problems, alternatives, test questions, explanation elicitations, and data aggregation, their conclusions regarding such consistency are quite tenuous. The next section elaborates this point with a focus on alternative ways to assess and approach subjects' consistency.

OTHER WAYS TO CONSIDER OR EXHIBIT THE PRESENCE OR ABSENCE OF RESPONSE CONSISTENCY

The Consistency Hypothesis is Not the Null Hypothesis

Implicit in C&B's article (and some others) is the idea that subjects' consistency may be rejected by showing that their responses do not follow some patterns (here, a criterial set of "impetus" responses). We seem in control: the data do not fit a certain "model" (e.g., with $p < .1$), we are tempted to reject "consistency" along with our null (model) hypothesis. But this claim is too global; subjects may be perfectly consistent, yet use unanticipated principles and/or task features. This is the curve-fitting problem (e.g., Harman et al., 1988); any finite data set may be generated or modeled with enough parameters (and subjects wield many "parameters"). C&B can at best claim (i.e., modulo their method's flaws) that subjects did not seem to consistently apply their proposed (aggregated) impetus model. But, due to its infinite breadth, we cannot reject the consistency hypothesis. We can reject the "*inconsistency* hypothesis," though, if the data are low in noise, by finding features or principles that predict subjects' responses beyond random levels. C&B offer no random consistency values, but they suggest that some subjects' data can be accounted for with (1) a Newtonian model (but see below) or (2) an "impetus" model (hence, the $-.48$ cited above).

How Full is the "Glass" of Consistency?

Many arguments over consistency—in any domain—are based on reactions to the contrast between the prior and posterior "gut" expectations of those viewing the data. If a model can account for, say, 50% of (individual) subjects' variance, one might (1) highlight the subjects' surprising consistency, (2) praise the model's predictive power, (3) note the "missing" 50%, or (4) wish that more consistency were observed; Reactions 1 and 3 are optimistic (the "half-full glass"), whereas Reactions 2 and 4 are less optimistic (the "half-empty glass"). Results from Hojnacki and Resnick's work (e.g., Hojnacki, 1988) are instructive: They too assessed consistency in naive motion conceptions, considering many problem situations, features, and (both a priori and a posteriori) dimensions. With a metric of consistency we developed, subjects were found to be only a fraction of the way (1.7 on the metric) from random consistency (about 1.5) to fully consistent (3.0 across Newtonian isomorphs). Hojnacki could thus "optimistically" term them "consistent" (non-random), yet I could claim that their responding was far from consistent (at least regarding the considered models) and thus highly context-specific across tasks. (See the caveat below, though, regarding temporal consistency.)

Model-Centered Versus Individual-Centered Consistency

Consistency can also be thought of either in terms of a researcher-imposed model (from past studies' results, hunches, etc.) or in terms of emergent principles based on individuals' responses (cf. C&B). Colleagues and I have used measures that span much of this continuum (e.g., Ranney, 1987/1988, 1988, 1994; Ranney & Thagard, 1988; Schank & Ranney, 1992; see also Chi, 1992, pp. 161-162), and one comes closer to treating consistency *as if* it were a "rejectable" null hypothesis as one uses individuals' responses to develop consistency metrics. Hence, the following examples (a-j; Ranney, 1987/1988, 1988,

1994) illustrate progressively individual-centered consistency notions (although the results show that subjects' consistency levels were generally *low*): a and b are the most model-centered (e.g., Newtonian) examples; c, d, e, and f are less so; and g, h, i, and j are fairly individually centered. In each case, these fairly naive adults offered drawn trajectories and oral explanations for pendular releases and various other (often isomorphic) tasks, with dropped, thrown, released, pushed, and swung objects: (a) Correlations between Newtonian accuracies among task sets were seldom significant. (b) Subjects correctly transitively used (nontheoretical) feedback only about half of the time. (c) Response consistency over isomorphic pendular and dropping/throwing tasks was only 20%. (d) Consistency among isomorphic swinging tasks' drawn paths was only 19%. (e) Oral descriptions of pendular motion were often inconsistent with near-transfer isomorphs' predictions. (f) Asymmetrical responses were offered for mirror-image pendular tasks 26% of the time. (g) Most subjects predicting a vertical path from a pendular swing's nadir gave inconsistent predictions for a wrecking-ball task. (h) Most subjects who predicted nonvertical apex-release paths also said that pendulums rest at the apex (cf. "Hal" in Ranney & Thagard, 1988). (i) Multidimensionally scaled similarity judgments show that subjects viewed the dropping/throwing tasks as fairly unrelated. (j) Only 31% of the time did individuals draw the same path form (e.g., as some are coded in the present Appendix) for task pairs they indicated to be isomorphic; this highly individual-centered "behavioral agreement" measure even welcomes pairs that are isomorphic from neither Newtonian nor standard-impetus perspectives.

Our lab has also simulated motion beliefs without (essentially) relying on *any* physical model, using what I call the "bifurcation/bootstrapping" method (Ranney et al., 1993; Schank & Ranney, 1992; cf. Ranney & Thagard, 1988) to predict subjects' believability ratings for both their verbalized propositions and a set of alternative trajectories (including their initial prediction). Our "reasoner's workbench" (Ranney, in press) automates this method as it helps subjects explicate their naive physics (and other) arguments.[4] Thus, we link subjects' on-line theorizing with a general belief evaluation model (ECHO; Ranney & Thagard, 1988).

Semistructured Interviews With Verbal Protocols as an Alternative Rigorous Method

Unlike C&B's, almost all of the relative consistency measures in the prior subsection rely on oral protocols from semistructured interviews, which are even critical for accuracy measures; they narrow interpretations of subjects' drawings, yielding (1) very high intercoder reliabilities and (2) good ways to compare trajectory drawings of both true and (e.g., subject-) alleged isomorphs (Ranney, 1987/1988, 1994). Without the richness of protocols, comparing subjects' kinematic and/or dynamic (e.g., speed vs. force) descriptions with their drawings can be imprecise or nondiagnostic. Verbal responses to semistructured interview probes also allow for more precise tallies of subjects' evidenced impetus types (e.g., Ranney, 1987).

Oral protocols are difficult to garner and process. Developing rigorous scoring rubrics and analyses requires creativity, and verbal probes must be constrained, nonsuggestive, and consistent (Ericsson & Simon, 1984/1993; cf. Nisbett & Wilson, 1977). But oral protocols from proper interviews offer converging measures (with drawings, choices, etc.) that are almost irreplaceable. With them, for instance, I found that each of one study's 28 experimental subjects (on a pretest to a 3-h session) evidenced some features of both dissipation and internal force, and that 61% also evidenced curvilinear impetus aspects. (Note that the proportion of subjects' reliance on impetus was $< .5$ overall for each of the three types; e.g., Ranney, 1987, 1987/1988.) This relative ubiquity of (three types of) impetus beliefs contrasts with C&B's interpretation of their explanation data—yet it is more in keeping with their physics test data.

Written explanations take fewer resources and reduce some worries of experimenter-subject bias. But they only roughly approximate oral responses. Subjects quickly tire of repeatedly writing similar rationales and then cut corners in explication. C&B's desire for longer explanations shows this, as does the fact that almost half of their explanations contained omissions or mere descriptions. Vigilant well-trained interviewers obviate this, and they are sensitive to new explanations that subjects might not write. Meanings of terms like "impetus," "momentum," "inertia," "energy," "oomph," "force," "acceleration," "power," "dying out," and "overcoming" are also clearer with the extra context and potential decomposition that such interviews provide—especially regarding motion, with its difficult-to-verbalize perceptual character (see Ranney, 1989, and related articles). Written explanations are often useful, but they are more retrospective (planned) than are oral verbal protocols, and thus more suspect. In sum, without the greater diagnosticity offered by oral responses to properly contingent probes, C&B's claims of low consistency among impetus-using subjects are, again, plausible yet tenuous.

Theoretical Stability, Dissociations, Coherence, and Temporal Consistency in Laypeople

Recall that we might attribute a naive (e.g., impetus) motion theory to a subject if his/her behavior was well predicted by a proposed model, but that much of the above data (including C&B's) indicate, as yet, no such satisfactory model. Furthermore, describing the relative theoryness of naive motion beliefs presents even more difficulties: Context-driven dissociations between perceptuo-kinesthetic-motoric knowledge and verbally displayed knowledge are also disconcerting (e.g., Anderson et al., 1992). Examples can be anecdotal (e.g., some of my subjects were both poor motion-predictors and elite ballplayers), testimonial (e.g., colleagues' stories on overriding one's cognitions based on muscular feedback), or formal (e.g., some of my subjects who best predict a projectile's location used

very non-Newtonian explanations and drawings). It is not clear, then, if a theory rests in one's actions, one's conceptions, or both.

Another danger lies in assuming, as C&B seem to do, that apparent theoretic/conceptual inconsistency means "on-the-fly" theorizing. It might be that subjects are doing very little on-line theorizing; perhaps a proposed model is just mainly orthogonal to the subjects' theories. Were this true, temporal consistency (e.g., for identical tasks over a delay) should be high, even if (from the researchers' view) "theoretical consistency" were low. Hojnacki (1988) showed this: After nearly a month's delay, subjects were (unexpectedly) given the same (and new) qualitative motion problems. In contrast to the low (just above random) consistency from a modeling perspective, subjects were much more temporally consistent (i.e., closer to perfect than to random). So, subjects (who were not simply remembering prior responses, it seems[5]) appeared to use somewhat stable personal "theories" that still defy description (although one can characterize the theories more so, either a posteriori or with many parameters; see Hojnacki, 1988).[6] Is not temporal consistency a hallmark of a theory? These data show that individuals can be affected in the same way by the same (but perhaps not merely "isomorphic") problem/situation characteristics. (Note that diSessa's 1983 and 1988 "p-prims" offer a plausible view of this puzzle, but the approach does not readily lend itself to formal prediction or empirical falsification; but see Anderson et al., 1992, for a suggestive attempt.)

Hojnacki's remarkable finding is supported by the work in which we use subjects' verbal protocols to "blindly" predict individuals' beliefs and predictions (Ranney et al., 1993; Schank & Ranney, 1992). Again, no physical theory is imposed as we encode subjects' arguments in a basic explanatory coherence model (Ranney & Thagard, 1988; Schank & Ranney, 1991; Thagard, 1989). (Note that "coherence" does not imply "no contradictions"; we all live with contradictions among our beliefs, since we do not have the resources to be globally consistent; see Ranney, in press.) In sum, these efforts, and those of Hojnacki and Resnick (above), suggest that the physical "theories" of laypeople may be somewhat systematic and temporally stable, while seeming to be highly idiosyncratic when viewed as a group. Clearly, this picture differs from that offered by C&B (since even their estimates of Newtonian consistency among their subjects may be challenged[7]). But it differs even more from those who claim, as Caramazza et al. (1981) did, that individuals adhere to a small set of "basic models of motion."

CONCLUSIONS

These comments make clear that studying naive physics entails much of what makes cognition research difficult. Methodological pitfalls seem to ring each construct, including the vague and polysemous notions of "theory," "consistent," and "concept" (e.g., Ranney, 1994). Experimental materials (e.g., problems, foils, and correct alternatives) must be excruciatingly precise. Procedures must involve minimal bias. The most sensitive and convergent measures are needed for diagnostic rigor. One's models must (1) guide a study's design, a priori, with explicit principles or (2) come from subjects via sensitive analyses of data that are rich enough to foster such emergence. Although C&B display fine efforts, their experiments seem to fall short on most of these criteria, relative to the useful methods of several studies described above.

C&B's conclusion that naive impetus theory plays a minor role in subjects' performance is suspect, due to their imprecise and effectively monolithic theoretical construal of the rich variety of types of impetus ideas, as well as their dubious measures, categorizations, and analyses. Furthermore, as illustrated with work by myself and others, there is little doubt of C&B's other conclusion that contextual cues are critical in the genesis of some of their garnered responses. Still, fewer of subjects' explanations may be constructed on-line than C&B suggest (since a stronger position comes close to improperly accepting the "null hypothesis of inconsistency"). Given the aforementioned findings of considerable temporal consistency (Hojnacki, 1988) and the relative coherence of subjects' beliefs (Ranney et al., 1993; Ranney & Thagard, 1988; Schank & Ranney, 1991, 1992), it seems more plausible to suggest that, although the layperson's physics may be idiosyncratic, such subjects' judgments and explanations may yet be found to be theory driven—depending on one's meaning of "theory."

REFERENCES

ANDERSON, T., TOLMIE, A., HOWE, C., MAYES, T., & MACKENZIE, M. (1992). Mental models of motion. In Y. Rogers, A. Rutherford, & P. A. Bibby (Eds.), *Models in the mind: Theory, perspective and application* (pp. 57-71). London: Academic Press.

CARAMAZZA, A., MCCLOSKEY, M., & GREEN, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition*, **9**, 117-123.

CHI, M. T. H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. N. Giere (Ed.), *Cognitive models of science* (pp. 129-186). Minneapolis: University of Minnesota Press.

CLEMENT, J. (1983). A conceptual model discussed by Galileo and used intuitively by physics students. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 325-339). Hillsdale, NJ: Erlbaum.

COOKE, N. J., & BREEDIN, S. D. (1994). Constructing naive theories of motion on the fly. *Memory & Cognition*, **22**, 474-493.

DISESSA, A. A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 5-33). Hillsdale, NJ: Erlbaum.

DISESSA, A. A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), *Constructivism in the computer age* (pp. 49-70). Hillsdale, NJ: Erlbaum.

DISESSA, A. A. (1993). Toward an epistemology of physics. *Cognition & Instruction*, **10**, 105-225.

DONLEY, R. D., & ASHCRAFT, M. H. (1992). The methodology of testing naive beliefs in the physics classroom. *Memory & Cognition*, **20**, 381-391.

ERICSSON, K. A., & SIMON, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: MIT Press. (Original work published 1984)

GUTWILL, J., FREDERIKSEN, J. R., & RANNEY, M. (in press). Seeking the causal connection in electricity: Shifts among mechanistic perspectives. *International Journal of Science Education*.

HALLOUN, I. A., & HESTENES, D. (1985). Common sense concepts about motion. *American Journal of Physics*, **53**, 1056-1065.

HARMAN, G., RANNEY, M., SALEM, K., DORING, F., EPSTEIN, J., & JAWORSKA, A. (1988). A theory of simplicity. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 111-117). Hillsdale, NJ: Erlbaum.

HOJNACKI, S. K. (1988). *Consistency in naive physical reasoning*. Unpublished master's thesis, University of Pittsburgh, Learning Research and Development Center.

KAISER, M. K., JONIDES, J., & ALEXANDER, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition*, **14**, 308-312.

MASSON, M. E. J., HILL, W. C., CONNER, J., & GUINDON, R. (1988). Misconceived misconceptions? In E. Soloway, D. Frye, & S. B. Sheppard (Eds.), *Proceedings of CHI'88 Human Factors Conference*. New York: ACM.

MCCLOSKEY, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299-324). Hillsdale, NJ: Erlbaum.

MCCLOSKEY, M., WASHBURN, A., & FELCH, L. (1983). Intuitive physics: The straight-down belief and its interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **9**, 636-649.

MINSTRELL, J., STIMPSON, V., & HUNT, E. (1992, April). *Instructional design and tools to assist teachers in addressing students' conceptions and reasoning*. Paper presented at the meeting of the American Educational Research Association, San Francisco.

NERSESSIAN, N. J., & RESNICK, L. B. (1989). Comparing historical and intuitive explanations of motion: Does "naive physics" have a structure? In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 412-417). Hillsdale, NJ: Erlbaum.

NISBETT, R. E., & WILSON, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, **84**, 231-259.

RANNEY, M. (1987, April). *Restructuring conceptions of motion in physics-naive students*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

RANNEY, M. (1988). Changing naive conceptions of motion. (Doctoral dissertation, University of Pittsburgh, Learning Research and Development Center, 1987.) *Dissertation Abstracts International*, **49**, 1975B.

RANNEY, M. (1988, November). *Contradictions and reorganizations among naive conceptions of ballistics*. Paper presented at the annual meeting of the Psychonomic Society, Chicago.

RANNEY, M. (1989). Internally represented forces may be cognitively penetrable: A comment on Freyd, Pantzer, and Cheng (1988). *Journal of Experimental Psychology: General*, **118**, 399-402.

RANNEY, M. (1994). *Individual-centered vs. model-centered approaches to consistency: A dimension for considering human rationality*. Manuscript submitted for publication.

RANNEY, M. (in press). Explorations in explanatory coherence. In E. Bar-On, B. Eylon, & Z. Schertz (Eds.), *Designing intelligent learning environments: From cognitive analysis to computer implementation*. Norwood, NJ: Ablex.

RANNEY, M., SCHANK, P., MOSMANN, A., & MONTOYA, G. (1993). Dynamic explanatory coherence with competing beliefs: Locally coherent reasoning and a proposed treatment. In T.-W. Chan (Ed.), *Proceedings of ICCE'93 International Conference on Computers in Education: Applications of Intelligent Computer Technologies* (pp. 101-106).

RANNEY, M., & THAGARD, P. (1988). Explanatory coherence and belief revision in naive physics. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 426-432). Hillsdale, NJ: Erlbaum.

RIDGEWAY, D. D. (1992, April). *Knowledge is not always what we take it to be: Issues in the assessment of students' understanding of motion*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 344 892)

SCHANK, P., & RANNEY, M. (1991). The psychological fidelity of ECHO: Modeling an experimental study of explanatory coherence. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 892-897). Hillsdale, NJ: Erlbaum.

SCHANK, P., & RANNEY, M. (1992). Assessing explanatory coherence: A new method for integrating verbal data with models of on-line belief revision. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 599-604). Hillsdale, NJ: Erlbaum.

SCHANK, P., & RANNEY, M. (1993). Can reasoning be taught? *Educator*, **7**, 16-21.

SHANNON, B. (1976). Aristotelianism, Newtonianism, and the physics of the layman. *Perception*, **5**, 241-243.

SMITH, J. P., DISESSA, A. A., & ROSCHELLE, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, **3**, 115-163.

THAGARD, P. (1989). Explanatory coherence. *Behavioral & Brain Sciences*, **12**, 435-502.

NOTES

1. "Overcoming" can also be used in a fairly Newtonian way. For example, subjects offer explanations such as (1) "the force of gravity overcomes the *fully* horizontal initial velocity of a ball rolling off a cliff," and (2) "gravity's acceleration overcomes the upward velocity component of a pendulum during an upswing."

2. I agree with C&B that the motion condition in Experiment 1 was flawed due to the anomalous dynamics. This usually occurs when subjects' data are aggregated, as C&B have done, since the foils' paths have inherently ambiguous dynamics that are based on the subjects' varied models (cf. earlier remarks on variants). For example, some may even describe a fully vertical path as either accelerative or not (e.g., Shannon, 1976). But with sensitivity to subjects' explanations, one can diversify the paths' dynamics and assess the display type factor again.

3. For the perceptual-set tube problem, for example, subjects were told to ignore gravity while the tube lies flat—an impossibility, since the tube's "start" must be elevated to be realizable and consistent with the text (e.g., with the ball acceleratively "shot out" while merely "put" in one end.) Also, in C&B's appendices, the standard problem's rocket seems initially more displaced than the orientation problem's, appearing to have an implied (e.g., diagonal/curvilinear) trajectory.

4. This computer program is called "Convince Me" (Ranney et al., 1993; Schank & Ranney, 1993). Compare recent related work by Jim Minstrell, Earl Hunt, and their colleagues (e.g., Minstrell, Stimpson, & Hunt, 1992).

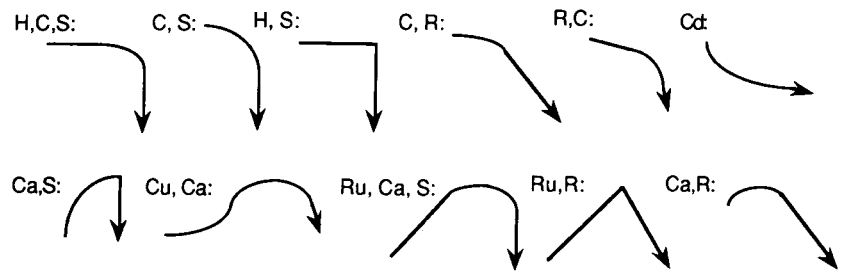
5. The old tasks followed the new tasks, and their ordering was reversed, providing memory interference (Hojnacki, 1988). Also, subjects even forget paths predicted only minutes earlier, even with external memory aids (as in "f" above; Ranney, 1987/1988, 1994).

6. Similarly, one measure of relative inconsistency and/or incoherence ("e" above) seemed fairly low (only 29% for some aspects of near-transfer isomorphs; e.g., see Ranney, 1988, 1994).

7. With their method, C&B's criteria for Newtonian consistency may be too generous (but cf. above and Hojnacki, 1988), and a random consistency measure is needed. Here, we might be less concerned about whether, for example, the experienced are fairly Newtonian than whether the *inexperienced* have consistent sorts of impetus theories. (Hence, it would be interesting to know the proportion of C&B's "Newtonians" that were from their experienced group.)

(Continued on next page)

APPENDIX
A Sample of the Many Incorrect Trajectory-Codings from Ranney (1987/1988; Experiment 2)



Note—This includes some trajectory-aspect codes—such as H (horizontal), C (curvilinear), S (straight-down), R (rectilinear/diagonal), Cd (concave-down), Ca (curvilinear-arch), Cu (concave-up), and Ru (rectilinear/diagonal-up)—for some lateral, downward, and upward releases.

(Manuscript received June 15, 1993;
revision accepted for publication August 30, 1993.)