IOP Publishing European Journal of Physics
Eur. J. Phys. **39** (2018) 015705 (12pp) https://doi.org/10.1088/1361-6404/aa940f

Evidence on the coherence–pieces debate from the force concept inventory

D Badagnani[1,2,3], **D Petrucci**[3,4] **and O Cappannini**[3,5]

[1] Departamento de Física, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Argentina
[2] UIDET CeTAD, Facultad de Ingeniería, Universidad Nacional de La Plata, Argentina
[3] Grupo de Didáctica de las Ciencias, IFLYSIB, Argentina
[4] Departamento de Ciencias Básicas, Facultad de Ingeniería, Universidad Nacional de La Plata, Argentina
[5] Instituto de Física de Líquidos y Sistemas Biológicos (IFLYSIB), CONICET-UNLP, Argentina

E-mail: daniel@fisica.unlp.edu.ar

Received 12 June 2017, revised 14 September 2017
Accepted for publication 17 October 2017
Published 18 December 2017

Abstract

We use force concept inventory (FCI) data to probe the consistency of commonsense physics as a knowledge system. The source of this data is the administration of the FCI to first-year science university students. Data quality was checked using item response theory and studying answer distributions for each question. We find apparently paradoxical results: depending on how the data is analysed, answers seem highly systematic or almost random-like. These results are compatible with others found in the literature and can be construed as arising either from a coherent knowledge system or from knowledge in pieces. We hypothesise as a possible source of this apparent contradiction that predictions and explanations use different resources: the former would use reflex, low-cost cognitive resources while the latter would involve conceptualisations. We show that the articulation of both resources may be crucial for expert thinking productivity (the ability to apply a theory to novel situations). We sketch some consequences of the proposed structure of commonsense thinking for teaching and further research.

Keywords: knowledge system, predictions in commonsense physics, explanations in commonsense physics

Introduction

It has been acknowledged for decades now that, prior to formal instruction, human beings already have a vast commonsense knowledge system on movement and interactions which is used to make predictions and explain phenomena, usually in conflict with Newtonian mechanics. Nonetheless, the structure of this system and its evolution as a learner acquires a Newtonian perspective remains controversial. Understanding this structure is crucial for drawing pertinent recommendations for teaching. The force concept inventory (FCI) (Hestenes *et al* 1992) allowed an impressive accumulation of evidence about the superiority of interactive engagement courses over traditional ones (Hake 1998). But we still do not have a deep understanding of the reasons for this superiority, which surely are to be found in the understanding of commonsense knowledge and its interplay with formal instruction mentioned above. There are two dominant views about commonsense mechanics: a 'coherentist' perspective defended for instance in Ionnides and Vosniadou (2002), and a 'pieces' perspective exposed for instance in diSessa (2004). For an overview of its instructional consequences see, for instance, Özdemir and Clark (2007).

There is little doubt about the reliability and validity of the FCI: certified Newtonian thinkers find it trivial (Hestenes *et al* 1992), FCI and mechanics baseline scores correlate well (Hestenes and Wells 1992), there is a good correlation with the force and motion conceptual evaluation (Thornton and Sokoloff 1998), (Thornton *et al* 2009), it has been validated with extensive interviews and by comparison with the former mechanics diagnostic test (Halloun and Hestenes 1985, Hestenes *et al* 1992), it shows global retest stability (Lasry *et al* 2011a) and low global context dependence (Stewart *et al* 2007), it has been shown to measure a single construct (Wang and Bao 2010), and students scoring 85% or higher exhibit a fair degree of coherence in their conception of force (Halloun and Hestenes 2009). Certainly, explorations with the FCI have led to the discovery of striking regularities for courses following similar pedagogy (Hake 1998).

Nevertheless, the analysis of FCI data keeps producing puzzling results: in spite of the fact that interviews confirm that students interpret its responses in terms of a few 'misconceptions', a factor analysis leads to almost no factorisation (Heller and Huffman 1995). There is also a significant switch among distractors between test and retest (Lasry *et al* 2011a) and individual questions are affected by context dependence especially in beginners (Bao and Redish 2001). In this work we are presenting new evidence in this direction: the pattern of answers by a non-Newtonian population is random-like from the point of view of most of the 'misconceptions' from the taxonomy in which the FCI is based, and none of them is used with any degree of consistency. All this sheds light on the structure of commonsense physics. There are some precedents in this line of reasoning: Huffman and Heller (1995) attributes the lack of factorisation to a possible piece-like structure of commonsense physics, citing works by Minstrell (1991) and diSessa (1993), Lasry *et al* (2011a) claim that their results support the idea of resource activation (Redish 2004, Hammer *et al* 2005, Sabella and Redish 2007). Liu and MacIsaac (2005) used data from FCI to compare the use of mental models and knowledge in pieces in questions involving impetus distractors.

As a way to make sense of these puzzling results we have proposed (Badagnani *et al* 2012) that commonsense predictions and explanations use separate cognitive resources: while predictions would occur through reflex, low-cost operations with minimal intervention of conceptual structures, and thus piece-like, explanations would be longer processes of rationalisation than the reflex-like predictions and would imply theory-like structures as in coherentist accounts of commonsense physics. This is not a compromise between the two points of view but a completely different view of the internal workings of commonsense

physics, which poses interesting questions about learning and the structuring of expert knowledge. This view is compatible with the observation that response times in FCI post tests are larger than in pre test: those longer times could be evidencing the subjects' efforts of using concepts acquired during instruction instead of low-cost commonsense predictions (Lasry *et al* 2013). A similar interpretation can be found in Wood *et al* (2016), where they use the theory of Kahneman (2011) that postulates the existence of two separated cognitive systems with characteristics broadly similar to our 2012 proposal. Wood *et al* (2016) find evidence of the usage of the so-called 'system 1' (low-cost, fast cognitive system in the theory of Kahneman) at solving the FCI by correlating it with an instrument designed to test the usage of each system in problem solving.

In this work we will review the evidence presented in our previous Spanish-language publication (Badagnani *et al* 2012) which led us to postulate a dual 'pieces-for-predictions, coherent-for-explanations' commonsense knowledge system, expose in some detail how we conceive the workings of such a system and its relation to expert knowledge, and sketch ways to further probe our hypothesis with experiments and interviews. Finally, we discuss briefly the implications for teaching.

Materials and methods

A reduced version of the FCI (Spanish version, with questions 8–11 omitted) was administered to 352 first-year university students from Exact Sciences Faculty (Universidad Nacional de La Plata, Argentina) prior to any physics teaching at university as part of an institutional evaluation. It was anonymous and students were not compelled to answer all the questions (we will call this sample CIBEX). We filled 1000 tests at random on a computer to be used as control (we will call this sample RANDOM).

We analysed the distribution of answers from the perspective of Newtonian thinking and according to each of the 'misconceptions' in the taxonomy in Hestenes *et al* (1992), the rationale being that differences between distributions in CIBEX and RANDOM would signal a systematic tendency to answer according (or against) the selected idea or group of ideas.

Results

For CIBEX the average score was 24.8% (as a reference, the score for random answering is 20%). Scores distribution for CIBEX and RANDOM are very similar, confirming that our population is highly non-Newtonian (see figure 1).

Only two tests had to be discarded due to obvious unengaged performance (in one, all answers were A, and in the other they showed a repetitive pattern). For the remaining 350, as a consistency check, we compared the Newtonian responses for each question with the prediction by the item response theory (IRT, see details in appendix B) metric for FCI from Wang and Bao (2010). The correspondence (shown in figure 2 right) is quite good. So, in spite of language, institutional and contextual differences, CIBEX response pattern is similar to that obtained from English speaking, low-Newtonian college populations, and we can assume that results from (Halloun and Hestenes 1985, Hake 1998, Heller and Huffman 1995, Huffman and Heller 1995, Wang and Bao 2010, Lasry *et al* 2011a) and others apply to our sample, and our results apply to theirs as well. In particular, figure 2 shows that the distribution of answers is far from random: if they were random a given option should receive $(20 \pm 2.5)\%$ answers with probability 99.7% for a sample of 350 respondents, as can be seen from applying the central limit theorem to the binomial distribution with $p = 0.2$.

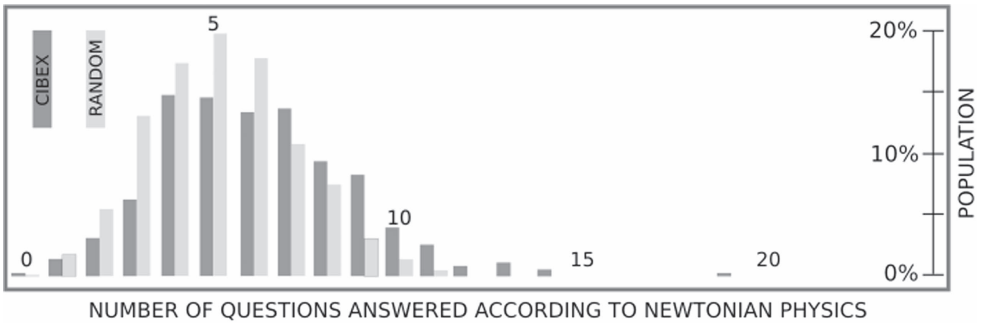

Figure 1. Distribution of Newtonian answers for our sample (CIBEX) and a random sample (RANDOM).

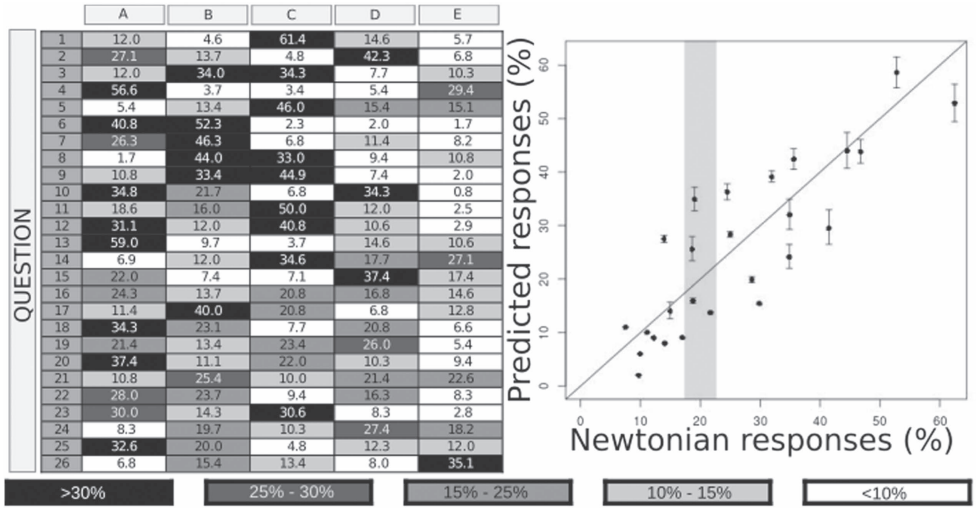

Figure 2. *Left*: Distribution (%) of answers for each question. If answered at random, for a sample of 350 respondents each question would receive $(20 \pm 2.5)\%$ of the answers with a 99.7% probability. *Right*: Percentage of Newtonian answers for each question versus prediction from an IRT analysis. The grey band shows the region where each Newtonian answer rate value would have taken, with a 99.7% probability, if answered at random.

The number of questions left unanswered (see figure 3 left) is negligible for questions 1–18 (8–11 omitted), about 10% for questions 19–25 and rose up to 20% for the last five questions. This is possibly due to tiredness. To check that this is not attributable to question difficulty we show in figure 3 (right) the 'guess' parameter versus proportion of unanswered questions. Both parameters are clearly uncorrelated (see details in appendix B).

In figure 4 we show histograms of distribution of answers compatible with each of the 'misconceptions' from the taxonomy introduced in Hestenes *et al* (1992) (see appendix A). The aim is to observe to what extent those ideas are used with some consistency. If consistency was high we should expect part of the population (those subjects holding that belief) to answer all questions compatibly with the idea, while there would be no compatible answers for the rest of the population. We did not expect that level of consistency, because it is known

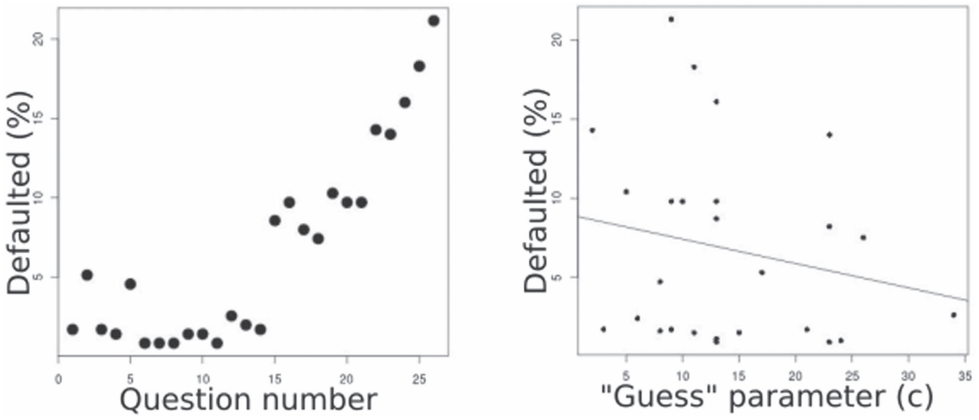

Figure 3. *Left*: Percentage of questions left unanswered as a function of question number (in the order in which they were answered). *Right*: Percentage of questions left unanswered versus 'guess' parameter from IRT analysis. Both graphs suggest that lack of answer is not related to the question contents but to tiredness.

that incompatible ideas seem to be activated depending on the context, but since competing ideas are few we still expected some level of consistency. What we found instead was a distribution hardly different from RANDOM. If the idea is essentially absent, as in I4 where almost no respondent answers any question compatibly the difference with RANDOM is noticeable. But in cases where the idea seems to be present (almost all respondents answer at least one question compatibly), as in I5 there is a difference with RANDOM but it is never spectacular. Observe that we are not trying here to check whether the distributions are compatible or not with random answering, since we already know that answering is far from random. What we find remarcable is that, when answer decisions are analysed from the perspective of the ideas involved rather than from individual questions, the information seems to blur, showing that the misconceptions are a poor organiser of answering decisions. This is in sharp contrast with the extent to which these ideas are organisers of answer justifications, which form part of the extensive validation studios of the FCI.

Discussion

At first sight, results like those shown in figures 1 and 4 could lead one to believe that students are just guessing their answers. We believe there is hardly any guessing, as can be seen from the fact that students answer virtually all questions (at least until they become tired for the last few), the pattern of item responses is far from random, students can always give reasons for their choices in interviews (Halloun and Hestenes 1985, Thornton and Sokoloff 1998) and there is no significative correlation between the 'guess parameter' from IRT fit and test-retest change in answers (Lasry *et al* 2011b). We have not found correlation between the same parameter and the number of questions left unanswered (see figure 3 left and appendix B). Then, the 'guess parameters' are indeed guesses only from the perspective of Newtonian theory.

Once commonsense answers to the FCI are established not to come just at random, the apparent inconsistency is puzzling. The very concept of 'misconception' as a belief brings the implicit idea that it is normative, but then the FCI would not measure a single construct for

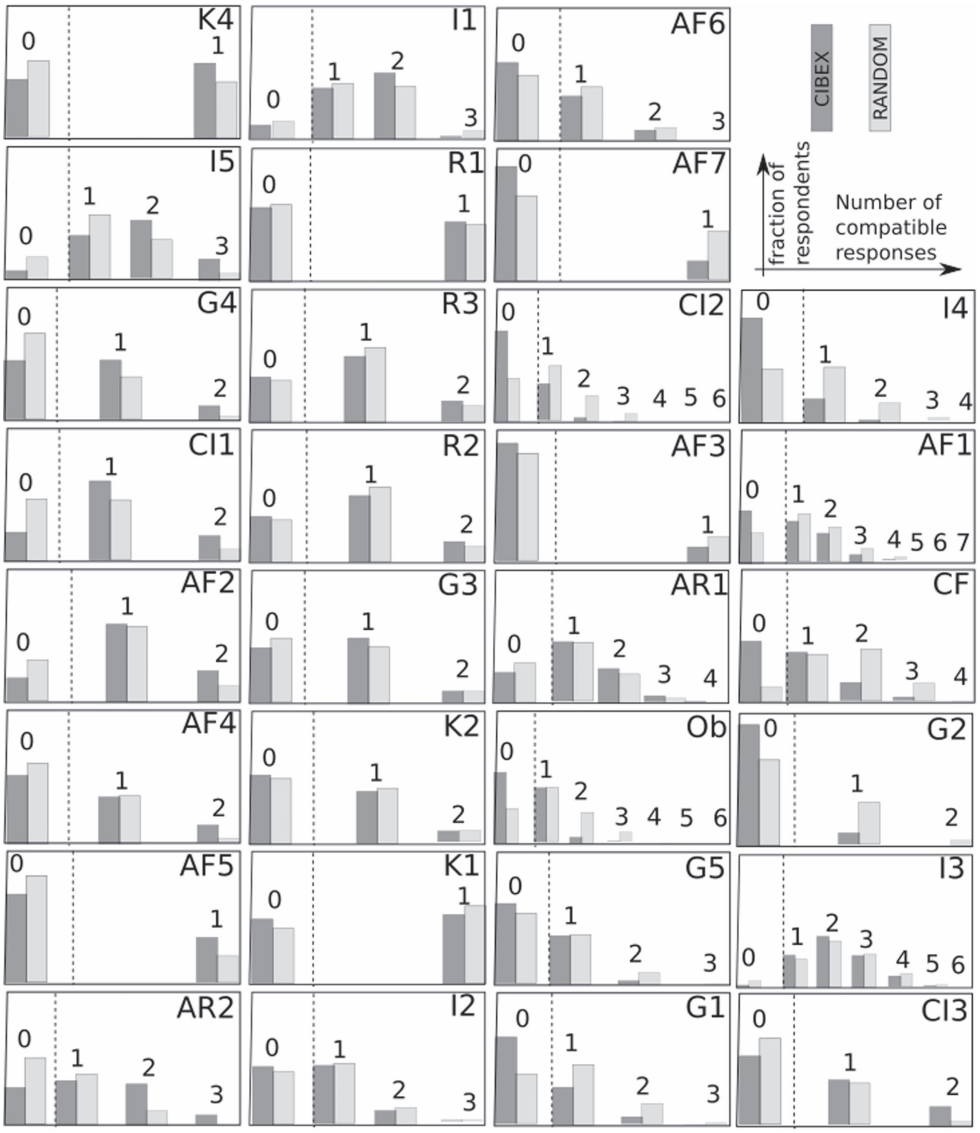

Figure 4. Answer distributions from the perspective of each of the 'misconceptions' used as distractors in the construction of the FCI, both for our sample (CIBEX) and the random sample (RANDOM). The horizontal axis is the number of questions answered compatibly with the misconception.

scores below 85%, which is the proposed threshold for Newtonian thinking (Halloun and Hestenes 2009). Even for quite low score populations one-dimensional IRT models work quite well (figure 2 right shows that this is so even for extremely non-Newtonian populations like CIBEX). The hypothesis of resources activation accounts for the success of one-dimensional IRT models for low proficiency (Wang and Bao 2010), the high context sensitivity (Bao and Redish 2001), test-retest answer switching (Lasry *et al* 2011a), problem classification by 'novices' according to surface characteristics of the problems instead of by

solving strategies (Chi *et al* 1981), and the apparent random use of 'misconceptions' shown in this work.

But, on the other hand, the systematic use of arguments compatible with the 'misconception' taxonomy from Hestenes *et al* (1992) in interviews, and researches such as that of Ionnides and Vosniadou (2002) revealing high levels of conceptualisations since early childhood, makes it difficult to ignore the hypothesis of existence of theory-like structures in the commonsense physics knowledge. In fact, it seems impossible to explain the occurrence of such a bounded taxonomy as the one presented in Hestenes *et al* (1992) only from the hypothesis of resources activation. There is empirical evidence of the possible coexistence of both knowledge systems: Anderson *et al* (1992) found considerable coherence in problem solving involving explanations and low coherence when problems involved predictions. Liu and MacIsaac (2005) found evidence in this sense from the analysis of FCI data. Results in Wood *et al* (2016) point in the same direction.

Our hypothesis that commonsense predictions are essentially reflex is compatible with the model of contextual resources activation. On the other hand, it requires a sophisticated conceptual apparatus to make sense of the huge and disparate mass of reflex responses and express them in words. As a matter of fact, the analysis of mutual forces between a small, fast moving car and a large, static truck is not a phenomenon (the forces cannot be observed), and it is impossible to make sense of the question of which is larger without a conceptualisation of the term 'force'. Our point is then that there are complex conceptual theory-like structures in commonsense physics, but such structures are not normative, and are not productive for predictions.

So, how could we characterise learning of a scientific theory? Observe that coherentist approaches tend to take for granted that 'thinking' is 'reasoning', that is, operating at the level of the meanings of propositions. On the other hand, the 'pieces' approaches (which originated from efforts in the field of artificial intelligence) tend to think of 'thinking' as a sort of hierarchical sets of reflex-like responses, where the hierarchies are reorganised by teaching. We are postulating here that even for the most conspicuously commonsense thinkers there is an interplay between a pre-reflexive structure (that might consist on the pieces postulated in diSessa (1993) or any other reflex-like system like Kahneman's 'system 1') and a complex conceptual apparatus. Observe that anyone able to comunicate orally about movement and interactions is already 'instructed', in the sense that they have aquired conceptual categories from their culture, and only through those categories questions can be interpreted and answers can be produced. So, the interplay between culturally aquired concepts and private, reflex-like representations during instruction should be quite intrincate, and we should wonder how, in experts, reflexes and concepts get integrated in a single system. What is the role of reflex-like proceses in expert thinking? Are they simply suppressed or are they somehow important? An expert thinker is convinced that he thinks in a formal logical way, but since reflex processes are low cost, fast and mostly subconscious processes, that conviction is far from making us rule out an important, or even key, role for these processes.

We can give a strong argument in favour of a key role of reflex-like processes in expert thinking, which lies at the root of the very concept of 'applying' a theory, which is that the theory can be applied to potentially infinite systems and situations, including scenarios unconceivable at the time it was proposed. Such 'applicability' can be thougth of as the colloquial expression of its normativity, and implies virtually infinite productivity (in the same sense that in linguistics). How does an expert 'apply' a theory to a specific situation? Observe that a formal theory poses a few abstract concepts, so there is a process of interpretation and modelling that takes commonsense descriptions of systems to a set of variables amenable to a theoretic treatment. There is no systematic procedure for doing so. Typically,

teaching involves a few 'examples of applications', and the students are expected to produce for themselves other applications. Their ability for doing so is typically viewed by educators as a discriminator between rote and meaningful learning. What is going on in the expert's mind during this non-systematic process of aplying a theory to novel contexts? If the process is mainly unconscious, such processes are strong candidates for reflex-like processes playing a role in expert thinking. If this is the case, experts should exert a sort of control on such reflexes.

Since fast reflexes are mainly unconscious, in order to see them manifested thinking aloud interviews are not enough. The hallmark of such processes is short duration. In order to study them we designed experiment-like interviews where fast responses are singled out. Those interviews, which support our hypothesis of a dual resource structure in novice thinkers and gives further clues on the underlying processes, will be shown elsewhere (they are briefly exposed in our Spanish-language publication Badagnani *et al* 2012).

Understanding this dual structure and its evolution from novices to experts is crucial to make pertinent teaching recommendations based on this kind of research. It seems immediate that neither mere cognitive conflicts nor mere training can work for most learners. The superiority of interactive engagement courses evidenced in studies like Hake (1998) suggests that social interactions play a key role in expert learning, maybe supplying each student ways of controlling or directing their reflex-like processes, but working out the details of how this happens (if it does) remains a challenge.

Conclusions

We have proposed as a possible explanation for the apparently paradoxical results from the FCI that there are two different knowledge structures operating in commonsense physics: a pre-reflexive fast, low-cost, unconscious and context-dependent structure for predictions and a 'coherent' slow, concept-based, conscious structure for explanations and rationalisations of predictions. In order to explain the paradoxes the second structure must not be normative, that is, explanations should not influence or guide predictions. We point out that if this is the case, commonsense physics is not simply just a compromise between pre-reflexive and coherent knowledge but something quite different in which the two systems work simultaneously (it is not either one or the other that is activated, but rather both are necessary to produce commonsense answers). This poses the question of what happens with both structures as novices become experts, and we argue that is is a reorganisation of both, where conceptual thinking becomes normative through some sort of feedback on the pre-reflexive processes. If this is the case, learning does not occur by mere training nor by mere resolution of cognitive conflicts. Understanding this is the key for drawing conclusions for teaching.

In order to probe this proposed knowledge structure we have designed experiment-like interviews in which we isolated fast time processes from the standard thinking aloud protocols. We have described briefly the procedure and some results in Badagnani *et al* (2012) and provide supporting evidence for our hypothesis: predictions coming from fast imagery, not challenged by their concept-based explanations.

Acknowledgments

We thank Monica Manceñido for style correction of the manuscript. This work was partially supported by grant PIP 03032 from CONICET (Argentina), 'Programa de Apoyo a las Propuestas de Mejoramiento de la Enseñanza', FCE-UNLP, and the research projects

'Perspectivas históricas y del presente en las representaciones de la materia y sus interacciones en estudiantes y cursos de la UNLP' and 'Sistematización de innovaciones didácticas y representaciones, actuales e históricas, en la Facultad de Ciencias Exactas' from Programa de Incentivos, Universidad Nacional de La Plata.

Appendix A. The 'misconceptions taxonomy' of Hestenes

The FCI questionnaire had some changes: three questions were removed, four added, and their order was very much altered since the 1992 version. The page www.modeling.asu.edu, where the FCI is officially distributed, mantains a table with the set of 'misconceptions' which the FCI is intended to probe, together with a suggested table of answers compatible with each. This is open to some interpretation, and the page even encourages users to send proposals for improvement. The table added even a 'misconception' absent in Hestenes *et al* (1992): K4 (ego-centred reference frame). We here list the 'misconceptions' in the present state of the table and list, for each 'misconception', the list of compatible answers that we used for constructing the histograms of figure 4 in our reduced 26 items version. Observe that the question numbers in the list below correspond to that in the complete current Spanish version. To make them to correspond to question numbers used in this paper, you should substract 4 to numbers from and above 12.

0. Kinematics	
K1. Position-velocity undiscriminated	19 BCD
K2. Velocity-acceleration undiscriminated	19 A; 20 BC
K3. Nonvectorial velocity composition	—
K4. Ego-centred reference frame	14 AB
1. Impetus	
I1. Impetus supplied by 'hit'	5 CDE; 27 D; 30 BDE
I2. Loss/recovery of original impetus	7 D; 21 A; 23 AD
I3. Impetus dissipation	12 CDE; 13 ABC; 14 E; 23 D; 24 CE; 27 B
I4. Gradual/delayed impetus build-up	21 D; 23 E; 26 C; 27 E
I5. Circular impetus	5 CDE; 6 A; 18 CD
2. Active forces	
AF1. Only active agents exert forces	15 D; 16 D; 17 E; 18 A; 28 B; 29 A; 30 A
AF2. Motion implies active force	5 CDE; 27 A
AF3. No motion implies no force	29 E
AF4. Velocity proportional to applied force	22 A; 26 A
AF5. Acceleration implies increasing force	3 B
AF6. Force causes acceleration to terminal velocity	3 A; 22 D; 26 D
AF7. Active force wears out	22 CE
3. Action/reaction pairs	
AR1. Greater mass implies greater force	4 AD; 15 B; 16 B; 28 D
AR2. Most active agent produces greatest force	15 C; 16 C; 28 D
4. Concatenation of influences	
CI1. Largest force determines motion	17 AD; 25 E
CI2. Force compromise determines motion	6 D; 7 C; 12 A; 14 C; 21 C
CI3. Last force to act determines motion	21 B; 23 C

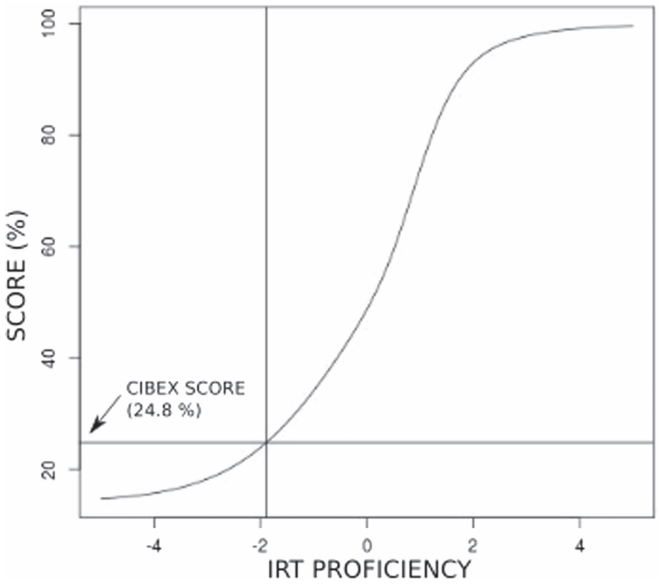

Figure B1. FCI score (% of Newtonian responses) as a function of proficiency as predicted by IRT (Wang and Bao 2010) for our 26 questions version.

(Continued.)

5. Other influences on motion	
CF Centrifugal force	5 E; 6 CDE; 7 CDE; 18 E
Ob. Obstacles exert no force	4 C; 5 A; 15 E; 16 E; 18 A; 29 A
Resistance	
R1. Mass makes things stop	27 AB
R2. Motion when force overcomes resistance	25 ABD; 26 B
R3. Resistance opposes force/impetus	26 B
Gravity	
G1. Air pressure-assisted gravity	3 E; 17 D; 29 CD
G2. Gravity intrinsic to mass	3 D; 13 E
G3. Heavier objects fall faster	1 A; 2 BD
G4. Gravity increases as objects fall	3 B; 13 B
G5. Gravity acts after impetus wears down	12 D; 13 B; 14 E

Appendix B. IRT analysis of our data

In order to estimate the proficiency parameter for our sample we calculated the predicted score (as percentage of Newtonian responses) for our 26 item test as a function of proficiency. This is simply the average of the IRT probability for all 26 questions at the given proficiency. We show this function in figure B1. Observe that, for extreme low proficiency, linearity must be badly broken: in our sample the average of Newtonian responses is about 6.5 out of 26, while for proficiency minus infinity that figure is given by the average of 'guess' parameters

and leads to about 3.6 Newtonian answers out of 26. From the function shown in figure B1, our sample should correspond to a proficiency of −1.85. The plot shown in figure 2(A) was produced using that proficiency, while the error bar corresponds to values between proficiencies of −2 and −1.75, wich correspond to scores between 24% and 26% respectively. In the horizontal axis of figure 2 (right) we represent the observed Newtonian responses for each question renormalized not considering the defaulted questions, which is not a possibility contemplated in the IRT parametrization.

The figure 3 (right) shows the relation of the 'guess' parameter c of IRT with the questions left unanswered. Since the subjects were not asked to answer all questions, some of them may decide not to answer instead of guessing. So, if c is indeed strictly a 'guess level' we should observe both quantities positively correlated. Instead of that, both sets of data are slightly negatively correlated. The correlation is about −0.2, and since there exist a strong correlation between defaulted questions and question order number (figure 3 left), this mild correlation could be understood as due to the purely casual correlation between the 'guess' parameter c and question number, which is about −0.15. The clear correlation between default and question number, which grows quite abruptly at the end, might be construed as due to tiredness (most subjects employed far more than half an hour in completing the questionnaire, and showed quite a compromise). So we think of the label 'guess parameter' as a nickname: our subjects nearly did not guess and were convinced of most of their responses.

ORCID iDs

D Badagnani https://orcid.org/0000-0001-9766-4222

References

Anderson T, Tolmie A, Howe C, Mayes T and MacKenzie M 1992 Mental models of motion *Models in Mind: Theory, Perspective, and Application* ed Y Rogers *et al* (London: Academic) pp 57–71

Badagnani D, Petrucci D and Cappannini O 2012 Sobre los recursos cognitivos en pensadores no-Newtonianos, Actas del SIEF XI http://hdl.handle.net/10915/

Bao L and Redish E 2001 Model Analysis: Assessing the Dynamics of Student Learning http://arxiv.org/pdf/physics/0207069.pdf (Accessed 17 September 2014)

Chi M T H, Feltovich P J and Glaser R 1981 Categorization and representation of physics problems by experts and novices *Cogn. Sci.* **5** 121–52

diSessa A 1993 Toward an epistemology of physics *Cogn. Instr.* **10** 105–225

diSessa A 2004 Coherence versus fragmentation in the development of the concept of force *Cogn. Sci.* **28** 843–900

Hake R 1998 Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses *Am. J. Phys.* **66** 64–73

Halloun I and Hestenes D 2009 The search for conceptual coherence in FCI data http://modeling.asu.edu/R&E/CoherFCI.pdf (Accessed 17 September 2014)

Halloun I A and Hestenes D 1985 The initial knowledge state of college physics students *Am. J. Phys.* **53** 1043–8

Hammer D, Elby A, Scherr R E and Redish E F 2005 Resources, framing and transfer *Transfer of Learning from a Modern Multidisciplinary Perspective* ed J Mestre (Greengwich, CT: Information Age Publishing) pp 89–120

Heller P and Huffman D 1995 Interpreting the force concept inventory, a reply to Hestenes and Halloun *Phys. Teach.* **33** 503–11

Hestenes D and Wells M 1992 A mechanics baseline test *Phys. Teach.* **30** 159–66

Hestenes D, Wells M and Swackhamer G 1992 Force concept inventory *Phys. Teach.* **30** 141–58

Huffman D and Heller P 1995 What does the force concept inventory actually measure? *Phys. Teach.* **33** 138–43

Ionnides C and Vosniadou S 2002 The changing meaning of force *Cogn. Sci. Q.* **2** 5–62
Kahneman D 2011 *Thinking, Fast and Slow* (London: Macmillan)
Lasry N, Rosenfield S, Dedic H, Dahan A and Reshef O 2011a The puzzling reliability of the force concept inventory *Am. J. Phys.* **79** 909–12
Lasry N, Rosenfield S, Dedic H, Dahan A and Reshef O 2011b Reply to 'Comment on 'the puzzling reliability of the force concept inventory,' by N Lasry, S Rosenfield, H Dedic, A Dahan, and O Reshef [Am. J. Phys. 79, 909–912 (2011)] *Am. J. Phys.* **80** 170–3 http://aapt.scitation.org/doi/abs/10.1119/1.3660664?journalCode=ajp
Lasry N, Watkins J, Mazur E and Ibrahim A 2013 Response times to conceptual questions *Am. J. Phys.* **81** 703–6
Liu X and MacIsaac D 2005 An investigation of factors affecting the degree of naive impetus theory application *J. Sci. Educ. Technol.* **14** 101–16
Minstrell J 1991 Facets of student knowledge and relevant instruction *Research in Physics Learning: Theoretical Issues and Empirical Studies* ed R Duit *et al* (Germany: University of Bremen) pp 110–28
Özdemir G and Clark D 2007 An overview of conceptual change theories *Eurasia J. Math. Sci. Technol. Educ.* **3** 351–61
Redish E F 2004 A theoretical framework for physics education research: modeling student thinking *Proc. 2004 Enrico Fermi Summer School, Course CLVI* ed E Redish (Italian Physical Society) pp 1–63
Sabella M S and Redish E F 2007 Knowledge activation and organization in physics problem-solving *Am. J. Phys.* **75** 1017–29
Stewart J, Griffin H and Stewart G 2007 Context sensitivity in the force concept inventory *Phys. Rev. Spec. Top.: Phys. Educ.* **3** 010102
Thornton R K, Kuhl D, Cummings K and Marx J 2009 Comparing the force and motion conceptual evaluation and the force concept inventory *Phys. Rev. Spec. Top.: Phys. Educ. Res.* **5** 010105
Thornton R K and Sokoloff D 1998 Assesing student learning of Newton's laws: the force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula *Am. J. Phys.* **66** 338–51
Wang L and Bao L 2010 Analyzing force concept inventory with item response theory *Am. J. Phys.* **78** 1064–70
Wood A K, Galloway R K and Hardy J 2016 Can dual processing theory explain physics students' performance on the force concept inventory? *Phys. Rev. Phys. Educ. Res.* **12** 023101