# How many factors underly cognitive mechanics?

**Wei Li**
Department of Psychology, Boston College

**Samantha Gutierrez**
Department of Communication Sciences and Disorders, MGH Institute of Health Professions

**Joshua K. Hartshorne (joshua.hartshorne@hey.com)**
Department of Communication Sciences and Disorders, MGH Institute of Health Professions

**Abstract**

How people reason about the mechanics of the physical world is an important question for several different cognitive sciences. Education, cognitive psychology, and developmental psychology have each conducted large numbers of studies over the last several decades, largely in isolation from one another (especially in the last quarter century). The results have suggested that cognitive mechanics may be subserved by a number of mechanisms that are differentially involved in different tasks. Here, we report converging results from factor analysis of a large compendium of mechanics questions.

**Keywords:** intuitive physics; cognitive mechanics; concept inventory; factor analysis

## Introduction

Characterizing what humans know about the mechanics of the physical world is a focal area of study in several cognitive sciences, including cognitive psychology, which endeavors to understand how people navigate, predict, and manipulate the physical world around them; human development, which investigates how babies learn to do the same; and education, which focuses on how to convert untutored beliefs about mechanics into explicit knowledge of the veridical theories. Within cognitive psychology, this research often goes under the heading "intuitive physics", but here we use the term "cognitive mechanics" in order to a) encompass both untutored "intuitive' ' beliefs as well as the beliefs acquired in the classroom, and b) make clear that our focus is on mechanics, not electromagnetism, relativity, or other branches of physics.

All three literatures trace their origins to same seminal studies (Karmiloff-Smith & Inhelder, 1974; McCloskey, 1983; Siegler, 1976). However, over the last quarter-century, they have arrived a very different conclusions based on strikingly divergent data (for review see Hartshorne & Jing (in press)). As a result, the theoretical debates in one literature frequently make no sense in the context of the others. Education studies have documented persistent misconceptions about mechanics among both everyday people and trained physicists, with theoretical debates focusing on explaining why even extensive intervention (e.g., a semester or two of college physics) moves the needle so little (Brown & Hammer, 2009; Resbiantoro, Setiani, et al., 2022; Vosniadou, 2019). If accuracies on cognitive mechanics assessments hover near chance in the education literature, they often — though not always — near ceiling in the cognitive psychology literature. Theoretical debates are focused whether the cases where humans make systematic errors are better explained by mistaken beliefs or inevitable computational limitations (Bass, Smith, Bonawitz, & Ullman, 2021; Kubricht, Holyoak, & Lu, 2017; Ludwin-Peery, Bramley, Davis, & Gureckis, 2020; Sanborn, Mansinghka, & Griffiths, 2013; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). In contrast with the first two groups, developmentalists find that babies are systematically mistaken about mechanics but that these misconceptions resolve during childhood (Hast & Howe, 2015; Hespos & VanMarle, 2012; Lin, Stavans, & Baillargeon, 2022). Theories are introduced to explain that transition from error to ceiling accuracy — something the other literatures agree does not exist.

Hartshorne & Jing (in press) considered and rejected a number of deflationary explantions for the divergence between the literatures. It is unlikely that two of them are simply wrong: all three literatures are large and have compelling arguments for the solidity of their empirical foundations and can point to mutiply-replicated key findings. There is extensive overlap in the physical laws studied — all three literatures have paid particular attention, for instance, to torque and balance — so it is not a matter of simply studying different mechanical phenomena. Systematic attempts to show that poor performance is due to confusing stimuli or that high performance is specific to perceptual or motor tasks have failed.

Instead, Hartshorne & Jing (in press) suggest that just as humans (for instance) have over a half-dozen separable cognitive mechanisms for depth perception, cognitive mechanics may involve a number of different cognitive mechanisms that are differentially involved in different tasks. The reason that different studies obtain different results is that — again as is the case in depth perception — the different systems are more or less good at different kinds of problems. This is not an entirely new idea: a number of researchers have suggested dual-mechanism accounts of various sorts (Dandurand & Shultz, 2014; Heckler, 2011; Hubbard, 2022; Smith, Battaglia, & Tenenbaum, 2023; Wood, Galloway, & Hardy, 2016). However, Hartshorne & Jing (in press)'s review suggests that two mechanisms are unlikely to be enough.

### Factor analysis on concept inventories

One of the primary tools used in the education literature to study cognitive mechanics is the concept inventory: a short

standardized test (usually multiple-choice) intended to probe students' understanding of basic concepts and detect common misconceptions. The first concept inventory was the Mechanics Diagnostic Test (Halloun & Hestenes, 1985), though this was soon replaced in popularity by the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992), which remains the most popular — though far from the only — concept inventory for physics today. Mechanics concept inventories are widely used not only in basic research but also to assess pedagogical interventions, compare effectiveness of instructors, and monitor student progress (Libarkin, n.d.; Sands, Parker, Hedgeland, Jordan, & Galloway, 2018). Concept inventories have also been developed for many other fields, including other subfields of physics, as well as topics in fields as diverse as biology, computer science, and psychology (D'Avanzo, 2008; Taylor et al., 2014; Veilleux & Chapman, 2017).

There has been an admirable amount of attention paid to the psychometric properties of these inventories in order to clarify what exactly it is that they measure and how well they measure it. One question of particular interest is whether variation in performance is best characterized by a single underlying factor (i.e., just being better at cognitive mechanics in some holistic sense) or by multiple separable factors? Does acquiring veridical knowledge involve improving one skill or piece of knowledge, or does it require several?

Results have been mixed. Correlations between mechanics concept inventories are often close to the noise ceiling, which would suggest the each measure the same thing Riener, Proffitt, & Salthouse (2005). Factor analyses, on the other hand, tend to find evidence of as many as five underlying factors, though the structure of these factors varies a great deal across studies and analytic methods, and are difficult to compare across studies of different inventories (Eaton & Willoughby, 2018; Huffman & Heller, 1995; Terry F. Scott & Schumayer, 2017; Terry F. Scott, Schumayer, & Gray, 2012; Stewart, Zabriskie, DeVore, & Stewart, 2018; Yang, Zabriskie, & Stewart, 2019).

In the current study, we combine five multiple-choice mechanics concept inventories — all the ones we could obtain — into a single test. This directly addresses the problem of comparing factor analyses across different studies involving different concept inventories, as all the data is fit in a single model. It also addresses a related issue, which is that most concept inventories are quite short: the popular Force Concept Inventory consists of 29 items, and the Force, Velocity, and Acceleration Test is only 17. Many are designed to have a diversity of items. This can make it difficult to distinguish between one or two "funny" items and a coherent subset of items that behave differently from the rest.

We analyze the data with a combination of Item Response Theory and Principal Component Analysis in order to test for clear evidence of more than one underlying factor driving performance on these tests. To preview, we do find such evidence and conduct a preliminary investigation of the nature of those different factors.

### Scope and Limitations

We focus on concept inventories. The underlying factor structure of concept inventories may well underestimate the number of mechanisms playing a role in cognitive mechanics more broadly. In future work, we intend to conduct a similar study that includes other kinds of tasks that have been used to study cognitive mechanics in the development and cognitive psychology literatures.

Principal Components Analysis (PCA) itself has limitations. It assumes that the underlying factors are additive, that there are no ceiling or floor effects, etc. — none of which is likely to be strictly true. Moreover, PCA and other factor analysis methods do not readily distinguish between multiple separable **systems** for cognitive mechanics and multiple distinct **components** of a single system that are differentially important for different stimuli.

In many cases, authors of concept inventories included lures that are intended to reveal misconceptions. PCA, which requires dichotomous outcome data, will not allow us to pick up on separable mechanisms that are differentially involved in different misconceptions. The are factor analysis methods that would take into account which answer an individual chose, but these methods require a great deal more subjects than we have tested so far. Moreover, for many of the concept inventories, the nature of the different lures is not well-documented, so interpretation would not be straightforward.

This last point raises one additional salient limitation, which is that currently there are no theories sufficiently well-specified so as to make quantitative predictions. As mentioned above, there are several different dual-mechanism proposals, but none make clear predictions about which items would load on which factor. We are unaware of any accounts predicting specifically three factors (or four, etc.). Thus, the present study is exploratory and data-driven. The hope is that it will help generate theory which can then be tested in future work.

## Methods

### Participants

Participants were recruited on Prolific. A total of 573 English-speaking participants aged 18 or older completed the experiment. After excluding subjects who missed more than one catch trial, 463 remained.

### Materials and Procedure

The survey consisted of instructions, a comprehension check, the Force Concept Inventory (*FCI*) (Hestenes et al., 1992), the Dynamics Concept Inventory Assessment (*DCIA*) (Gray, Costanzo, Evans, Cornwell, & Self, 2005), the Mechanics Diagnostic Test (*MDT*) (Halloun & Hestenes, 1985), the Force and Motion Conceptual Evaluation (*FMCE*) (Ronald K. Thornton & Sokoloff, 1998), and the Force, Velocity, and Acceleration Test (*FVAT*) (Rosenblatt & Heckler, 2011), attention checks, and a short demographics questionnaire asking age, gender, and native language. The full Dynamics Con-

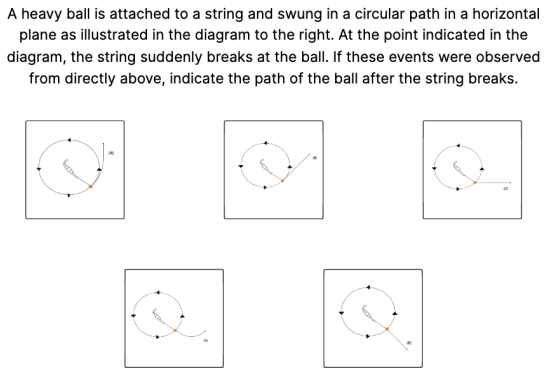


Figure 1: A classic question from the Force Concept Inventory.

cept Inventory (Gray et al., 2005) was not available online, so only questions that were listed as examples on the paper cited were included in the augmented survey. Duplicate questions between two separate concept inventories were omitted. For example, Question 13 and 14 from the Force Concept Inventory were removed from the augmented survey since questions 35-38 in the Force and Motion Conceptual Evaluation were similar. Examples are shown in the figures.

Because these concept inventories were designed several years ago to be taken as paper-and-pencil exams, several modifications were made to adapt it to be taken on the computer. For example, questions were displayed one at a time on their own screen. The wording of the questions was adapted to fit the experimental survey as well. For example, the phrase "you have chosen in question (14)" in question 16 from the Mechanics Diagnostic Test was changed to "you have chosen in question before". Many of the drawings and figures were redrawn for improved aesthetics and ease of reading. On the Force and Motion Conceptual Evaluations, one of the nine options on questions 14-21 was removed. When selecting from a set of figures, subjects could respond by clicking one of the pictures, rather than entering a letter or number corresponding to the picture. There were a total of 121 critical items.

We additionally included 8 catch trials (trivially easy questions) to ensure that participants were paying attention (Ex: Please indicate your agreement with the statement: If you drop a large iron ball and a small iron ball from teh top of a tower, the small iron ball can fly away instead of falling. Strongly disagree, Disagree, Agree, Strongly Agree). Participants who answered fewer than 7 of these questions correctly were excluded from the analysis.

Finally, because we expected overall accuracy to be quite low on the critical items, we also included six easy (but not trivial) questions. These were used for a variety of sanity checks — for instance, confirming that participants who failed the catch trials also did worse on the easy items. (Accuracy on critical items is so low it would be difficult to detect any relationship between performance on the catch trials and on the critical items.) Easy items were not used in the factor analysis.

The experiment was implemented using Gorilla (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020).

## Results

As expected, accuracy on the critical items was low ($M$ = 0.29, 95% CI[0.28, 0.30], $SD$ = 0.11, $N$ = 463). This comports well with what was been reported in the literature. However, this was not because subjects could not answer mechanics questions; accuracy on the added "easy" questions was reasonable ($M$ = 0.81, 95% CI[0.79, 0.82], $SD$ = 0.19, $N$ = 463).

We conducted a Principal Components Analysis (PCA) on the critical items. The scree plot suggested inclusion of four factors.

There was substantial range in factor loadings across items (see histograms). We estimated item difficulty for each item using a three-parameter Item Response Theory model. The correlation between difficulty and loading was not significant for any of the four factors (ps $> .05$). There was, however, a positive relationship with the likelihood of correctly guessing the answer to a question for the first factor ($r = .25$, 95% CI $[.07, .41]$, $t(119) = 2.81$, $p = .006$) and the third ($r = .24$, 95% CI $[.07, .40]$, $t(119) = 2.74$, $p = .007$).

| CI | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| DCIA | -0.03 (0.1) | -0.03 (0.14) | 0.13 (0.13) | -0.04 (0.08) |
| FCI | 0.21 (0.22) | 0.11 (0.2) | 0.11 (0.15) | 0.11 (0.13) |
| FMCE | 0.23 (0.24) | 0.08 (0.21) | 0.22 (0.2) | 0.11 (0.17) |
| FVAT | 0.05 (0.14) | 0.17 (0.18) | 0.16 (0.12) | 0.05 (0.2) |
| MDT | 0.11 (0.19) | 0.14 (0.21) | 0.09 (0.12) | 0.13 (0.14) |

Table 1: Mean (sd) factor loadings for each concept inventory

Otherwise, there were few immediately obvious broad patterns. For the most part, there was no difference in average factor loadings across the different concept inventories, with the exception of FMCE, which on average loaded mildly on the first and third factors. This patterns was stronger, however, when we separated out the large minority of items that involved interpreting graphs depicting velocity, acceleration, or force over time.

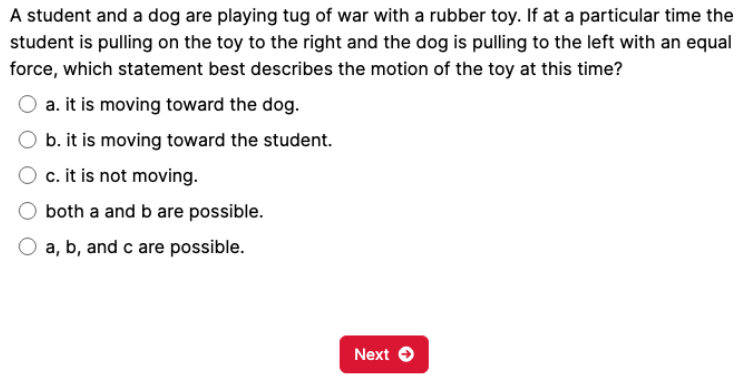


Figure 2: Another example of a concept inventory question.

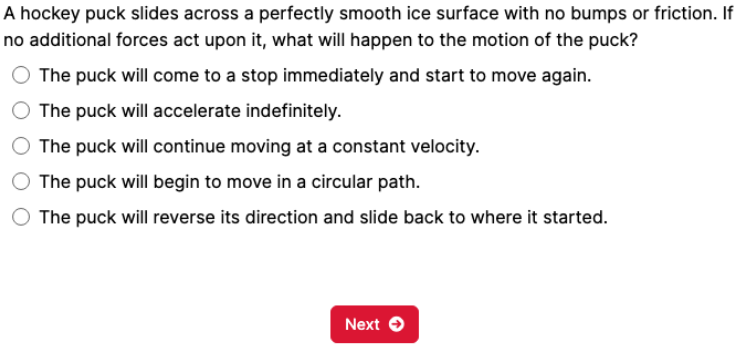


Figure 3: An example of an easy question.

| graphs | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| FALSE | 0.35 (0.23) | 0.04 (0.15) | 0.32 (0.17) | 0.03 (0.12) |
| TRUE | 0.04 (0.06) | 0.15 (0.27) | 0.04 (0.11) | 0.24 (0.17) |

Table 2: Mean (sd) factor loadings for FMCE items involving or not involving graphs

There were a relatively large number of questions that dealt a ball exiting circular motion (N = 7) or an object tossed into the air (N = 12). These two basic problem types have played an outsized role in the history of the study of cognitive mechanics. Interestingly, they do not pattern particularly similarly with respect to the PCA.

| type2 | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| circular.motion | 0.02 (0.06) | 0.13 (0.22) | -0.01 (0.06) | 0.11 (0.17) |
| toss | 0.28 (0.17) | 0.01 (0.17) | 0.37 (0.19) | 0.06 (0.08) |

Table 3: Mean (sd) factor loadings for two classic types of problems

## Discussion and Conclusions

We tested subjects understanding of mechanics using five concept inventories. We found evidence for four factors underlying performance. This is consistent with growing evidence that performance in cognitive mechanics is not a unitary skills but likely involves multiple underlying mechanisms.

Characterizing those mechanisms remains an open challenge. There were no obvious patterns. Some prior work has tried to tease apart understanding of different mechanical laws, formulas, and concepts. This is not straightforward, however, in that many questions invole more than one. There is also a question of how fine-grained to make the distinctions. Too fine-grained, and there are potentially a couple dozen categories and only a few items in each. Too coarse-grained, and we are likely to miss patterns even if they are there.

While trying different classification schema for the items may be helpful, another potentially useful direction would be to include other kinds of tasks, such as those commonly used in cognitive psychology, and again use factor analysis. This would provide more angles from which to look at the problem. Another potentially useful option is to use more complex forms of factor analysis, as discussed above. This does, however, require a lot more participants.

## Acknowledgements

Funding was provided by NSF 2449029.

## References

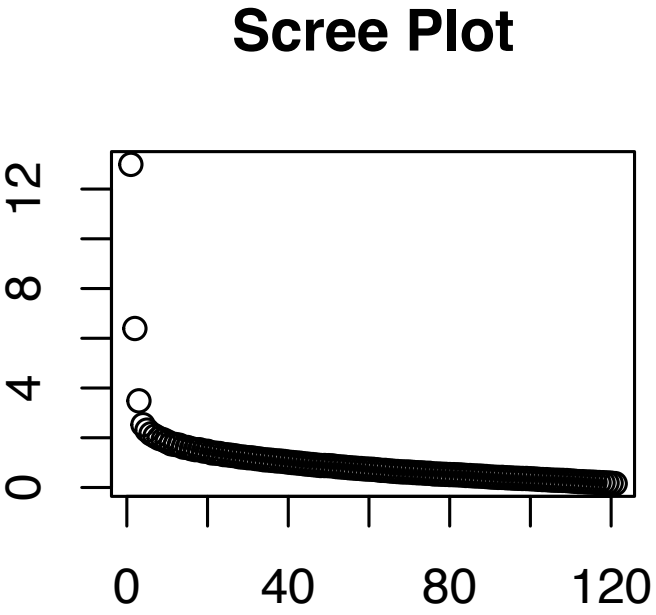


Figure 4: Scree plot

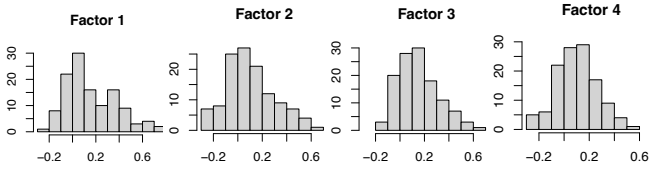


Figure 5: Histograms of factor loadings


Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407.

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*(7-8), 413–424.

Brown, D. E., & Hammer, D. (2009). Conceptual change in physics. In *International handbook of research on conceptual change* (pp. 155–182). Routledge.

D'Avanzo, C. (2008). Biology concept inventories: Overview, status, and next steps. *BioScience*, *58*(11), 1079–1085.

Dandurand, F., & Shultz, T. R. (2014). A comprehensive model of development on the balance-scale task. *Cognitive Systems Research*, *31*, 1–25.

Eaton, P., & Willoughby, S. D. (2018). Confirmatory factor analysis applied to the force concept inventory. *Phys. Rev. Phys. Educ. Res.*, *14*, 010124. http://doi.org/10.1103/PhysRevPhysEducRes.14.010124

Gray, G., Costanzo, F., Evans, D., Cornwell, P., & Self, B. (2005). The dynamics concept inventory assessment test: A progress report and some results. *Proceedings of The IEEE - PIEEE*.

Halloun, I. A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, *53*(11), 1056–1065. http://doi.org/10.1119/1.14031

Hartshorne, J. K., & Jing, M. (in press). Insights into cognitive mechanics from education, developmental psychology, and cognitive science. *Nature Reviews Psychology*.

Hast, M., & Howe, C. (2015). Children's predictions and recognition of fall: The role of object mass. *Cognitive Development*, *36*, 103–110.

Heckler, A. F. (2011). The ubiquitous patterns of incorrect answers to science questions: The role of automatic, bottom-up processes. *Psychology of Learning and Motivation-Advances in Research and Theory*, *55*, 227.

Hespos, S. J., & VanMarle, K. (2012). Physics for infants: Characterizing the origins of knowledge about objects, substances, and number. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(1), 19–27.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*, 141–158. http://doi.org/10.1119/1.2343497

Hubbard, T. L. (2022). The possibility of an impetus heuristic. *Psychonomic Bulletin & Review*, *29*(6), 2015–2033.

Huffman, D., & Heller, P. (1995). What does the force concept inventory actually measure? *Phys. Teach.*, *33*(3), 138–143.

Karmiloff-Smith, A., & Inhelder, B. (1974). If you want to get ahead, get a theory. *Cognition*, *3*(3), 195–212.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759.

Libarkin, J. (n.d.). Concept inventories in higher education science. In.

Lin, Y., Stavans, M., & Baillargeon, R. (2022). Infants' physical reasoning and the cognitive architecture that supports it. *Cambridge Handbook of Cognitive Development*, 168–194.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, *31*(12), 1602–1611.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–131. Retrieved from http://www.jstor.org/stable/24968881

Resbiantoro, G., Setiani, R., et al. (2022). A review of misconception in physics: The diagnosis, causes, and remediation. *Journal of Turkish Science Education*, *19*(2).

Riener, C., Proffitt, D. R., & Salthouse, T. (2005). A psychometric approach to intuitive physics. *Psychonomic Bulletin & Review*, *12*(4), 740–745. http://doi.org/10.3758/BF03196766

Rosenblatt, R., & Heckler, A. F. (2011). Systematic study of student understanding of the relationships between the

directions of force, velocity, and acceleration in one dimension. *Phys. Rev. ST Phys. Educ. Res.*, *7*, 020112. http://doi.org/10.1103/PhysRevSTPER.7.020112

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411.

Sands, D., Parker, M., Hedgeland, H., Jordan, S., & Galloway, R. (2018). Using concept inventories to measure understanding. *Higher Education Pedagogies*, *3*(1), 173–182.

Scott, Terry F., & Schumayer, D. (2017). Conceptual coherence of non-newtonian worldviews in force concept inventory data. *Physical Review Physics Education Research*, *13*(1), 010126.

Scott, Terry F., Schumayer, D., & Gray, A. R. (2012). Exploratory factor analysis of a force concept inventory data set. *Phys. Rev. ST Phys. Educ. Res.*, *8*, 020105. http://doi.org/10.1103/PhysRevSTPER.8.020105

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481–520.

Smith, K., Battaglia, P., & Tenenbaum, J. (2023). Integrating heuristic and simulation-based reasoning in intuitive physics.

Stewart, J., Zabriskie, C., DeVore, S., & Stewart, G. (2018). Multidimensional item response theory and the force concept inventory. *Phys. Rev. Phys. Educ. Res.*, *14*, 010137. http://doi.org/10.1103/PhysRevPhysEducRes.14.010137

Taylor, C., Zingaro, D., Porter, L., Webb, K. C., Lee, C. B., & Clancy, M. (2014). Computer science concept inventories: Past and future. *Computer Science Education*, *24*(4), 253–276.

Thornton, Ronald K., Kuhl, D., Cummings, K., & Marx, J. (2009). Comparing the force and motion conceptual evaluation and the force concept inventory. *Phys. Rev. ST Phys. Educ. Res.*, *5*, 010105. http://doi.org/10.1103/PhysRevSTPER.5.010105

Thornton, Ronald K., & Sokoloff, D. R. (1998). Assessing student learning of newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics*, *66*(4), 338–352. http://doi.org/10.1119/1.18863

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Veilleux, J. C., & Chapman, K. M. (2017). Development of a research methods and statistics concept inventory. *Teaching of Psychology*, *44*(3), 203–211.

Vosniadou, S. (2019). The development of students' understanding of science. In *Frontiers in education* (Vol. 4, p. 32). Frontiers Media SA.

Wood, A. K., Galloway, R. K., & Hardy, J. (2016). Can dual processing theory explain physics students' performance on the force concept inventory? *Physical Review Physics Education Research*, *12*(2), 023101.

Yang, J., Zabriskie, C., & Stewart, J. (2019). Multidimensional item response theory and the force and motion conceptual evaluation. *Phys. Rev. Phys. Educ. Res.*, *15*, 020141. http://doi.org/10.1103/PhysRevPhysEducRes.15.020141