

Confidence versus Performance as Indicator of the Presence of Alternative Conceptions and Inadequate Problem Solving Skills in Mechanics

Marietjie Potgieter^{a*}, Esther Malatje^b, Estelle Gaigher^c and Elsie Venter^d

International Journal of Science Education, 2010, 32(11), 1407 – 1429.

^aDepartment of Chemistry, UP; ^bDepartment of Education, ^cFaculty of Education, UP; ^dCentre for Evaluation and Assessment, UP

This study investigated the use of performance-confidence relationships to signal the presence of alternative conceptions and inadequate problem-solving skills in mechanics. A group of 33 students entering physics at a South African university participated in the project. The test instrument consisted of 20 items derived from existing standardized tests from literature, each of which was followed by a self-reported measure of confidence of students in the correctness of their answers. Data collected for this study included students' responses to multiple-choice questions and open-ended explanations for their chosen answers. Fixed response physics and confidence data were logarithmically transformed according to the Rasch model to linear measures of performance and confidence. The free response explanations were carefully analysed for accuracy of conceptual understanding. Comparison of these results with raw score data and transformed measures of performance and confidence allowed a re-evaluation of the model developed by Hasan, Bagayoko and Kelley in 1999 for the detection of alternative conceptions in mechanics. Application of this model to raw score data leads to inaccurate conclusions. However, application of the Hasan hypothesis to transformed measures of performance and confidence resulted in the accurate identification of items plagued by alternative conceptions. This approach also holds promise for the differentiation between over-confidence due to alternative conceptions or due to inadequate problem-solving skills. It could become a valuable tool for instructional design in mechanics.

Keywords: *Confidence; Performance; Mechanics; Alternative conceptions; Rasch model*

Introduction

In order to ensure a smooth transition between secondary and tertiary education, it is imperative for lecturers of physics to be well informed about the baseline knowledge and understanding of students upon entry to tertiary education. Furthermore, where lack of preparedness is identified lecturers should ideally be able to distinguish between a lack of knowledge about specific concepts and the presence of strong alternative conceptions. It is widely accepted that the instructional design for conceptual change of firmly rooted alternative conceptions will differ markedly from that aimed at acquisition of new knowledge or development of problem-solving skills. Hasan, Bagayoko, and Kelley (1999) proposed that a student's self-reported certainty of response could be used in conjunction with the answer to a conceptual test item to differentiate between lack of knowledge and the presence of misconceptions in mechanics. While acknowledging that their hypothesis is best utilised for individual students, they have expanded it for application to groups of students in an attempt to enable the lecturer to extract this information with relative ease for large groups of students. This paper investigated the use of the relationship between confidence and performance to signal not only the presence of alternative conceptions but also inadequate problem-solving skills in mechanics and as such serve as a practitioner's tool in instructional design.

Literature review

Over the past years, there have been a number of studies focusing on students' understanding of physics concepts (Ates & Cataloglu, 2007; Clement, 1982; Halloun & Hestenes, 1985; Hasan, et al., 1999; Jimoyiannis & Komis, 2001; Knight, 1997; McDermott & Redish, 1999; Planinic, Boone, Krsnik & Beilfuss, 2006). Jimoyiannis and Komis (2001) and Knight (1997) documented that secondary and university physics students have limited basic knowledge and thus have difficulties in the understanding of mechanics. This lack of basic knowledge and understanding has an impact on the understanding of other more complex topics at higher levels of physics. Not only have lecturers to deal with a lack of basic knowledge, but also with the presence of alternative conceptions about physical processes, which are developed from very young ages through experiences, observation and interactions with the environment. These ideas interfere with the learning and understanding of physics concepts. This is especially true of basic mechanics, which is considered to be the “cradle” of alternative conceptions because it is close to the students' daily life experiences of motion and forces (Planinic et al., 2006; Ramaila, 2000). Intuitive or alternative conceptions in mechanics require special attention because they are transferred to other physics topics where they create learning problems as well (Galili, 1995). Hasan et al. (1999) have defined student *misconceptions* as ‘strongly held cognitive structures that are different from the accepted understanding in a field and that are presumed to interfere with the acquisition of new knowledge’. Heller and Finley (1992) refer to these as *intuitive conceptions*, but Dykstra, Boyle and Monarch (1992) prefer the expression alternative conceptions because it implies that they are ‘rationally based on the students' experiences with the world and prove adequate for ... most everyday tasks ...’ (p. 621). We have used the general term alternative conceptions to include all of these perspectives.

In psychology literature, the relationship between confidence and performance has been studied extensively, primarily with the view to explore the dynamics of *confidence bias*, which is the systematic error made by individuals in assessing the correctness of their responses relating to intellectual or perceptual problems (e.g., Pallier, et al., 2002). The use of confidence levels in mathematics or science education research is limited, but has been applied in chemistry (Potgieter, Rogan & Howie, 2005), mathematics (Yazdani, 2006), biology (Bowen & Roth, 1999) and mechanics (Oliva, 1999; Reif & Allen, 1992). Planinic et al. (2006) have also used certainty of response (self-reported confidence) coupled with performance to explore relative strengths of misconceptions in different areas of physics. The probabilistic Rasch model (Rasch, 1960) was used for data analysis to facilitate objective comparison of students and items across different subject domains. They have shown that students show higher confidence levels on Newtonian mechanics

than on simple DC circuits for both correct and incorrect answers of similar difficulty levels. This result was interpreted to signal the presence of strongly held alternative conceptions in Newtonian mechanics as compared to the absence of strong conceptual models of any kind to support reasoning and problem solving in the case of DC circuits.

Hasan Model

According to the *hypothesis* of Hasan et al. (1999) regarding decision-making in mechanics, a low certainty of response would indicate lack of knowledge when combined with either an incorrect answer or a correct answer (a lucky guess). If a student chose a correct answer and reported a high certainty of response, such a student is classified as having adequate knowledge and understanding of the concept, but if a high certainty of response accompanies an incorrect answer it would signify the presence of alternative conceptions (Figure 1). The instrument used for their study consisted of multiple-choice items in which distractors were based on known alternative conceptions. The choice of any of these as the answer to a question was interpreted to signal the presence of that specific alternative conception. Every answer was accompanied by a self-reported certainty of response index (CRI), expressed on a six-point Likert type scale, where a total guess is given a 0 rating and 5 indicates complete confidence in the knowledge of the principles and laws required to arrive at the selected answer (Hasan et al., 1999). A CRI value above the numerical average of 2.5 was considered to be high and below 2.5 to be low.

	Low CRI	High CRI
Correct answer	<p>Correct answer and low CRI Lack of knowledge (lucky guess)</p>	<p>Correct answer and high CRI Knowledge of correct concepts</p>
Wrong answer	<p>Wrong answer and low CRI Lack of knowledge</p>	<p>Wrong answer and high CRI Misconceptions</p>

Figure 1. Decision matrix for an individual student and for a given question, based on combinations of correct or wrong answers and of low or high CRI (Hasan et al., 1999).

In this paper we distinguish between the Hasan *hypothesis* for decision-making by an individual student (as shown in Figure 1) and the Hasan *model* for student groups (Hasan et al., 1999) which will now be described. When dealing with a group of students, the identification of alternative conceptions was done in a similar manner as the analysis for an individual student. For a given test item, correct and incorrect responses were grouped separately and the *average* CRI associated with each calculated. The average CRI for correct and incorrect answers were utilized in conjunction

with the fraction of students choosing the correct answer, to decide whether the student group in general had alternative conceptions or were lacking knowledge and conceptual understanding. For example, a high average CRI for a correct answer and a high average CRI for an incorrect answer coupled with a low fraction of students choosing the correct answer was interpreted to signal a presence of alternative conceptions. A low average CRI for a correct answer and low average CRI for an incorrect answer coupled with low fraction of students choosing the correct answer was interpreted to suggest lack of knowledge of suitable principles and scientific laws. Even those respondents who answered correctly felt uncertain about their responses. In situations where, for a given item, the average CRI value for correct and incorrect answers were close to the numerical average, the authors utilized the fraction of correct answers to decide whether the CRI value is judged to be high or low, and hence decide on the presence of alternative conceptions. For example, if the CRI for incorrect answers is close to the numerical average, and a large fraction of students have chosen incorrect answers, then the implication is the large number of students who have chosen wrong answers were quite confident about their choices. This situation thus signals that a large fraction of students had alternative conceptions. Teaching should then be geared toward addressing the specific alternative conceptions reflected in the response frequencies to the multiple-choice physics question. On the other hand, if the CRI for correct answers is very close to the numerical average, and a large proportion of students have chosen the correct answer, then the implication is that students felt insecure about their choices despite its accuracy. In this case the authors suggest that nothing special needs to be done for the whole class apart from reinforcement, since only a small fraction of students seem to have alternative conceptions (Hasan et al. 1999). The application of this model is not without complications, but it offers the attractive possibility of determining areas where special attention is required without having to resort to the demanding task of interviews or detailed analysis of written responses.

The multiple-choice format of assessment as employed by Hasan et al. (1999) is widely used for large first-year cohorts, but its limitations are well known. In the context of this study the most serious issues related to the use of multiple-choice test items are detection of logical reasoning, valid assessment of conceptual understanding, restriction of guessing, and prevention of plagiarism. In order to address these issues and explore the scope of the Hasan model of interpretation of multiple-choice data as a means to distinguish between students' misconceptions and lack of knowledge, we propose that unstructured explanations of answers to multiple-choice questions are required in addition to answers to multiple-choice items. These explanations must then be analysed for indications of either appropriate conceptual understanding or of conceptions that are not scientifically acceptable. The next step would then be to check whether alternative conceptions that

were identified in this way would have been reliably signalled by the average CRI values for correct answers and incorrect answers, respectively, combined with average performance as suggested by the Hasan model.

Methodology

This study forms part of a bigger project to document the baseline knowledge and understanding in mechanics of South African students upon entry to tertiary education. We have developed a test instrument consisting of multiple-choice items taken from existing standardized tests from the literature, mainly from the Force Concept Inventory (FCI) (Hestenes, Wells & Swackhamer, 1992) and the Mechanics Baseline Test (MBT) (Hestenes & Wells, 1992). The two tests, FCI and MBT, were chosen because they are complementary to each other. They probe for the students' understanding of the most basic mechanics concepts and for the mastery of basic problem-solving skills in mechanics (Hestenes & Wells, 1992). Items in the test instrument require minimal computation to arrive at correct answers and their scope is limited to the concepts that are addressed in elementary physics at the secondary school level in South Africa. Table 1 specifies the origin of test items used in our instrument.

Table 1. Origin of items used in the instrument for this study

Item number in our instrument	Corresponding item number in source documents ^a	Item number in our instrument	Corresponding item number in source documents ^a
6	FCI 1	16	FCI 21
7	FCI 2	17	MBT21
8	FCI 5	18	FCI 28
9	FCI 23	19	MBT 1
10	MBT13	20	MBT 2
11	MBT 14	21	The Physics Classroom (2005)
12	MBT 17	22	MBT 7
13	FCI 18	23	FCI 24
14	FCI 19	24	FCI 4
15	FCI 20	25	FCI 10

^aFCI (Hestenes et al., 1992) and MBT (Hestenes & Wells, 1992)

The instrument was piloted in 2005 with senior secondary physical science teachers, physics lecturers from University of Limpopo, and Foundation Year students at the same institution. Based on the results of the pilot study and the comments of the educators, the test was refined, resulting in the removal of five items from the instrument. The final paper-and-pencil test consisting of 25 items

had two sections. Section A, made up of 5 items, required the students to report on their educational background. Section B, consisting of 20 items, probed for students' conceptual understanding of concepts in mechanics. It included 12 items from FCI, seven from MBT and a single item that was obtained and adapted from a question in 'The Physics Classroom' (2005) to strengthen the selection of items on Newton's second law. Each item in Section B of the test instrument had three parts. The first part was a statement in the form of a physics question followed by four or five options (A, B, C, D and E) to choose from. The second part required that the students give written explanations for their chosen options. This part was included so that the student's knowledge and understanding of relevant concepts could be explored. The third part required that the students indicate their confidence in the correctness of their answers on a four-point scale similar to that used by Planinic et al. (2006): certain (D), almost certain (C), almost a guess (B), or a totally guessed answer (A). The final version of the test instrument was administered at the beginning of the 2006 academic year to first entering physics students at three South African universities. The reliability of the instrument was established on the combined dataset that was obtained ($N = 982$ students).

We selected the best performing cohort for this study so that an in-depth analysis of all of their responses could be performed. This approach was chosen to minimise the influence that a lack of knowledge of basic mechanics concepts or random guessing may have to obscure alternative conceptions, which is the focus of the study reported here. The cohort of 33 mainstream students registered for the first course in physics offered at the University of Pretoria was thus selected for further study. These students had indicated upon registration that they may opt to take physics as a major course in their degree programmes. Their written explanations for each item were then coded by one of the authors and the associated frequencies determined. For each of the items, five to six codes were required, including "no response" and "uncodable response", to accurately reflect all of the responses obtained. Independent verification of coding by a second author confirmed that an intercoder reliability of above 95% was achieved.

The physics test data and confidence data were analysed according to the probabilistic Rasch model (Bond & Fox, 2007). According to this model, raw scores for persons and items are transformed logarithmically from ordinal data to linear measures of proficiency in physics and confidence in their responses on the test items. The basic requirements for the Rasch measurement model as summarised by Bond and Fox (2007) are that each person and each item are characterised by a proficiency (or ability) and a difficulty respectively, both of which can be expressed on a continuum of the underlying construct and that the probability of observing a particular scored response can be computed from the difference between the proficiency and difficulty. Fundamental measurement requires the construction of reliable and valid measures. The rigour of the Rasch

measurement model satisfies this requirement. Data were analyzed using the Ministep Rasch Measurement software (Linacre, 2006).

Validity and Reliability

Both the Force Concepts Inventory (FCI) and the Mechanics Baseline Test (MBI) are standardised tests for which content and face validity have been established (eg. Savinainen & Scott, 2002; Hestenes & Wells, 1992). After selection of some of the items from these instruments for our purposes the reliability and validity of the new instrument had to be re-established. During the piloting of our instrument with physics educators at the University of Limpopo and the University of Pretoria content and face validity of the instrument for the South African context for which it was intended, were confirmed. The Cronbach alpha values for physics and confidence data as reported in Table 2 below confirms that the level of internal consistency was acceptable and comparable to those associated with the source instruments (alpha reliability coefficient for FCI = 0.80 and KR-20 coefficient for MBT = 0.70, Ates and Cataloglu, 2007).

One of the important characteristics of the Rasch model is that construct validity points to the fact that, in the words of Andrich (1988), ‘the actual properties are not observed – only their manifestations are observed’ (p.14). The fit statistics, which provides empirical evidence about whether this requirement of unidimensionality and a latent variable or construct is upheld, were checked for the items; Item 22 proved to be problematic, and Item 10 to a lesser extent. As will be discussed below, Item 22 posed a complex problem where students were handicapped by both a physics misconception and inadequacy of mathematical insight (components of forces). Similarly Item 10 requires sophisticated analysis and accurate application of Newton’s laws, so-called Newtonian thinking (Hestenes & Halloun, 1985). Explanations provided for correct answers to Item 10 revealed that students often grasped the key idea, but lacked the in-depth understanding needed to provide a scientifically meaningful explanation. After scrutinising the free responses for these two items we were certain that a high level of construct validity is exhibited, in other words all the items contribute on a continuum to the underlying construct of basic mechanics knowledge and skills. The Rasch person separation index reliability (similar to the ‘test’ reliability of traditional test theory) was 0.73 and the Rasch person confidence reliability was 0.89, both indicating a fairly strong reliability.

Results

Analysis of the written responses revealed that the instrument included one item that had an ambiguous problem statement. This was the only item obtained from an outside source (The

Physics Classroom, 2005). It was not clear in Item 21, which dealt with a ticker tape trace representing the motion of a car, whether the ticker tape was attached to the car with the ticker timer stationary or whether the ticker timer was attached to the car with the ticker tape stationary. We have decided to accept both scenarios as valid, which meant that this item had two correct answers. The free response explanations were interpreted accordingly. This decision was validated by the results of data analysis according to the Rasch model. The unacceptable misfit statistics associated with Item 21 disappeared when the raw data was corrected to accommodate both answers. The item we used in our test has since been updated on “The Physics Classroom” (2009).

Fixed Response Raw Data

The overall performance and confidence levels of students participating in the study are shown below in Table 2. The performance of each student in Section B of the test was calculated by allocating a score of 1 to the scientifically correct answer and a score of 0 to each of the incorrect options. The scores for performance of individual students were added to obtain an average performance score for the cohort. Fifty five percent of students (18 of 33) obtained a performance score above ten (10) out of a maximum of 20. The confidence levels of the respondents were calculated by allocating a score of 0, 1, 2 or 3 to the response categories A (a totally guessed answer), B (almost a guess), C (almost certain), and D (certain), respectively. The range of the Likert confidence scale from zero to three resembles the scale of Hasan et al. (1999) and reflects the fact that the choice of category A indicates a complete absence of confidence. The confidence score or CRI for an individual student was calculated as the average of the scores obtained by a student in all 20 items.

Table 2. Performance and confidence results based on raw data

Number of students	33
Average test performance	11.36 (max = 20) Std dev = 3.72
Average confidence level	2.22 (max = 3.0) Std dev = 0.57
Pearson correlation coefficient: performance and confidence per student	0.56
Cronbach alpha (physics answers) ^a	0.72
Cronbach alpha (confidence responses) ^a	0.93

^aCronbach (1951).

Fifty five percent of students (18 of 33) obtained a performance score above 10 out of a maximum of 20. The average confidence level of 85% of respondents (28 of 33) was above the

numerical average value of 1.5. The fairly weak correlation between performance and CRI for students in this cohort indicates that only 31% of the variance in students' confidence levels, CRI, is accounted for by their test performance.

Table 3 provides a summary of the most important information obtained from raw score data and open response explanations. Each item is briefly described and the performance of the cohort reported in terms of the percentage of correct answers recorded. The average CRI values associated with correct and incorrect answers, respectively, are reported for each item as well as the frequencies of alternative conceptions and incorrect explanations determined from analysis of the free response component of the item. Figure 2 shows the average CRI values for correct and incorrect responses in a graphic presentation similar to that used by Hasan et al. (1999).

Table 3. Performance of the UP mainstream cohort with average CRI for correct and incorrect answers, and frequencies of alternative conceptions and incorrect explanations as revealed by their written responses

Item	Description	Performance (% correct)	Average CRI		Alternative conception/ <i>Incorrect explanation</i> ^a	Freq (%)
			Correct answers	Incorrect answers		
6	Acceleration of falling bodies is independent of the mass of objects.	78.8	2.7	2.0	<i>Gravity exerts the same force on objects. No discrimination between gravitational force and gravitational acceleration.</i>	15.2
					A heavy object falls faster than a light object.	18.2
7	Impulsive forces: The magnitude of forces of the same interaction does not depend on the masses of the objects involved.	57.6	2.7	2.2	A bigger mass exerts a bigger force.	36.4
8	For vertical motion (both upwards and downwards), the gravitational force always acts downwards on objects	36.4	2.7	2.2	Impetus dissipation followed by increasing gravity as the object falls.	45.5
9	Parabolic motion: Trajectory motion of an object dropped from a moving aeroplane, viewed from the ground as frame of reference.	57.6	2.4	2.0	<i>Trajectory curves backwards due to friction.</i>	21.2
10	Cancelling forces on a mass when it is pulled upwards at constant speed.	63.6	2.2	1.8	None (No explanation: 12%)	--
11	Cancelling forces on a mass hanging at rest, when another mass is hanging from it.	75.5	2.2	1.5	None (No explanation: 15%)	--
12	Inverse relationship between the mass of the object and its acceleration at constant force.	36.4	2.3	2.2	<i>Oversight: assumed that mass doubled instead of tripled.</i>	48.5
13	The vector sum of vertical forces on an object moving upwards at a constant speed.	63.6	2.6	2.1	Motion implies active force.	33.3

Item	Description	Performance (% correct)	Average CRI		Alternative conception/ <i>Incorrect explanation</i> ^a	Freq (%)
			Correct answers	Incorrect answers		
14	The resultant of forces acting on a crate that is being pulled on a rough floor by a man and a boy.	57.6	2.3	2.2	<i>Diagram mistaken for a vector diagram.</i>	36.4
15	Interpretation of ticker tape trace: Differentiation between speed and position.	57.6	2.5	1.5	<i>Inadequate understanding of instantaneous velocity</i>	18.2
16	Interpretation of ticker tape trace: Discrimination between speed and acceleration.	66.7	2.7	1.2	Acceleration and velocity undiscriminated.	15.2
17	Relationship between mass and acceleration, when the applied force remains constant.	87.9	2.4	2.0	None	--
18	A comparison of the magnitudes of the forces acting on a box that is being pushed across a rough floor at constant speed.	54.5	2.4	2.0	Applied force overcomes friction. (No explanation: 12%)	18.2
19	Interpretation of multiframe diagram and the transformation of the multiframe into a velocity-time graph.	66.7	2.5	1.9	<i>Multiframe: Poor interpretation and/or translation to velocity-time graph.</i>	30.4
20	Interpretation of multiframe diagram and the transformation of the multiframe into an acceleration-time graph.	69.7	2.3	1.9	<i>Multiframe: Poor interpretation and/or translation to acceleration-time graph.</i> (No explanation: 12%)	15.2
21	Interpretation of ticker tape trace: direction of the applied force and the acceleration.	63.7	2.4	2.4	Motion implies active force.	27.3
22	Balancing components of forces, acting on a block that is being pulled across a rough horizontal surface.	6.1	3.0	2.6	<i>Incorrect application of components of forces.</i> Applied force overcomes friction.	29.7 54.5
23	Trajectory of a rocket drifting in outer space, if a constant thrust is applied perpendicular to the original velocity.	30.3	2.2	1.7	Loss of original impetus Force compromise determines motion. (No explanation: 33%)	12.1 12.1
24	The selection of the trajectory followed by a ball that initially followed a circular path, when the string breaks.	42.2	2.1	1.6	Circular impetus. (No explanation: 21%)	21.2
25	The selection of the path followed by a ball, when it leaves a semicircular channel.	57.6	2.4	1.8	Circular impetus. (No explanation: 15%)	27.3

^a Faulty explanations with frequencies less than 10% are not included in the table. Incorrect explanations that do not represent known alternative conceptions are shown in italics.

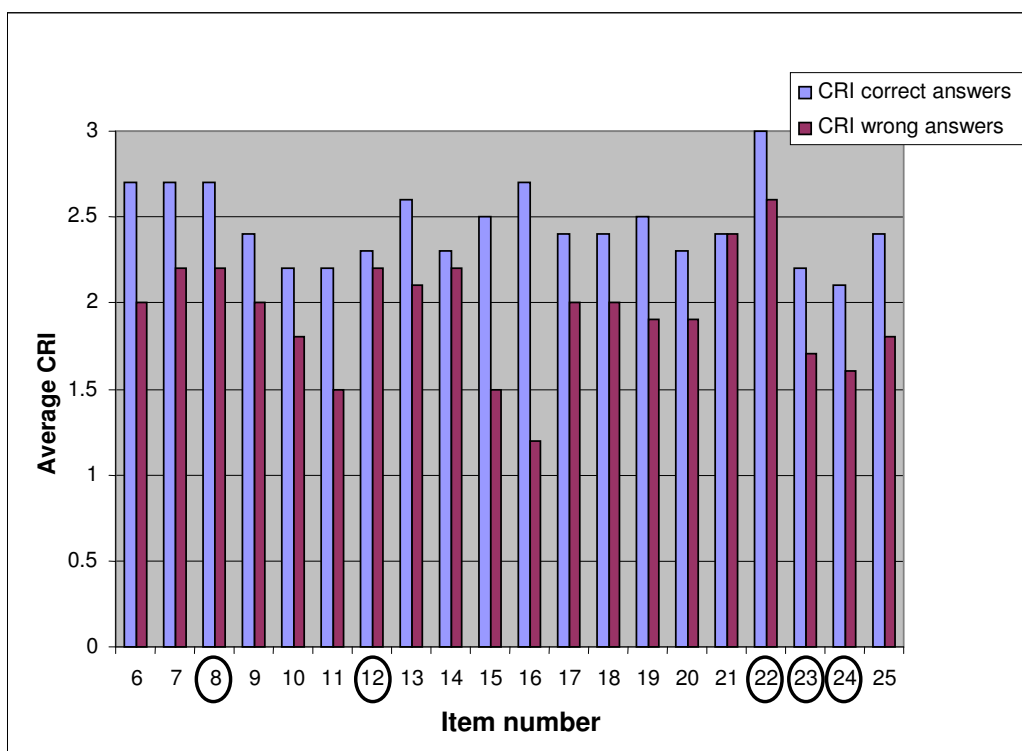


Figure 2. Bar chart of average CRI values per item for correct and incorrect responses. Items with below average performance (<50% correct answers) are circled.

Analysis of Free Response Explanations

The most prevalent alternative conceptions and/or incorrect explanations as revealed by the free response answers are also presented in Table 3. Faulty explanations with frequencies less than 10% are not included in the table since this indicates that only one, two or three students had given the incorrect explanation. This explains the absence of frequency values for Items 10, 11 and 17 in Table 3. The number of uncodable responses on any item did not exceed 9%. However, a number of items were plagued by the absence of explanations. Item 23 represents the extreme case in this regard where 33% of respondents did not provide an explanation for their answer, presumably due to unfamiliarity with the setting of the problem statement (a rocket in outer space). This is followed by Item 24 (21%), Items 11 and 25 (15%) and Items 10, 18 and 20 (12%).

It was clear that many of the incorrect explanations that were given did not arise from strong alternative conceptions. Respondents lacked problem-solving skills, for example the interpretation of graphic representations, or made a critical error in their analysis by overlooking key facts or misinterpreting the information provided. In such cases, the open response explanations revealed that they understood the relevant underlying mechanics concepts, but were unable to apply them correctly. A significant number of items included schematic representations, ticker tape traces or velocity/acceleration graphs (Items 9 - 11, 13 - 17, 19 - 25). Many of these schematic

representations were included in items from the MBT (Hestenes & Wells, 1992), which was designed to assess basic problem-solving skills in mechanics, such as the interpretation of graphic representations. Free responses that were not scientifically accurate were therefore categorised as either *alternative conceptions* or *incorrect explanations* to capture this important difference. The category *incorrect explanations* included error of analysis, incorrect interpretation of graphs, incorrect application of components of forces, and inadequate understanding of concepts for successful application. All of these can be viewed as manifestations of inadequate problem-solving skills in mechanics.

Analysis of Raw Score Confidence-Performance Relationships

In our first line of analysis, the difference between average CRI values for correct and incorrect answers to a specific item is considered. This approach is based on the reasoning that the context of a specific item (familiar or unfamiliar) or the graphic representations included in the problem statement may influence the confidence with which thinking occurs. It provides the platform from which either correct or incorrect conclusions are made. For all items in our instrument, respondents providing correct answers were as confident as or more confident about their answers than those providing incorrect answers. This means that in general, respondents were making good judgments about their knowledge when expressing their confidence in the correctness of answers. However, the degree to which their judgment was accurate would depend on how big the difference between average confidence levels was. From Table 3 and Figure 2, it can be seen that the confidence associated with correct and incorrect answers, respectively, was typically characterised by a difference in average CRI values of 0.4 – 0.7. However, two unique sets of items can be identified that do not follow this general trend. Items 12, 14 and 21 are flagged by their small values for difference in average confidence. On the opposite extreme, Items 15 and 16 display exceptionally large differences between average confidences. From an instructor's point of view there would be more concern about the high confidence associated with incorrect answers to Items 12, 14 and 21 than that of Items 15 and 16. The cause of unjustified confidence must be identified in order to be successfully addressed. Instruction on the application of principles of kinematics (Items 15 and 16) would not require anything more than exposure and reinforcement from the instructor.

Our second line of analysis based on the Hasan model considers performance together with confidence in incorrect answers (Table 3 and Figure 2). From an instructor's point of view, all items with a percentage of correct answers below the average for the test are suspect and the cause of incorrect judgments needs to be determined. When the average CRI for incorrect answers to these questions is below the average for the test, a lack of knowledge is suspected and nothing special is required. However, if the average CRI for incorrect answers is above average, the conclusion will

be that alternative conceptions are to blame and the instructional design will have to be adjusted accordingly. When these criteria are applied to our data, a different list of problem items is obtained. Items 8, 12 and 22 are flagged, based on below average performance ($\leq 57\%$) and relatively high confidence for incorrect answers (≥ 2.2) and Items 7 and 14 are borderline cases for consideration. By contrast, respondents were well aware of their lack of mastery of the concepts necessary to solve the problems posed in the two other items where performance was below average (Items 23 and 24), as judged by the low CRI values for incorrect answers.

Our third line of analysis considers the results of the analysis of free response explanations for answers provided as evidence for the presence of alternative conceptions or inadequate problem-solving skills (Table 3). Several items show a disturbing prevalence of alternative conceptions, for example Item 22 (55%), Item 8 (46%), followed by Item 7 (36%) and Item 13 (33%). A high prevalence of incorrect explanations was revealed in the free responses for Item 12 (49%), Item 14 (36%) and Item 19 (30%). This sends a warning signal regarding the accuracy of conclusions drawn when the Hasan model is applied to raw score confidence-performance data. Because of reasonably good performance, Item 13 would have been completely missed in the search for prevalent alternative conceptions and Item 19 would not have been flagged for its associated lack of problem-solving skills. Furthermore, the Hasan model does not make provision for differentiation between the source of over-confidence as either alternative conceptions or lack of problem-solving skills.

Rasch Analysis

Planinic et al. (2006) have utilised the Rasch model for the evaluation of both attitudinal data (confidence levels) and test data. The raw score data for item difficulty calculations consisted of the percentage of correct answers obtained per item in the test. These percentages were determined by assigning a value of 1 for correct answers and a value of 0 for incorrect answers. A correct answer to any item contributed equally to overall performance irrespective of whether the item was difficult or easy. The Rasch measurement model converts these raw scores into linear measures of item difficulty as well as person ability or proficiency. Raw score performance data and item difficulty measures are presented in Table 4. Item difficulty measures are expressed in terms of logits (log-odds units) which range from negative to positive values, where an item of, for example, 0 logits represents a 50% probability that a respondent with matching ability would answer the item correctly (Bond & Fox, 2007). This hypothetical respondent will have a higher probability of answering items with negative logit values correctly and a lower probability to answer items with positive logit values correctly. Item difficulties can also be transformed to fit into any specified range. It seemed appropriate in our case to specify the range as 0 – 100, with a set mean of 50, to

correspond with the range of 0 – 100% commonly used for raw data on performance (Table 4). It should be noted that high values for raw score performance data are associated with low values for item difficulties. This is to be expected, because the higher the percentage of correct answers for a specific item the easier the item and therefore the lower the corresponding item difficulty.

Table 4. Raw score data and transformed scores for item difficulty and associated confidence values (item difficulty and item endorsability measures as generated by the Rasch model)

Item	Item difficulty			Confidence		
	Raw score (% correct answers)	Item measure (Logit)	Item measure (Normalised 0 – 100)	Raw score (Ave CRI)	Item endorsability (x) ^a (-x) ^a	
6	78.8	-1.46	35.43	2.58	-0.96	0.96
7	57.6	-0.07	49.30	2.52	-0.75	0.75
8	36.4	1.01	60.11	2.36	-0.30	0.30
9	57.6	-0.07	49.30	2.24	0.01	-0.01
10	63.6	-0.53	44.68	2.00	0.56	-0.56
11	75.5	-1.00	39.99	2.00	0.53	-0.53
12	36.4	1.01	60.11	2.18	0.16	-0.16
13	63.6	-0.37	46.26	2.42	-0.47	0.47
14	57.6	-0.07	49.30	2.27	-0.06	0.06
15	57.6	-0.07	49.30	2.06	0.43	-0.43
16	66.7	-0.53	44.68	2.18	0.16	-0.16
17	87.9	-1.98	30.20	2.33	-0.22	0.22
18	54.5	0.08	50.79	2.21	0.09	-0.09
19	66.7	-0.53	44.68	2.24	0.01	-0.01
20	69.7	-0.69	43.05	2.15	0.23	-0.23
21	63.7	-0.37	46.26	2.42	-0.47	0.47
22	6.1	3.83	88.25	2.61	-1.08	1.08
23	30.3	1.37	63.68	1.85	0.87	-0.87
24	42.2	0.53	55.31	1.79	0.99	-0.99
25	57.6	-0.07	49.30	2.12	0.30	-0.30

^a Item endorsability expressed in logits (x) and inversed sign logits (-x).

The conceptualisation of the transformation of raw score data for the confidence levels associated with specific items (average CRI values) into linear measures is more challenging. The Rasch polytomous model was used and the transformation was done according to the procedure that was carefully described by Planinic et al. (2006). One of the strengths of the application of the

Rasch model for rating scales is that the distances between categories are actually estimated rather than assumed, as is the case with traditional theory (The University of Western Australia, 2008). Application of the Rasch model to the confidence level data resulted in linear confidence measures called *item endorsabilities*, which reflect the difficulty (or ease) of endorsing the answer to the item with confidence. According to the Rasch polytomous model, there are three threshold values if a four-point Likert scale is used: between *certain* and *almost certain*, between *almost certain* and *almost a guess*, and between *almost a guess* and *a totally guessed answer* in our case. By Rasch convention, these threshold values for item endorsability add up to 0 for each item. The zero value is therefore a reference point and for convenience, the range of endorsabilities is reflected on a scale from -1.5 to 1.5. The endorsability value is the pivot point where a choice of either the lowest or highest categories is equally likely. In order to simplify reasoning about the meaning of numerical values, and to facilitate interpretation, the Rasch measures for endorsability were multiplied by -1 as suggested by Planinic and co-workers (2006). Following this adjustment, larger numerical values for item endorsability will indicate that the answers provided for a specific test item were associated with higher confidence (Table 4 and Figure 3). A scatter plot of item difficulty versus adjusted item endorsability using the linear measures generated by the Rasch model as reported in Table 4 is shown in Figure 3.

Analysis of relationships between Rasch measures of proficiency and confidence

We are using item difficulty measures as a reflection of student performance and item endorsability as a measure of student confidence for the purpose of analysing the relationship between Rasch measures of performance and confidence. If students were able to make a perfect judgment regarding the correctness of their answers one would expect easy items to be associated with high item endorsability and difficult items with low endorsability. In other words, one would expect a linear relationship between item difficulty and item endorsability. However, where specific items are plagued by the presence of strong alternative conceptions over-confidence is expected and the endorsability of the item should be higher than justified by performance on the item (upper right corner of the graph in Figure 3). In the case where students are able to apply the correct thinking to a problem statement, but do so with hesitation or uncertainty items will appear in the lower left corner of the graph.

The hypothetical situation of perfect judgment can be represented by a diagonal line on the scatter plot in Figure 3. It crosses through the zero value for item endorsability and 50 for item difficulty and has as anchor points the x, y coordinates (0, 1.5) and (100, -1.5). This line merely provides a pragmatic distinction between under-confidence and over-confidence. Figure 3 shows that several of the items appear in fairly close proximity to the hypothetical diagonal line. The most

obvious exception is item 22, which was the most difficult item to answer correctly, but solicited the highest confidence in the correctness of the answer, confidence that was clearly completely unjustified. The meaning of the placement of the rest of the items will be discussed below.

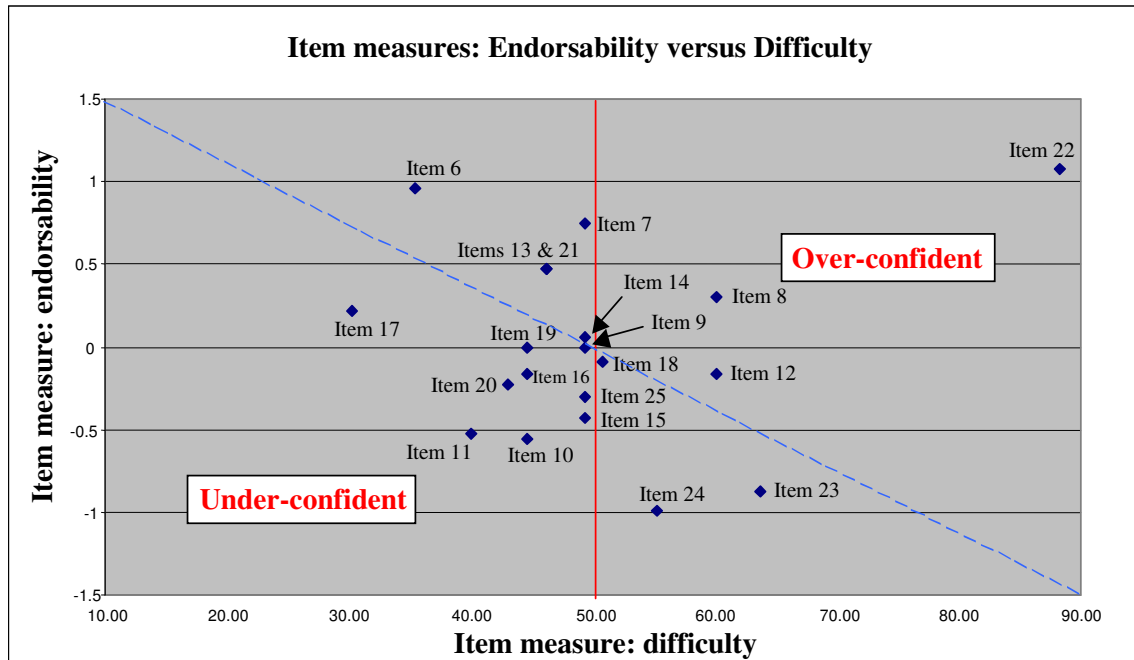


Figure 3. Scatter plot of item endorsability (item confidence) versus item difficulty measures

In order to evaluate whether the Hasan hypothesis is applicable to linear Rasch measures of performance and confidence the results presented in Table 4 and Figure 3 will now be compared with that reported for the analysis of free response explanations in Table 3. Item 22 stands out very clearly in Figure 3 as an item associated with strong misplaced confidence. Items 7 and 8 were associated with moderate over-confidence, followed by Items 6, 12, 13 and 21. On the other hand, Items 10, 11 and 24 are characterised by a larger measure of under-confidence than seven other items that appear in the under-confidence zone closer to the imaginary trend line. From the instructor's point of view both the positive and negative outliers in terms of confidence are of concern.

Without exception all of the items with a high prevalence of alternative conceptions (Items 22, 8, 7 and 13 in decreasing order) appear in the over-confident zone of Figure 3. Also present in this zone is the item with the highest prevalence of analytical inadequacy (Item 12) and Item 6 for which the combined total of unacceptable explanations is high (33%). The reasons for low confidence associated with Items 10, 11 and 24 were investigated. Items 10 and 11 are coupled quantitative problems based on the same setting of a mass suspended on a rope from an elevator

ceiling with another mass hanging from it. In general, respondents grasped the underlying concepts, but were not competent in the application of these principles in order to arrive at a quantitative answer. They were aware of their incompetence as evidenced by the low endorability of this item. Item 24, which deals with the trajectory of a moving object that leaves a circular path, was plagued by poor overall performance and absence of explanations, both indicative of a high degree of uncertainty about the principles involved.

Our test instrument included five items based on multiframe diagrams (Items 15, 16, 19, 20 and 21). Performance on all of these items was above average to reasonably good (58 – 70%). It can safely be assumed, in our opinion, that the typical physics student will not have encountered a ticker tape or a multiframe diagram except during formal instruction and will therefore not have formed strong alternative conceptions about the meaning or interpretation of these diagrams. Indeed, with the possible exception of Item 21, no inflation of confidence was observed for any of these items in Figure 3. No firm conclusions can be drawn for Item 21 because of ambiguities in the problem statement as stated above.

Item 22

Item 22 warrants further discussion, because it is clearly an outlier and was flagged as a problem item by all of the analyses discussed above. The very poor performance for Item 22 is associated with the highest average CRI values for both correct and incorrect responses in our dataset. The problem states that a block is pulled at constant speed along a rough surface and shows a diagram of the direction of the forces acting on the block (Figure 4). Item 22 was designed to test the ability to qualitatively analyse an equilibrium problem in terms of force components. However, the fact that the block was pulled at a constant speed seemed to distract students towards the ‘overcoming friction’ misconception. The correct option is C, and a correct explanation could be that the horizontal component of the applied force balances friction, so the applied force has to be larger than friction. The vertical component together with the normal force balances the weight, so the normal force has to be smaller than the weight. Only 6% of the students chose the correct option. The most popular distractor was D, chosen by 76% of the students.

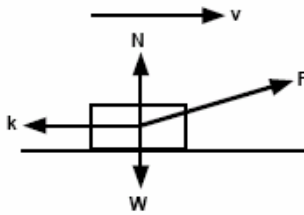
The frequency distribution for the multiple-choice component does not provide a clear picture of student understanding, because an incorrect answer could arise from either an alternative conception or from incorrect application of force components, or from a combination of the two. The most popular choice D was mostly explained by the coded explanation Q6, the ‘overcoming friction’ misconception, or Q3, which correctly applied the horizontal component, but overlooked the contribution of the vertical component of F. The inflated confidence on this difficult item can

primarily be attributed to the large incidence of the misconception that the applied force overcomes friction as reflected by explanations Q05 and Q06.

Item 22

A person pulls a block across a rough horizontal surface at a constant speed by applying a force F . The arrows in the diagram correctly indicate the directions, but not necessarily the magnitudes of the various forces on the block. Which of the following relations among the force magnitudes W , k , N and F must be true?

(15%) A. $F = k$ and $N = W$
 (3%) B. $F = k$ and $N > W$
 (6%) C. $F > k$ and $N < W$
 (76%) D. $F > k$ and $N = W$
 (0%) E. None of the above choices



Coded explanations:

Q01: No response (0%)
 Q02: Uncodable response (6%)
 Q03: The block does not move up or down, therefore $N = W$. The horizontal component of force F must be equal to force k ; therefore force F must be greater than force k . (24%)
 (Q04): When the box moves at constant speed, it means that all forces acting on it balance each other. (9%)
 Q05: The applied force must be greater than the frictional force, since motion is in the direction of a bigger force. The weight of an object is greater than the upward force by the surface on an object. (3%)
 Q06: The applied force must be greater than the frictional force, since motion is in the direction of a bigger force. The weight of an object is equal to the upward force by the surface on an object. (52%)
 (Q07): The sum of x components of forces is zero, and the sum of y components of forces is zero. (6%)

Figure 4. Item 22 with frequency distribution for multiple-choice answers and free response explanations for answers.

Discussion

We embarked on this study to investigate the capacity of the Hasan model for the accurate signalling of alternative conceptions in mechanics for a group of students. This model is based on the relationship between performance and confidence (certainty of response) as evidenced by responses to multiple-choice conceptual test items (Hasan et al., 1999). We used a test format where fixed response answers to physics questions were complemented by open response explanations for answers given so that these free responses could be used to verify whether strong alternative conceptions were indeed responsible for some of the wrong answers given to the multiple-choice questions. We also employed the method reported by Planinic et al. (2006) for the identification of the presence of alternative conceptions based on linear measures of performance and confidence as generated by the Rasch measurement model. The conclusions drawn from application of the Hasan

model to either raw score data or transformed measures of confidence and performance was subsequently checked against the results obtained from analysis of open responses in order to evaluate the success of each of these approaches.

The application of the Hasan model using average confidence levels associated with correct and incorrect answers was complicated by the fact that students in this cohort (and all of the others in the bigger project) displayed general over-confidence irrespective of the accuracy of their responses. This is evident from the average CRI values reported in Table 3. Hasan et al. (1999) used the numerical average of their Likert scale as the threshold to decide whether confidence was high or low. Item 16 is the only item with an average CRI value that is below the threshold value of 1.5 for the four-point scale used in our project. Since the choice of a threshold value of 1.5 would severely limit discrimination in our case, we decided to use the average confidence level obtained for the test as a whole (2.2, see Table 2) for the interpretation of the confidence levels reported in Table 3.

The Hasan model has clearly identified three problem items (Items 22, 12 and 8), but only two of these were affected by alternative conceptions. Similarly, only one of the two borderline cases in terms of the criteria used for the Hasan model (Items 7 and 14) was plagued by alternative conceptions. It did not flag Item 19 and Item 13, two items on which overall performance was reasonably good (67% and 64%, respectively), but where unacceptably high frequencies of incorrect explanations and alternative conceptions were documented. Based on these results, we conclude that inaccurate conclusions are drawn when the Hasan model is applied to raw score data. The items that were flagged were a subset of problem items plagued by either alternative conceptions or inadequate problem-solving skills and not a clean selection of either group.

The application of the Hasan hypothesis to Rasch measures, however, leads to a significantly more reliable identification of problem items. All of the items with high prevalence of alternative conceptions were flagged, even items where performance was acceptable or good. Analysis is greatly facilitated by the use of a scatter plot of item confidence measures (adjusted item endorsabilities) against item difficulty measures as shown in Figure 3. All of the outliers in the over-confidence zone are problem items and the majority of them are associated with strong alternative conceptions, the rest show serious deficiency in problem-solving skills. Outliers in the under-confident zone are associated with inadequate conceptual understanding and/or skills of application.

Alternative Conception or Lack of Skill?

The question remains of whether it is possible to use confidence-performance relationships to distinguish between over-confidence due to alternative conceptions or inadequate problem-solving skills. The design of our test instrument has enabled us to explore the application of the Hasan model to a wider selection of test items, namely to items specifically designed to reveal the presence of alternative conceptions (derived primarily from the Force Concept Inventory, Hestenes et al., 1992) and those aimed at probing the mastery of problem-solving skills and analytical thinking in mechanics (obtained from the Mechanics Baseline Test, Hestenes & Wells, 1992).

As reported in Table 3 the two items that were most severely affected by inadequacy of problem-solving skills are Items 12 and 14, followed by Items 19 and 22. Item 12 dealt with a car towing another vehicle twice its mass. It became evident from the open responses to this item that students understood the basic principle of the inverse relationship between acceleration and mass, but neglected to take the mass of the first car into account when calculating the acceleration with a second car in tow. Item 14 addressed the resultant forces acting on a crate that is being pulled on a rough floor by a man and a boy. The graphical representation used to illustrate the problem may have contributed to the flawed reasoning, because the ropes used by the man and the boy to pull the crate at different angles were of the same length in the drawing, thereby suggesting a vector diagram rather than a mere sketch of the problem setting.

Figure 3 indicates that the inflation of confidence because of a lack of skill is indeed significantly lower than that due to the presence of alternative conceptions. For example, the lack of skill reported for Item 12 (49%) did not inflate confidence nearly as much as the alternative conception reported for Item 22 (55%) despite comparable prevalence (Figure 3). A similar comparison can be made for item pair 13 and 14. This result makes intuitive sense. Students receive feedback on mistakes due to inaccuracy in analysis or graphical interpretation and they are more likely to accept the feedback as valid. Alternative conceptions, on the other hand, are much more resistant to change, presumably because they have intuitive value and are self-constructed. This difference in confidence inflation depending on its cause may prove to be useful in distinguishing between the two scenarios. In other words, the further removed outliers in the over-confidence zone are from the imaginary trend line in Figure 3, the more likely it is that answers to these items are based on strong alternative conceptions. Incorrect answers for items closer to the trend line in the over-confidence zone may result from analytical inaccuracy rather than from alternative conceptions in a mixed test similar to ours, but further study is required to determine whether this guideline is generally applicable.

Conclusions

The Hasan hypothesis is intuitively sound. The working definitions for misconceptions and alternative conceptions state that these are strongly held conceptions that are not in line with accepted scientific thought (Hasan et al. 1999; Nakhleh, 1992). The combination of high certainty of response with incorrect answers to conceptual test items, where distractors represent known alternative conceptions, can therefore be expected to signal the presence of such alternative conceptions. However, it is clear from the results discussed above that the extension of this thinking from individual responses to the collective responses of a group of students is problematic.

We have found that application of the Hasan model to raw score data for a group of students can at best be expected to flag a selection of the most important problem items, but whether the problem arises due to alternative conceptions, lack of analytical or graphical skills or ambiguity in the problem statement can not be determined with only multiple-choice data at your disposal. In addition, test items with above average performance that are associated with a significant presence of alternative conceptions will be missed completely. The model is also not simple to apply unambiguously, especially where general over-confidence is observed or in borderline cases where both confidence and performance on specific items is at the numerical average level for the cohort.

When the Hasan hypothesis is applied to data transformed according to the Rasch measurement model (Planinic et al., 2006), problem items are identified with much greater accuracy. Items were reliably flagged when associated with a significant prevalence of alternative conceptions, even in cases where performance on the item was good. Analysis was facilitated by the use of a specific graphical representation, i.e. a scatter plot of item confidence (adjusted item endorsabilities) against item difficulty measures. This approach can become a valuable tool for instructional design in mechanics. It also holds promise for the differentiation between over-confidence due to alternative conceptions or due to inadequate problem-solving skills.

References

- Andrich, D. (1988). *Rasch Models for Measurement*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-068. Newbury Park. Sage Publications, Inc.
- Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

- Bowen, G. M. & Roth, W.-M. (1999). Confidence in performance on science tests and student preparation strategies. *Research in Science Education*, 29, 209 – 226.
- Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, 50, 66 - 70.
- Duit, R., & Treagust, D.F. (2003). Conceptual change: a powerful framework for improving science teaching and learning. *International Journal of Science Education*, 25, 671 – 688.
- Dykstra, D.I. Jr., Boyle, C.F., & Monarch, I.A. (1992). Studying conceptual change in learning physics. *Science Education*, 76, 615–652.
- Galili, I. (1995). Mechanics background influences students' conceptions in electromagnetism. *International Journal of Science Education*, 17, 371-387.
- Gunstone, R.F., & White, R. (1981). Understanding gravity. *Science Education*, 65, 291 - 299.
- Halloun, I.A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American journal of Physics*, 53, 1043-1048.
- Hasan, S., Bagayoko, D., & Kelley, E.L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education*, 34, 294-299.
- Heller, P., & Finley, F. (1992). Variable uses of alternative conceptions: A case study in current electricity. *Journal of Research in Science Teaching*, 29, 259–75.
- Hestenes, D., & Halloun, I.A., (1995). Interpreting the Force Concept Inventory. *The Physics Teacher*, 33, 502-506.
- Hestenes, D., & Wells, M. (1992). A Mechanics Baseline Test. *The Physics Teacher*, 30, 159-166.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, 30, 141-158
- Jimoyiannis, A., & Komis, V. (2001). Computer simulation in physics teaching and learning: a case study on students' understanding of trajectory motion. *Computers & Education*, 36, 183-204.
- Knight, R.D. (1997). *Physics: A Contemporary Perspective*. Addison-Wesley, New York.
- Linacre, J. M. (2006). MINISTEP (WINSTEPS®) Rasch Measurement. Version 3.63.2. Retrieved 15 February, 2008, from www.winsteps.com.
- McDermott, L.C., & Redish, E.F. (1999). Resource letter: PER-1: Physics education research. *American Journal of Physics*, 67, 755 – 767.
- Nakhleh, M.B. (1992). Why some students don't learn chemistry. Chemical misconceptions. *Journal of Chemical Education*, 69, 191 – 196.
- Oliva, J. M. (1999). Structural patterns in students' conceptions in mechanics. *International Journal of Science Education*, 21, 903 – 920.

- Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, D.R. (2002). The Role of Individual Differences in the Accuracy of Confidence Judgments. *The Journal of General Psychology*, 129, 257-299.
- Planinic, M., Boone, W.J., Krsnik, R., & Beilfuss, M.L. (2006). Exploring alternative conceptions from Newtonian dynamics and simple DC circuits: links between item difficulty and item confidence. *Journal of Research in Science Teaching*, 40, 150 – 171.
- Potgieter, M., Rogan, J. M., & Howie, S. (2005). Chemical concepts inventory of Grade 12 learners and UP foundation year students. *African Journal for Research in Mathematics, Science and Technology Education*, 9, 121-134.
- Ramaila, S.M. (2000). The kinematic equations: An analysis of students' problem-solving skills. Master of Science Degree Dissertation. School of Science Education, University of Witwatersrand, South Africa.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institute.
- Reif, F., & Allen, S. (1992). Cognition for interpreting scientific concepts: A study of acceleration. *Cognition and Instruction*, 9, 1-44.
- Savinainen, A., & Scott, P. (2002). The FCI: a tool for monitoring student learning. Retrieved 3 July, 2008 from www.iop.org/Journals/PhysEd.
- Scott, P. H., Asoko, H. M., & Driver, R. H. (1991). Reprinted in *Connecting Research in Physics Education with Teacher Education* (1997), Andrée Tiberghien, E. L. Jossem, & J. Borgas (Eds.), International Commission on Physics Education.
- The Physics Classroom. Retrieved 29 May, 2005 from <http://www.physicsclassroom.com/Newtonlaws/html>.
- Yazdani, M. A. (2006). The exclusion of the students' dynamic misconceptions in college algebra: a paradigm of diagnosis and treatment. *Journal of Mathematical Sciences and Mathematics Education*, 1, 32 – 38.