Understanding the role of reasoning ability in mathematical achievement

Caren A Frosch (c.frosch@le.ac.uk)
School of Psychology, University of Leicester, LE1 9HN, UK

Victoria Simms (v.simms@ulster.ac.uk)
School of Psychology, Ulster University, BT52 1SA, UK

Abstract

A theoretical link between reasoning and mathematical ability has been supported by some recent empirical evidence. We argue that some of this evidence is indirect and measure selection may have influenced this relationship. We report one study in which mathematical ability was measured using a fluency and a calculation measure (Wood-cock Johnson-III, 2001) and reasoning ability was measured using the extended cognitive reflection test (Toplack, West & Stanovich, 2014) and a belief bias conditional reasoning task. Results from 68 undergraduate students suggest that mathematical ability is predicted by performance on the cognitive reflection test but not conditional reasoning performance. We discuss the implications of these findings for research on the link between mathematical ability and reasoning skills.

Keywords: reasoning; mathematical ability; CRT (cognitive reflection test); conditional reasoning

Introduction

Research has established that a variety of general cognitive skills are necessary for mathematical success, such as working memory, inhibitory control and shifting skills (Cragg & Gilmore, 2014). More recently it has been suggested that logical reasoning skills are an important aspect of good mathematical reasoning abilities. We report a study which examines this link between reasoning and mathematical ability.

Reasoning and mathematical ability

Recent research on reasoning and mathematical ability suggests a close relationship between these two skill sets. Ko and Knuth (2013) outline the type of reasoning skills used by mathematics majors, including informal deductive reasoning, example-based reasoning and further education in mathematics, compared to English, is associated with gains in reasoning ability, specifically in being able to reject invalid conclusions (Attridge & Inglis, 2013).

However, careful examination of some of the evidence suggests that the measures of reasoning ability are in fact frequently tapping into domain specific mathematical reasoning skills. For example, Nunes et al (2007) present evidence of a causal link between logical reasoning and mathematical ability in primary school children. However, logical competence was operationalized as the application of logical concepts in mathematics e.g. their understanding of the inverse relation between addition and subtraction, additive composition, one-to one and one-to-many correspondences, and seriation. Therefore, this specific study provides evidence for children's mathematical knowledge being based on their understanding of its underlying logic, but this is not the same as suggesting that logical reasoning skills per se are required for good performance on mathematical tasks. This research merely provides evidence for close relationships between domain specific reasoning skills.

An additional factor that should also be acknowledged is the impact of task selection when measuring mathematical ability. Numerous studies attempting to establish a link between reasoning and mathematical ability have not used gold-standard measures of mathematical ability, and instead have used self-report questionnaires, brief experimental maths tasks or school achievement measures (Gomez-Chacon, Garcia-Madruga, Vila, Elosua, & Rodriguez, 2014). These issues with measure choice are problematic in terms of validity of results and decreasing the possibility of comparing between studies.

Gomez-Chacon, et al. (2014) report that mathematics performance, as measured by mathematics scores at the end of a secondary school mathematics course, is correlated with high cognitive reflection (as measured by the Cognitive Reflection Test, Frederick, 2005) and overall reasoning performance (as measured by a battery of reasoning tasks, including propositional reasoning, syllogistic reasoning and probabilistic reasoning). However, there is little information about the tests used to assess mathematical ability and it is unclear how the individual reasoning tasks are related to mathematical performance.

Handley, Capon, Beveridge, Dennis and Evans (2004) established a relationship between teacher-reported numeracy levels, a standardized mathematics measure and reasoning skills in a small sub-sample (N=32) of 10 - 11 year old children. They used relational and conditional reasoning problems in which they varied the believability of the conclusions. Several indexes relating to logic and believability of the problems were highly correlated with numeracy skills. However, again it is unclear whether reasoning performance is related to specific numeracy skills.

Vamvakoussi, van Dooren, and Verschaffel (2012) present evidence of intuitive reasoning in processing mathematical problems which are incongruent with natural number rules. The authors argue that this points to a distinction between system 1 and system 2 processing within mathematical reasoning. However, it is unclear from this research whether the skills demonstrated by participants in these studies are domain specific or whether they draw on domain general reasoning processes.

A recent study by Morsanyi, Devine, Nobes and Szucs (2013) attempts to shed some light on the question of whether mathematical achievement is related to the ability to engage in system 2 type analytical reasoning. Children with developmental dyscalculia were compared to typically developing children as well as children with outstanding mathematical abilities. While the authors report significant differences between the three groups it is important to note that the measure of reasoning performance (transitive inferences with believable and unbelievable premises and conclusions) was based on quantitative relations and hence it is not surprising that the children with developmental dyscalculia had difficulties with this task.

Some evidence suggests that deductive and mathematical processing draw on distinct neural substrates which implies that the two skills may be drawing on different cognitive systems (Kroger, Nystrom, Cohen & Johnson-Laird, 2008). Based on our assessment of the current evidence, we conclude that the question of whether mathematical ability is related to domain general reasoning abilities has not been answered inconclusively.

Cognitive reflection test (CRT)

One measure that has recently attracted a lot of attention, both within the reasoning community but also as a measure of reasoning abilities in relation to mathematical abilities is the Cognitive reflection test – CRT (Frederick, 2005). The original CRT consists of three problems such as:

A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost?

The intuitive, but incorrect, answer is 10 cents whereas the correct answer is 5 cents. These problems are meant to tap into a person's tendency to either engage in heuristic Type I processing or analytical Type II processing.

However, the CRT has been criticized, as all of the problems also have a large mathematical component to them. For example, Welsh, Burns, and Delfabbro (2013) report a factor analysis of a series of decision style measures, cognitive tasks and the CRT which revealed that the CRT primarily loaded onto a factor with numerical ability measures. Hence, the authors conclude that the CRT is a primarily numerical measure as it only has predictive value for reasoning tasks for which numerical skill is required to reach the correct answer. Although, others have concluded that the CRT is not 'just another numeracy scale' (e.g., Liberali, Reyna, Furlan, Stein, & Pardo, 2012). However, the CRT has also been criticized as it does not generalize to reasoning tasks which do not contain an arithmetic component, such as belief bias and matching bias syllogism tasks (Stupple, Gale & Richmond, 2013). Therefore, it is difficult to establish the precise relationship between mathematical ability and reasoning skills due to the content overlap in measures.

A better measure of reasoning ability

A well-established deductive reasoning task is the conditional reasoning task. Participants are typically presented with a conditional premise, such as:

If the weather gets warmer then more people will go to the beach.

followed by a minor premise such as:
The weather gets warmer.

Participants are then either asked to make an inference or to select an inference from a set of options such as:

More people go to the beach
More people do not go to the beach
More people may or may not go to the beach

Four different inferences can be made from a conditional by either affirming or denying either the first part (the antecedent) or the second part of the conditional (the consequent) in the minor premise that follows the conditional. Affirming the first part of the conditional, as in the example above, results in the modus ponens inference of 'more people go to the beach', which, according to propositional logic, is a valid inference. Similarly, denying the consequent (e.g. not more people go to the beach) results in the equally valid modus tollens inference (e.g. the weather does not get warmer). The other two inferences, denial of the antecedent (e.g., the weather does not get warmer) and affirmation of the consequent (e.g., more people go to the beach) are not considered valid inferences, hence, one should conclude for both inferences that the proposed conclusion may or may not occur.

The valid modus ponens inference is fairly easy, with endorsement rates ranging between 89% and 100%. There is considerable variability in endorsement rates of the valid modus tollens and the invalid inferences (denial of the antecedent and affirmation of the consequent), particularly when the content of the materials is arbitrary (Evans, Newstead, & Byrne, 1993).

Manipulating the believability of the initial conditional has an impact on the rate at which participants endorse all of the four inferences (e.g., Evans, Handley, & Bacon, 2009). When faced with an unbelievable conditional such as,

If fast food is taxed then childhood obesity will increase.

Followed by a minor premise such as
Fast food is taxed

A participant has to deal with the conflict between what is logically correct (childhood obesity increases) and what is believable (childhood obesity does not increase). In this sense, this task is similar to the cognitive reflection task as a person must overcome their intuitive response and engage in analytical thinking in order to arrive at the correct response. The advantage of this task over the cognitive reflection task

is that it does not require any mathematical processing, but like the cognitive reflection task it requires the suppression of an intuitively appealing response. We felt it was important to retain this aspect of the reasoning measure in order to assess whether it is this ability to detect a conflict between an intuitive and a correct response which is driving the relationship between reasoning abilities and mathematical abilities.

Being sensitive to conflict in stimuli could be potentially useful for mathematical achievement. Geary, Bailey & Hoard (2009) established that children who were able to detect errors in a puppet's counting procedure had higher mathematical achievement. Conflict detection could assist individuals to correctly establish the veracity of basic arithmetic problems, essential for self-monitoring and correct of performance (Campbell & Fugelsang, 2001).

The present study thus seeks to examine the relationship between mathematical and reasoning ability by employing gold-standard measures of mathematical ability and a reasoning task which does not rely on mathematical ability. However, we also include the CRT in order to examine its relationship with the mathematical measures and reasoning task we employ.

In order to improve validity, the study will utilize standardized measures of mathematical achievement, the Wood-cock Johnson-III Math fluency test and the Wood-cock Johnson-III calculation test (2001). The Math Fluency measure requires participants to complete as many simple arithmetic problems as possible in a set period of time, in contrast the un-timed calculation test requires more complex mathematical reasoning skills. We would anticipate a stronger association between performance on the general reasoning tasks and the calculation test, than with the fluency test due to the overlap in reasoning content between the two measures.

We therefore predict that mathematical ability as measured by the calculation task in particular will be related to performance on the cognitive reflection task and conditional reasoning task and that after taking mathematical fluency into consideration, the reasoning tasks will contribute to the prediction of mathematical ability.

Study

Method

Participants Seventy-four University of Leicester Psychology students participated in this study in return for course credit**.** However, six participants were removed prior to analysis as they had failed to follow the instructions for the Woodcock Johnson Mathematical Fluency task (they completed the calculations column by column rather than row by row, thus affecting the difficulty of the problems solved). The remaining 68 students had a mean age of 19.57 years (range 18-46 years) and there were 13 male and 55 female participants. All participants had a minimum grade C or equivalent GCSE in mathematics.

Materials and Design The study had a correlational design and all participants completed the same four measures. There were two measures of mathematical ability the Wood-cock Johnson-III Math fluency test and the Wood-cock Johnson-III calculation test (2001). The fluency test is a measure of people's ability to solve simple arithmetic calculations (e.g., 4 - 2) within a 3 minute time period. It includes 160 simple addition, subtraction and multiplication problems. The calculation test is a 45-item test used to assess more complex mathematical ability. The questions start with basic maths calculations and get progressively harder, reaching A-level (final year of secondary school) standard. Participants in this study started with item 14 on the calculation task due to the age of the sample.

Reasoning abilities were assessed using the extended Cognitive Reflection Test (Toplack, West & Stanovich, 2014) which consists of 7 items. However, we omitted the last of the seven items as it necessitated a multiple choice response option, which the other items did not (see Appendix for the full set of items included). The six items were each presented on a separate page in 8 different counterbalanced orders.

The second reasoning measure was a belief bias conditional reasoning task, for which the materials were provided by Roser (2012-2015). Participants were presented with 32 conditionals together with a minor premise and three response options. For example:

Given
If oil prices continue to rise then UK petrol prices will rise.

Suppose
Oil prices continue to rise

Does it follow that
☐ UK petrol prices rise
☐ UK petrol prices do not rise
☐ UK petrol prices may or may not rise

Each problem was presented on a different page of the 32 page booklet. Half of the conditionals had believable content and the other half had unbelievable content, that is participants were presented with conditionals with 32 distinct contents. Within each set of believable and unbelievable conditionals participants were presented with four modus ponens, four modus tollens, four denial of the antecedent and four affirmation of the consequent inferences.[1] The 32 conditional inferences were presented in different random orders to each participant.

Procedure Participants were tested in groups of up to 12 people. Participants were advised to work in exam conditions, meaning in silence, independently, and without

[1] Due to an administrative error four participants did not see the full set of 32 conditionals, two participants only received 3 instead of 4 inferences of one type and two participants were not given one set of inferences, e.g. they did not receive the four believable affirmation of the consequent inferences.

the use of calculators. The four tasks were completed in four different orders. Participants either started with the fluency measure or with the CRT and after completing the first two tasks either completed the calculation task or the conditional reasoning task. The fluency task was timed and participants had to complete as many items as possible in three minutes. All other tasks were completed at the participants' own pace. When completing the calculation task, as per standardized administration rules, the experimenters advised the participants to move on to the next task (or finish if this was their last task) when the experimenters noted that they had 6 consecutive incorrect answers. The test battery took approximately 40 – 45 minutes in total to complete, with a small minority of participants completing it in 30 minutes or 60 minutes.

Instructions for the conditional reasoning task:

This booklet contains 32 statements. The statements and associated tasks are about how people think in their daily lives and are not tests of intelligence. Please read each statement carefully and decide which conclusion follows. Please answer the questions in the order in which they are presented and do not try to change your answers once you have written them.

Instructions for the CRT:

This booklet contains six items that vary in difficulty. Answer as many as you can.

For the two maths tasks, participants received the standardised verbal instructions.

Results

Table 1 displays the mean scores on the four measures the participants completed. Participants scored a mean of 103.53 out of a possible 160 on the fluency measure and a mean of 15.26 out of a possible 32 on the calculation task. Performance on the cognitive reflection task was fairly poor with a mean number of correct responses of 1.18 out of 6 and as can be seen from Table 1 the majority of incorrect responses were the intuitive ones (mean 3.35).

Table 1: Mean Scores (standard deviations in parenthesis) for the four measures.

Task	Mean (SD)
Fluency	103.53 (23.24)
Calculation	15.26 (5.44)
CRT correct	1.18 (1.44)
CRT intuitive	3.35 (1.55)
CRT incorrect	1.47 (1.06)
Conditionals (proportion logically consistent inference)	.48 (.10)

For the conditional reasoning task we calculated a score of the number of logically consistent inferences (i.e., we counted the number of endorsements of the modus pones and modus tollens inferences, and the number of times participants did not endorse the affirmation of the consequent and denial of the antecedent inferences) and converted these scores into proportions due to the fact that a small number of participants had not received the same number of conditional inferences (see footnote 1). This score gave us a logic index which we could use for our analyses. However, we also present the responses to the conditional reasoning task in the customary way of endorsement rates for each inference in Figure 1. A 2x4 ANOVA on the endorsement rates confirmed that the believability manipulation of the conditional reasoning task had been successful. There was a main effect of believability; $F(1,65) = 26$, $p < .001$, $n^2 = .286$ indicating that the participants endorsed the believable inferences more than the unbelievable inferences. There was also a main effect of inference, $F(3,195) = 7.35$, $p < .001$, $n^2 = .102$, with participants endorsing the modus ponens inference more than any of the other inferences.

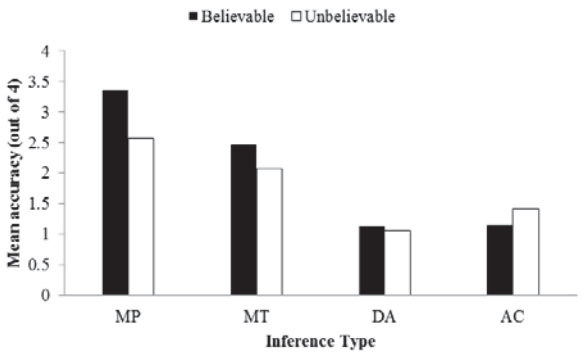

Figure 1. Mean accuracy rate (according to formal logic) for each of the four inferences by believable and unbelievable condition.

As can be seen from Table 2, the fluency measure only correlated with the calculation measure, $r = .49$, N = 68, $p < .001$. The calculation measure was correlated most strongly with the number of correct responses on the CRT, $r = .36$, N = 68, $p = .003$. However, the proportion correct responses on the conditional reasoning task was also moderately correlated with the calculation task, $r = .24$, N = 68, $p = .047$. But an examination of the correlations between the calculation task and the different conditional inferences revealed no significant correlations; rs< .12, ps > .16. The CRT and the conditionals task were also moderately correlated, $r = .271$, N = 68, $p = .026$. Interestingly, this correlation between CRT and the conditionals task is driven by a correlation between correct inferences on the believable conditionals, r = .257, N = 68, p = .034 and not by a correlation between unbelievable inferences and CRT, $r = .154$, $p = .21$.

Table 2. Correlations between the four measures

	1	2	3
1.Fluency			
2. Calculation	.487**		
3. CRT (correct)	.182	.357**	
4. Proportion Correct Conditionals	.122	.242*	.271*

Examination of the correlations between the CRT and the different conditional inferences suggested that there were no (for modus ponens inferences) or negative (for modus tollens inferences) correlations between CRT performance and the valid inferences and there were positive correlations between CRT and the invalid inferences (affirmation of the consequent and denial of the antecedent). Hence we examined whether participants who made the correct inference for the invalid inferences (i.e. chose the option that the conclusion may or may not be drawn) performed better on the CRT. Indeed there was a positive correlation between the CRT and the correct inferences for the invalid inferences, $r = .377$, N= 68 $p = .002$.

A hierarchical regression with fluency entered at step 1 and CRT correct and proportion correct on the conditionals task to predict calculation scores indicates that fluency accounted for 24% of the variance in calculation scores ($F(471.332, 22.938) = 20.55$, $p < .001$) and that the addition of the two reasoning measures added .089 to R^2, thus accounting for 33% of the variance in calculation scores (the R^2 change was significant, $F = 4.209$, $p = .019$). Inspection of the beta weights for the final model revealed that fluency was the best predictor; beta = .428 (t=4.085, $p < .001$) and CRT correct also added significantly to the prediction; beta = .246 (t =2.28, p = .026), but performance on the conditional reasoning task was not a significant predictor.

Discussion

The key finding from this study is that while conditional reasoning performance is correlated with mathematical ability as measured by the Wood-cock Johnson-III calculation test (2001), it does not predict performance on this task when mathematical fluency is taken into account. The cognitive reflection task however does predict performance on the calculation task. Hence, this study provides further support for the idea that the relationship between the CRT and mathematical achievement is driven by the mathematical component of the task.

We acknowledge a potential limitation of our study is that that mathematical education levels were not recorded for our participants, although we do know that all our participants had at least a grade C or equivalent in Mathematics at GCSE level. Furthermore, the measures of mathematical ability included in this study are sensitive standardized measures and so we are confident that these measures give us a fair reflection of the participants' mathematical abilities.

In terms of the relationship between mathematical ability and reasoning ability, we argue that future studies examining this relationship must take into consideration whether they are investigating a relationship between domain specific (mathematical) reasoning skills or more domain general reasoning skills. We acknowledge that our domain general reasoning measure relied heavily on verbal abilities and so we propose that future studies should also include a non-verbal reasoning measure in order to assess whether abstract reasoning skills are linked to mathematical calculation skills. It is thus still unclear whether reasoning per se is essential for mathematical achievement or whether these are independent cognitive skills as suggested by Kroger et al (2008) who reported that different brain areas are linked to mathematical and reasoning processes. An alternative hypothesis worth pursing is whether mathematical and reasoning abilities are linked due to their relationship with more domain-general characteristics of intelligence.

Another interesting finding from this study is the fact the CRT was only correlated with good performance on the believable conditionals and not with performance on the unbelievable conditionals. Arguably, good performance on the CRT requires a person to recognize a conflict when considering the intuitive answer which results in a more considered response. Similarly, good performance on the unbelievable conditionals requires a person to recognize the conflict between believability and logic before they can recognize the correct inference. The fact that this aspect of the conditionals task was not related to CRT performance suggests that further investigation into what the CRT is measuring is also warranted.

Acknowledgments

We thank Matt Roser for supplying the materials for the conditional reasoning task and Priscilla Baafi, Sach Dhesi and Mesel Teklebrhan for their assistance in data collection.

References

Attridge, N. & Inglis, M., (2013). Advanced mathematical study and the development of conditional reasoning skills. *PLoS ONE*, 8, e69399.

Campbell, J., & Fugelsang, J. (2001). Strategy choice for arithmetic verification: Effects of numerical surface form. *Cognition, 5*, 1-39.

Cragg, L. & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function skills in the development of mathematics proficiency. *Trends in Neuroscience and Education, 3,* 63-68.

Evans, J. S. B. T., Handley, S. J., & Bacon, A. M. (2009). Reasoning under time pressure. *Experimental Psychology, 56*, 77-83.

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). Human reasoning : the psychology of deduction. Hove: Lawrence Erlbaum.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19*, 24-42.

Geary, D., Bailey, D., & Hoard, M. (2009). Predicting Mathematical Achievement and Mathematical Learning Disability With a Simple Screening Tool. *Journal of Psychoeducational Assessment, 27*, 265-279.

Gomez-Chacon, I. M., Garcia-Madruga, J. A., Vila, J. O., Elosua, M. R., & Rodriguez, R. (2014). The dual process hypothesis in mathematics performance: Beliefs, cognitive reflection, working memory and reasoning. *Learning and Individual Differences, 29*, 67-73.

Handley, S. J., Capon, A., Beveridge, M., Dennis, I., Evans, J. St. B.T. (2004). Working memory, inhibitory control and the development of children's reasoning. *Thinking and Reasoning, 10*, 175-195.

Ko, Y. & Knuth, E. J. (2013). Validating proofs and counterexamples across content domains: Practice of importantce for mathematics majors. *The Journal of Mathematical Behavior, 32,* 20-35.

Kroger, J., Nystrom, L. E., Cohen, J. D., & Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Research, 1243*, 86-103.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making, 25*, 361-381.

Morsanyi, K., Devine, A., Nobes, A., & Szucs, D. (2013). The link between logic, mathematics and imagination: evidence from children with developmental dyscalculia and mathematically gifted children. *Developmental Science, 16,* 542-553.

Nunes, T., Bryant, P., Evans, D., Bell, D., Gardner, S., Gardner, A., & Carraher, J. (2007). The contribution of logical reasoning to the learning of mathematics in primary school. *British Journal of Developmental Psychology, 25*, 147-166.

Roser, M. (2012-1015). Dual processes in reasoning: A neuropsychological study of the role of working memory. ESRC grant RES-062-23-3285.

Stupple, E. J. N., Gale, M. & Richmond, C. (2013). Working Memory, Cognitive Miserliness and Logic as Predictors of Performance on the Cognitive Reflection Test In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1396-1401). Austin, TX: Cognitive Science Society.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*, 147-168.

Vamvakoussi, X., van Dooren, W., & Verschaffel, L. (2012). Naturally biased? In search for reaction time evidence for a natural number bias in adults. *The Journal of Mathematical Behavior, 31*, 344-355.

Welsh, M., Burns, N. & Delfabbro, P. (2013). The Cognitive Reflection Test: how much more than Numerical Ability? In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1587-1592). Austin, TX: Cognitive Science Society.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson Tests of Achievement (WJ III). Riverside Publishing.

Appendix

Examples of conditional reasoning task

Believable Modus Ponens
If car ownership increases then traffic congestion will get worse
Car ownership increases

Believable Denial of the Antecedent
If jungle deforestation continues then gorillas will become extinct
Jungle deforestation does not continue

Unbelievable Modus Tollens
If fast food is taxed then childhood obesity will increase
Childhood obesity does not increase

Unbelievable Affirmation of the Consequent
If the lottery prize-money increases then fewer people will buy tickets
Fewer people buy tickets

Additional items for extended CRT

(1) If John can drink one barrel of water in 6 days, and Mary can drink one barrel of water in 12 days, how long would it take them to drink one barrel of water together? _____ days [correct answer = 4 days; intuitive answer = 9]

(2) Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class? _____ students [correct answer = 29 students; intuitive answer = 30]

(3) A man buys a pig for $60, sells it for $70, buys it back for $80, and sells it finally for $90. How much has he made? _____ dollars [correct answer = $20; intuitive answer = $10]