Understanding Performance in Test Taking: The Role of Question Difficulty Order

Lina Anaya[1] **Nagore Iriberri**[2] **Pedro Rey-Biel**[3] **Gema Zamarro**[4]

April 2021

Abstract: Standardized assessments are widely used to determine access to educational resources with important consequences for later economic outcomes in life. However, many design features of the tests themselves may lead to psychological reactions influencing performance. In particular, the level of difficulty of the earlier questions in a test may affect performance in later questions. How should we order test questions according to their level of difficulty such that test performance offers an accurate assessment of the test taker's aptitudes and knowledge? We conduct a field experiment with about 19,000 participants in collaboration with an online teaching platform where we randomly assign participants to different orders of difficulty and we find that ordering the questions from easiest to most difficult yields the lowest probability to abandon the test, as well as the highest number of correct answers. Consistent results are found exploiting the random variation of difficulty across test booklets in the Programme for International Student Assessment (PISA), a triannual international test, for the years of 2009, 2012, and 2015, providing additional external validity. We conclude that the order of the difficulty of the questions in tests should be considered carefully, in particular when comparing performance between test-takers who have faced different order of questions.

Keywords: Question order, Difficulty, Test performance

JEL Codes: C93, D81, I20, J16.

[1] University of Arkansas, United States. Email: lanaya@uark.edu
[2] University of the Basque country and Ikerbasque, Spain. Email: nagore.iriberri@gmail.com
[3] ESADE, Ramon Lull University, Spain. Email: pedro.rey@esade.edu. Pedro Rey acknowledges funding from Ministerio de Ciencia e Innovación (PID2019-107108GB-I00),
[4] University of Arkansas, United States. Email: gzamarro@uark.edu

1. Introduction

Standardized assessments are one of the most frequently used ways to measure individuals' knowledge and aptitudes. On top of their regular academic use, performance on some of these tests plays a crucial role in shaping educational and labor market outcomes. For example, the Scholastic Aptitude Test (SAT) in the US, and the Graduate Record Examination (GRE) worldwide, are assessment tests that determine access to universities and graduate studies. Similarly, standardized assessments are used to license professionals, including among other professionals in Medicine and Law. Some examples are the bar exams in Law, the *United States Medical Licensing Examination* (USMLE), or *Medico Interno Residente* (MIR) in Spain. Additionally, standardized tests are used to compare education systems across countries, being the *Program for International Student Assessment* (PISA) the most prominent worldwide example.

The assessment of knowledge and aptitudes in tests depends on multiple design features that may potentially influence performance. These potential effects must be considered in light of the objectives and goals of the tests. For example, a topic that has received quite a lot of attention, both in education and in economics, is whether incorrect answers should be scored differently than omitted questions in multiple-choice tests.[5] In addition, in psychology, performance differences between essay type tests vs. multiple-choice tests have been studied (see for example the works by Bridgeman et al., 1996 and Chamorro-Premuzic, 2006), as well as potential effects of time limits, which may put more or less pressure on test-takers (see, for example, the reviews by McDonald, 2001, and Zeidner, 2010).

A topic that has received less attention in Education and Economics is to what extent question difficulty order within a test influences performance. A perfectly rational test taker should be immune to the actual presentation of the order of questions of different levels of difficulty. However, there exists ample evidence from Behavioral Economics showing that humans are affected by presentation effects (Tversky and Kahneman, 1981). For example, test fatigue

[5] Early observational studies include Swineford (1941), Anderson (1989), Atkins et al. (1991), Ramos and Lambating (1996), Tannenbaum (2012), Pekarinen (2014), Akyol, Key, and Krishna (2016) and Riener and Wagner (2017). Only recently there have been important advances in establishing causality by conducting randomized controlled trials in the laboratory (Baldiga, 2014), in the field (Ben-Shakhar and Sinai, 1991, Espinosa and Gardeazábal, 2013, and Funk and Perrone, 2016, Karle et al., 2019, Atwater and Saygin, 2020, and Iriberri and Rey-Biel, 2021) and using before-after quasi-controlled studies (Coffman and Klinowski, 2020).

increases, and performance decreases as the test progresses (Borghans & Schils, 2012; Zamarro et al., 2019) and so, it may be better to place harder questions at the start of the test given that they require greater analytical ability and effort. However, students might be disheartened by seeing a hard question early in the test, as a signal of the general difficulty of the rest of the test. In fact, exams like the GRE adapt the level of difficulty of subsequent questions to performance in the previous ones to obtain an accurate assessment of the test-takers knowledge.

In addition, potential psychological effects stemming from the order of difficult questions are particularly important in tests in which participants cannot access later questions until having answered the previous ones. A common practice by many academics is to simply randomize the order of questions or order questions in the order in which particular topics were explained during a course. However, these practices do not necessarily take into account potential psychological effects, if they do not explicitly incorporate the potential concern in the test design. In summary, if test designers have the goal of accurately measuring knowledge and not performance as affected by psychological effects, how should questions be ordered according to their difficulty levels? From easiest to most difficult, from most difficult to easiest, something in between? Does the difficulty order matter at all?

This paper studies whether the order of difficult questions is an important feature to take into account when designing tests. In particular, we present two studies that complement each other in internal and external validity by showing how the order of difficulty of questions affects performance and ultimately, may bias the assessment of knowledge and aptitudes. Our first study is a field experiment designed in collaboration with an online teaching platform, which publicly promoted a 10-question online math challenge among their potential clients. The 19,000 voluntary participants were randomized into four different treatments, all with the same ten questions but with variation in the order of presentation of questions. The two main treatments contained either the three more difficult questions at the beginning (or at the end), while the four questions in the middle were the same in all treatments. The two additional treatments combined difficult and easy questions throughout the assessment. We measured participants' number of completed questions, their number of correct responses, as well as self-reported demographic and self-perceived performance.

Our second study employs data from three years of the PISA exam (2009, 2012, and 2015) with a total of more than six hundred thousand student participants from around the world. We exploit the fact that test booklets vary in the order of groups of questions and that the allocation of test booklets to students in PISA is done randomly to estimate the effect of the order of difficulty of groups of questions on the number of blank responses and performance throughout the assessment.

In the field experiment, we find that ordering the questions from easiest to most difficult yields the lowest probability to abandon the test as well as the highest number of correct answers. In particular, among participants who take the test with difficult questions placed first, 44% of them do not complete the test, while only 30% of the participants who take the test with easy questions placed first drop it. Accordingly, participants answer on average more than one more correct question in the Easy-Difficult than in the Difficult-Easy treatment (3.53 vs 2.42 out of 10 questions). In addition, consistent with the findings of the field experiment, in PISA, we find that students coming from a part of the test that is about 10 percentage points more difficult, leave between 0.2 and 2 percentage points more questions blank and obtain between 0.4 and 4 percentage points fewer correct answers in the current part of the test.

Most of the available literature on the effect of question difficulty order comes from psychology with studies that involve mostly undergraduates and small convenience samples. Early studies found mixed effects. While some studies found no effect of question difficulty order on performance (Brenner, 1964; Smouse and Mung, 1968; Plake, Thompson, and Lowry, 1981; Gerow, 1980; Laffitte, 1984), others, consistent with our results, found that students perform better in tests where they have the most difficult questions at the end of the assessment test, as opposed to the beginning (Skinner, 1999; Hambleton and Traub, 1974).

The more recent literature in psychology finds overall null results of question difficulty order on performance, but positive results on self-perceived performance when the easier questions are at the beginning of the assessment. Weinstein and Roediger (2012) create two versions of a general knowledge assessment using the same 100 questions but changing their order. The first version has the most difficult questions at the beginning and the easiest questions at the end of the test (difficult-easy). In contrast, the second version starts with the easiest questions and ends with the most difficult questions (easy-difficult). The authors randomly assign either version of the

assessment among a sample of 50 college students. They find no statistically significant differences in performance between the two test versions as a function of question order. However, the participants who receive the easy-difficult test are more optimistic about their performance relative to their peers who receive the difficult-easy version. Another study from Bard and Weinstein (2017) follows a similar test design and obtains similar results. The authors employ a sample of about 270 college students and implement three experiments in which the order of questions, the type of test, and whether or not students are allowed to review the questions before answering them varies from one experiment to another. They find that there are no differences in performance as a function of question difficulty order. In addition, the authors find that students tend to be more optimistic about their performance when the easy questions are at the beginning of the test. One of the key limitations of all these studies in the psychology literature is their small sample sizes, which hinders power and external validity.

Our two studies contribute to the literature by studying the effect of question difficulty order using two large subject samples in field settings. A common feature of both studies is that performance is not explicitly incentivized. Hence, in principle, issues such as risk aversion, competitive pressure, or importance attributed to rewards should not influence answering strategies, and therefore, assessment performance. Study 1 allows for random treatment variation and control over the test design, which is crucial for identifying causality and therefore to guarantee the internal validity of our results. In study 2, using PISA data, we have less control over the test structure, even though the order of groups of questions varies by test booklets and booklets are randomly assigned to students, but it provides higher external validity as it uses a large representative sample of international students taking a relevant exam used for the comparison of educational systems in different countries.

The paper has the following structure. Section 2 describes and presents the results of our field experiment while section 3 focuses on the PISA study. Section 4 concludes.

2. Study 1: Field Experiment with *Smartick*

2.1. Experimental Design and Randomization

In collaboration with *Smartick*, an online platform aimed to teach math to students in Spain and Latin America, we launched a viral mathematical challenge consisting of 10 math questions.[6] The main criteria to choose the 10 questions was variation in their levels of difficulty, according to the past performance of a large number of students.[7] Please, see Appendix A for an English translation of the 10 questions.

We picked 3 "easy" questions (E), in which the percent of correct answers was above 80%, 4 "medium-difficulty" questions, with roughly 50% of correct answers, and 3 "difficult" questions (D), in which the percent of correct answers was below 25%. Our treatment design kept the medium-difficulty questions in the middle block and randomly varied the order in which the easy and difficult questions came before or after the medium-difficulty questions. We designed two main treatments, "Easy-Difficult" and "Difficult-Easy", alternating whether the block of the same 3 easy questions was presented before or after the medium difficulty block. Two additional treatments, "2Easy-1Difficult" and "1Easy-2Difficult", mixed two easy questions and one difficult question or two difficult and one easy, respectively, in the first three-question block of the test. Figure 1 shows the main features of the experimental design.

[Figure 1, here]

Smartick launched the challenge on July 2nd, 2019 through a message on social media. The message contained a direct link to the test, which automatically randomized those who accessed it to one of our four treatments. Two weeks later, on July 19th, *Smartick* published the challenge on their blog and sent a press release which was published by more than 70 media outlets. See Appendix B for an English translation of different announcements in social media.

[6] More information regarding the online platform can be found at https://www.smartick.es/. The math challenge can be found at https://retosmatematicos.smartick.es/.

[7] The questions were selected from the final stage of the 2016, 2017, and 2018 editions of *Concurso de Primavera de Matemáticas*, a regional math contest organized by the Math Department of the University Complutense of Madrid. In every edition, around 40,000 students participate in a first stage and about 3,000 reach the second and final stage of the math contest. Students from primary education, secondary education and High School participate in this regional contest (see Iriberri and Rey-Biel 2019 and 2021 for a description of the math contest).

Our sample consists of the 18,952 individuals who participated in the challenge before general results from the challenge went public in November 2019.[8] Participants were randomly assigned to the four different treatments, as shown in Figure 1. Before starting the test, participants revealed their gender, age group (according to eight categories, ranging from 4-11 years old to more than 60 years old), and educational attainment (five categories from primary schooling to postgraduate). No feedback about the individual performance was provided until all 10 questions had been answered. Participants could drop out from the assessment at any time, but we kept records of their performance until the time of abandoning the test. For each question, we recorded whether it was answered and whether the answer was correct or not. At the end of the test, when all 10 questions were completed, participants were asked to guess how many questions they believed they got right. Finally, the application provided participants with feedback on their number of correct answers and the average performance of participants in their treatment group.

2.2. Assessing the Design

We first check that we identified the difficulty of the questions properly. Table A1 in the Appendix shows the mean values of correct answers for each of the questions, ordered from the easiest to the most difficult. As shown by the mean values of correct responses, we were successful in identifying the difficulty of the questions correctly. The easiest (E1-E3) show the highest proportion of correct answers and the most difficult ones (D1-D3) the lowest, both when aggregating across all four treatments (columns 1-3) and when only focusing on the two main treatments in which the block of easy questions, or the block of difficult ones, were answered first (columns 4-6).

[Table 1. Randomization, over here]

We next check if the randomization went correctly. Table 1 shows the mean of the gender, age, and education variables overall and by treatment. About two-thirds are male participants and young participants are overrepresented compared to the general population, which was expected in a platform targeted for teaching Math to children. Accordingly, most participants are completing or have completed their secondary, high school, or undergraduate studies. Regression analysis

[8] We cannot fully distinguish whether the same individual participated more than once in the challenge, but anecdotal evidence provided from records in the servers shows that this behavior was very unusal.

from an ordered logit regression, included in Table A2 in the Appendix, shows that none of the control variables are statistically significant in explaining the assignment to the four treatments. Therefore, the randomization was successfully implemented.

2.3. Description of the Results

Table 2 includes the mean values of the main outcome variables by treatment. Our main analysis focuses on the overall sample of 18,952 participants. Furthermore, we will also describe the results for three interesting subsets of participants: Sample I includes all participants (14,433) who answered the first block of questions (questions 1 to 3); Sample II includes those participants (12, 719) who finished the first and medium blocks of questions (questions 1 to 7). Finally, sample III includes only participants (12,139) who completed the whole test.

[Table 2 over here]

Dropped takes value 1 when a participant does not complete the test. Equivalently, *Dropped_Q4_Q7* and *Dropped_Q8_Q10* take value 1 when participants abandon the test between the referenced questions. *Total Answers* measures the number of provided answers (equivalently for *Total Answers_Q4_Q7* and *Total Answers_Q8_Q10*). *Total Correct* measures the number of total correct answers provided (equivalently for *Total Correct_Q4_Q7)*. *Guessed Correct* measures the total number of correct answers the participant expects, while *Overconfidence* shows the expected number of correct answers minus the number of actually correct answers. *Total Time* measures time spent in doing the test in seconds (equivalently for *Total Time_Q4-Q7*).

Focusing on the overall sample, we observe statistical differences across all four treatments in all outcome variables. The two main treatments (Easy-Difficult, Difficult-Easy) always exhibit extreme values for all variables, while the treatments combining easy and difficult questions in the first block of questions always show intermediate values. The first and most important outcome variable is the probability to abandon the test. The proportion of participants abandoning the test is 50% higher in the Difficult-Easy treatment than in the Easy-Difficult one (44% vs 30%). Accordingly, the average number of questions answered is highest in the Easy-Difficult treatment (8.03) and lowest in the Difficult-Easy one (6.51). Participants answer on average more than one more correct question in the Easy-Difficult than in the Difficult-Easy treatment (3.53 vs 2.42). Regarding beliefs, participants expect to answer correctly the highest number of answers in the

Easy-Difficult treatment (6.24 questions) and lowest (5.75) in the Difficult-Easy one. Participants are also most overconfident when easy questions appear in the first block (1.97 vs 1.70). Finally, and fittingly given the previous results, participants take almost 25% more average time in solving the test in the Easy-Difficult treatment than in the Difficult-Easy one (207 vs 160 seconds).

The sample I is interesting because it shows that, among those who complete the first block of three questions, while there are no statistically significant differences in the proportion of subjects who abandon the test across treatments (*p*-value equal to 0.21) not even in the number of questions that they answer in the second block (*p*-value of 0.19), there are yet differences in the number of correct answers they provide in the second block of questions, those which are the same for all participants across treatments (1.52 correct answers on average in the Easy-Difficult treatment vs 1.39 in the Difficult-Easy one). Similarly, sample II shows a similar effect in the number of correct answers in the same second block of identical questions among those who complete the first and second blocks (1.66 vs 1.54 average correct answers). Finally, sample III confirms significant differences in the described direction in all outcome variables for the selected sample of subjects who complete the whole test. These results confirm that even among those who are not abandoning the test, the sequence of difficulty of the early questions (first block) affects performance in the second block, where the four questions are the same in all treatments.

[Figure 2 over here]

Consistently, Figure 2 focuses on the proportion of participants who do not answer all questions. By design, all participants answer the first question, but the probability of not providing an answer is increasing from question 2 onwards, as participants abandon the test. Figure 2 shows that the probability of not providing an answer is much larger and always above the rest of the treatments when the questions are ordered from the most difficult to the easiest ones.

[Figure 3 over here]

Figure 3 shows the average number of correct answers by the question and by treatment. Figure 3a validates our choice of questions in showing that, irrespective of treatment, easy questions were correctly answered in a higher proportion than difficult ones. Figure 3b amplifies the second block of questions, those of medium difficulty, which were the same for all participants,

to show that participants in the Difficult-Easy treatment answered correctly a lower number of questions in this second block than in the Easy-Difficult treatment.

Figure A1 in the Appendix shows the proportion of correct answers for the sample of participants who completed the test (sample III in Table 2), ordering identical questions together irrespective of the order in which they were presented in different treatments. Figure A1a shows how the proportion of correct answers decreases in all treatments as the questions were more difficult, as intended by the design. The following three figures disaggregate the results by the three blocks of questions, amplifying them to be able to see the patterns more clearly. Figure A1b shows that the proportion of correct answers among the easy ones (E1 to E3) is highest in the treatment where easy questions come first (Easy-Difficult) and that the difference increases as we move from E1 to E3. Similar to Figure A1b but for selected sample III, Figure A1c shows the same pattern for medium difficulty questions (M1 to M4). Finally, Figure A1d shows that the effect is less clear for the difficult questions (D1 to D3), which may be because there is less scope to being influenced by the order in which questions are presented when questions are already difficult.

These descriptive figures and results show that the order of difficulty affects performance: the best performance, both in terms of the completion rate and in the number of correct, occurs when the questions are ordered from easiest to most difficult.

2.4. Results: Regression Analysis

We now perform a regression analysis to study the effect of the order of difficulty when controlling for gender, age, and educational attainment. Table 3 shows OLS regressions for the overall sample having all the main outcome variables and the Easy-Difficult treatment as the omitted group. Odd columns (1, 3, 5, 7, 9, and 11) include all four treatments in the sample, while even columns (2, 4, 6, 8, 10, and 12) only compare the Easy-Difficult with the Difficult-Easy treatments and further includes the interaction between the female variable with the order of difficulty (*Female*Difficult-Easy*).

In the Difficult-Easy treatment, we find a higher proportion of participants abandoning the test (15 percentage points higher) and accordingly lower numbers of provided answers (1.5 fewer answers), correct answers (1.1 fewer answers), expectations of correct answers (0.45 fewer answers), overconfidence (0.26 smaller difference) and total time employed in the test (47 fewer

seconds), all highly significant with *p*-values < 0.01. The Easy-Difficult order, therefore, shows participants' best performance in all dimensions. Regarding the control variables, women do not significantly drop the test more than men do, although they provide a lower number of correct answers, believe they have performed worse, and show lower levels of overconfidence. These findings are consistent with the results in the literature regarding gender differences in confidence in performance in Math tests (see, for example, Bordalo et al., 2019, Rey-Biel and Iriberri, 2019, and Rey-Biel and Iriberri, 2021). When comparing Easy-Difficult with Difficult-Easy, the two extreme treatments, we do not find that women are differently affected by treatment Difficult-Easy except for the number of correct answers, where women are less negatively affected by the Difficulty-Easy treatment.

Tables A3, A4, and A5 in the Appendix show qualitatively similar results for each of the selected samples, described in Table 2, Section 2.3. That is, the order Difficult-Easy shows the participants' worst performance, from highest drop rates to lowest number of correct answers. The exception is given by the results in the sample of those who complete questions 1 to 3 (Sample I, shown in Table A3). Consistent with results in Table 2, those who complete the first block of questions do not have a higher probability of abandoning the test and do not provide a significantly different number of answers in the second block (columns 1 and 3 in Table A3). The remainder of coefficients in the three selected samples show that all performance measures are worse under the treatment Difficult-Easy, both in the second block of questions (where the questions are the same) and in the remainder of the test for each of the samples who complete each block of questions. All these results in different samples consistently show that the order difficulty matters and that the Easy-Difficult test structure shows the best performance measures.

3. Study 2: Observational Study Using PISA

3.1. PISA Data

The PISA assessment, sponsored by the Organization for Economic Co-operation and Development (OECD), is a triannual survey that evaluates educational systems around the globe in math, reading, and science. The participants are 15-year-old students approaching the end of their compulsory schooling in about 74 countries. For our study, we use the assessments and

student questionnaires from PISA 2009, 2012, and 2015 that have a final sample size between 178,00 and 270,000 for a total of about 600,000 participants.

The PISA assesses students in the areas of reading, math, and science. These assessments were all paper-based in the years of 2009 and 2012. In the year 2015, the main form of assessment was computer-based; paper-based tests were available for countries that did not have access to computers. The PISA test lasts about two hours and includes approximately 60 questions. After the assessment, students fill in a questionnaire that collects information about learning experiences, school environment, demographic and family factors, and student attitudes. These surveys are always administered immediately after the completion of the test.

Within each country, each participant is randomly assigned one of several test booklets to complete. Each booklet comprises four groups of questions (i.e., clusters) in different subjects that appear in different booklets and different positions (see Figure 4 panel [a]). Four different clusters of questions compose each of the test booklets (see Figure 4 panel [b]). The cluster rotation in PISA guarantees that most clusters appear four times, once on each of the possible positions from one to four (see Figure 4 panel [b]). Even though across booklets the clusters appear in different positions, the order of questions within clusters always remains the same.

Although in this analysis we have less control over the test structure than in our randomized controlled experiment described in section 2.1., the randomized assignment of PISA's test booklets allows us to exploit the different positions of clusters across booklets to estimate the effect of the order of difficult sets of questions on the proportion of blank responses and performance throughout the assessment. Importantly, with PISA we gain external validity of test-taking in low stakes contexts by having an international sample of 15-year-old students across multiple countries.

3.2. Methodology

3.2.1. Measuring Half-cluster Question Difficulty in PISA

For our analysis, we restrict our sample to those OECD countries whose students take the standard set of 13 booklets in PISA 2009 and 2012.[9] For PISA 2015, we restrict our sample to the OECD countries that took the computer-based assessment. We choose OECD countries to keep a more homogenous sample of countries that is the same across years.

In our empirical specification, we measure question difficulty at the half-cluster level by dividing PISA's cluster of questions into two halves based on the order of presentation. As it is explained above, question order only varies across clusters, not within clusters, but studying the difficulty at the half-cluster level allows us to have higher levels of variation in difficulty than if we consider the full cluster level. Table 4 presents some summary statistics of our main measures of interest such as difficulty, blank, and correct responses at the cluster and half-cluster levels. According to Table 4, the standard deviations of our main variables are mostly higher at the half-cluster level than at the cluster level. As a result, we decide to work with half-clusters instead of clusters because we want to maximize the variation of our variables of interest.

[Table 4 over here]

Our "baseline" measure of half-cluster difficulty corresponds to the average proportion of incorrect responses that students obtain in a given half-cluster when it appears at the beginning of the test (i.e., first and second positions). We use the information from all students to calculate the proportion of incorrect responses on the first and second half-clusters for each booklet, within each country. Focusing on measuring difficulty when the half-clusters appear at the beginning of the test allows us to obtain a cleaner measure of difficulty since at the beginning of the assessment students are less likely to have test fatigue and responses are not influenced by the difficulty level of prior questions.

[9] In 2006, and before this year, all countries participating in PISA received the same set of thirteen booklets. The booklets changed in 2009 when countries that achieved a low mean score in reading in 2006, or new countries that were expected to do so, had the option of taking an easier set of booklets.

3.2.2. Estimating the Effect of Half-cluster Question Difficulty

Our empirical approach tracks the effect of prior half-cluster difficulty on current half-cluster performance as students move along the test from one half-cluster to the next. We use ordinary least squares (OLS) and student fixed effects models for each year of PISA to obtain the effect of the difficulty of the prior half-cluster of questions on the proportion of blank questions and the proportion of correct answers in the current half-cluster.

In line with the analysis of our field experiment described in section 2.1., we estimate separate models depending on the position of half-clusters of questions in the test: the very beginning of the test (half-cluster in the second position); the beginning of the test (half-clusters in third and fourth positions); middle of the test (half-clusters in fifth and sixth positions), and end of the test (half-cluster in seventh and eighth position). We use linear regression models to estimate, separately, the effects of prior half-cluster difficulty on the proportion of blank and correct responses a student obtains in the current half-cluster. We control for current half-cluster difficulty, prior half-cluster difficulty (our main variable of interest), and the interaction of prior half-cluster difficulty and a dummy variable for the student being female.

For our estimates of prior half-cluster difficulty effects at the very beginning of the test (i.e., the half-cluster in the second position), we also control for country dummies to account for heterogeneity across countries. For half-clusters situated at the beginning (i.e., half clusters in third and fourth positions), middle (i.e., half-clusters in fifth and sixth positions), and end of the test (i.e., half clusters in seventh and eight positions) we follow a similar specification. However, in this case, we have two observations per student, and therefore, we estimate student fixed effects models to better control for student unobserved heterogeneity.

Our empirical approach aims to mimic the methodology of our field experiment described in subsection 2.4. Given that in the PISA students are required to stay for a minimum amount of time to complete the test, the closest situation to “dropping out” in the PISA test is to leave questions blank.

3.3. Results

Tables 5 to 7 present our estimation results from the PISA study. In particular, we assess how the probability of leaving questions blank varies throughout the test as a function of the difficulty level of the previous half-cluster group of questions, after controlling for the level of difficulty of the current half-cluster of questions (see columns 1 through 4 of Tables 5 through 7). We then study the effect of the half-cluster difficulty of the prior half-cluster set of questions on the proportion of current half-cluster correct responses in the whole sample (see columns 5 through 8 of Tables 5 through 7), as well as the sample restricted to only those students who do not leave any blank responses (see columns 9 through 12 of Tables 5 through 7). Figures 5 to 7 illustrate the estimated effect of prior half-cluster difficulty by gender. For the red bars, we show the statistical significance of gender differences on the effect of prior half-cluster difficulty.

The abbreviation HC in Figures 5 to 7 and Tables 5 to 7 represents different sections of the test: HC:2 represents the very beginning of the test (second half-cluster); HC: 3 and 4 represent the beginning of the test (third and fourth half-clusters); HC: 5 and 6 represent the middle of the test (half-clusters fifth and sixth) and HC: 7 and 8 represent the end of the test (half-clusters seventh and eight).

As shown by the results in Figure 5, and consistent with the results of the field experiment, the higher the difficulty of the previous half-cluster the higher the proportion of blank responses in the current half-cluster. We generally observe that prior half-cluster difficulty has the highest effects on the proportion of questions students leave blank towards the middle (HC: 5 and 6) and the end of the test (HC: 7 and 8). Overall, an increase of 10 percentage points in the difficulty of the previous half-cluster is associated with an increase between 0.4 to 1 percentage points in the proportion of half-cluster blank responses in HC: 5 and 6, while in HC: 7 and 8 this effect is between 1 and 2 percentage points. Also, female students often leave fewer blank responses than boys as the difficulty of the previous half-cluster increases (see Figure 5). These findings suggest that the difficulty of the previous half-cluster increases the proportion of blank responses in the current half-cluster.

When we analyze the effect of prior half-cluster difficulty on current half-cluster performance (see Figures 6 and 7), consistent with the findings from the field experiment, a higher

level of difficulty of the prior half-cluster of questions is associated with a lower half-cluster performance. According to Tables 5 through 7, the decline in performance is generally higher in the second half of the test (half-clusters 5 through 8). For example, using the whole sample, between half-clusters 5 through 8, when the difficulty of the previous half-cluster increases by ten percentage points, the decline in performance ranges from 0.6 to 5 percentage points. In the restricted sample, this decline is of a similar range. Altogether, our models explain between 15 to 22 percent of the within-student variation in the proportion of correct responses in a half-cluster. Finally, as presented in Figures 6 and 7, in most cases, boys and girls experience a similar decline in current half-cluster performance when the difficulty of the previous half-cluster increases.

4. Discussion and Conclusions

Performance in standardized assessments is compared to determine access to educational resources, labor market outcomes, and even to evaluate educational systems. However, many of these comparisons are done across similar tests which differ in the order in which the same questions are posed. In this paper we show, using two studies complementing each other in internal and external validity, that the order of questions with varying levels of difficulty can lead to different performance measures. In particular, placing the most difficult questions at the beginning of a test can lead to higher dropping out rates by participants and lower overall scores. Therefore, we argue that the order of the level of difficulty of the assessment questions should be taken into account both when designing the tests and when comparing performance across subjects who may have faced tests with different order of questions.

References

Akyol, S.P., Key, J. and Krishna, K. (2016). "Hit or Miss? Test Taking Behavior in Multiple Choice Exams." NBER Working Paper Nr. 22401.

Anderson, J. (1989). "Sex-Related Differences on Objective Tests among Undergraduates." *Educational Studies in Mathematics*, 20(2):165–177.

Atkins, W.J., Leder, G.C., O'Halloran, P.J., Pollard, G.H. and Taylor, P. (1991). "Measuring Risk Taking." *Educational Studies in Mathematics,* 22(3), 297-308.

Atwater, A., Saygin, P. O. (2020), "Gender Differences in Willingness to Guess on High-Stakes Standardized Tests". Mimeo.

Balart, P., and Oosterveen, M. (2019). "Females show more sustained performance during test-taking than males." *Nature communications* 10.1,1-11.

Baldiga, K. (2014). "Gender Differences in Willingness to Guess." *Management Science*, 60(2): 434-448.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). "Beliefs about gender." *American Economic Review*, *109* (3), 739-73.

Brenner, M. H. "Test difficulty, reliability, and discrimination as functions of item difficulty order." *Journal of Applied Psychology*, 1964, 48, 98–100.

Bridgeman, Brent, and Rick Morgan. (1996). "Success in college for students with discrepancies between performance on multiple-choice and essay tests." *Journal of Educational Psychology* 88.2: 333.

Coffman, Katherine B., and David Klinowski. (2020). "The Impact of Penalties for Wrong Answers on the Gender Gap in Test Scores." *Proceedings of the National Academy of Sciences of the United States of America, 117* (forthcoming). (Pre-published online April 6, 2020).

Chamorro-Premuzic, Tomas. (2006). "Creativity versus conscientiousness: Which is a better predictor of student performance?" *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 20.4: 521-531.

Espinosa M.P. and Gardeazabal J. (2013). "Do Students Behave Rationally in Multiple Choice Tests? Evidence from a Field Experiment." *Journal of Economics and Management*, Vol. 9, No. 2, 107-135.

Espinosa Alejos, María Paz & Gardeazábal, Javier, 2010. "Optimal Correction for Guessing in Multiple-Choice Tests." *Journal of Mathematical Psychology* 54(5), 415-425.

Funk, P., and Perrone, H. (2016). "Gender Differences in Academic Performance: The Role of Negative Marking in Multiple-Choice Exams." Working Paper.

Gerhard R., Valentin W. (2018). "Gender Differences in Willingness to Compete and Answering Multiple-choice Questions: The Role of Age." *Economics Letters*, 164: 86-89.

Gerow, J. R. (1980). "Performance on achievement tests as a function of the order of item difficulty." *Teaching of Psychology*, 7, 93–95.

Hambleton, Ronald K., and Ross E. Traub. (1974). "The effects of item order on test performance and stress." *The Journal of Experimental Education* 43.1: 40-46.

Iriberri, N., and Rey-Biel, P. (2019). "Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics." *The Economic Journal*, *129* (620), 1863-1893.

Iriberri, N., and Rey-Biel, P. (2021). “Brave Boys and Play-it-Safe Girls: Gender Differences in Willingness to Guess in a Large Scale Natural Field Experiment.” *European Economic Review.*

Karle, H., Engelmann, D., and M. Peitz, (2019) “Student Performance and Loss Aversion,” Rationality & Competition Discussion Paper No. 181.

Laffitte Jr, and Rondeau G. (1984). “Effects of item order on achievement test scores and students’ perception of test difficulty.” *Teaching of Psychology*, 11.4: 212-214.

McDonald, Angus S. (2001). “The prevalence and effects of test anxiety in school children.” *Educational psychology* 21.1: 89-101.

Pekkarinen, T., (2015). “Gender Differences in Behaviour under Competitive Pressure: Evidence on Omission Patterns in University Entrance Examinations”. *Journal of Economic Behavior and Organization*, 115: 94-110.

Plake, B. S., Thompson, P. A., & Lowry, S. (1981). “Effect of item arrangement, knowledge of arrangement, and text anxiety on two scoring methods.” *Journal of Experimental Education*, 49: 214–219.

Ramos, I. and Lambating, J. (1996). “Gender Difference in Risk-Taking Behavior and their Relationship to SAT-Mathematics Performance.” *School Science and Mathematics*, 96(4): 202-207.

Riener, G., Wagner, V. (2017). “Shying Away from Demanding Tasks? Experimental Evidence on Gender Differences in Answering Multiple-choice Questions.” *Economics of Education Review*, 59, 43-62.

Skinner, Nicholas F. (1999). “When the going gets tough, the tough get going: Effects of order of item difficulty on multiple-choice test performance.” *North American Journal of Psychology* 1.1: 79-82.Smouse, A. D., & Mung, D. C. (1968) “The effects of anxiety and item difficulty sequence on achievement test scores.” *Journal of Psychology*, 68:181–184.

Swineford, F. (1941). “Analysis of a Personality Trait.” *Journal of Educational Psychology*, 32(6):438–444.

Tannenbaum D. (2012). “Do Gender Differences in Risk Aversion Explain the Gender Gap in SAT Scores? Uncovering Risk Attitudes and the Test Score Gap.” Unpublished paper, University of Chicago, Chicago.

Tversky, Amos, and Daniel Kahneman. "The framing of decisions and the psychology of choice." *Science* 211.4481 (1981): 453-458.

Zeidner, Moshe. (2010). “Test anxiety.” *The Corsini encyclopedia of psychology*, 1-3.

Figures and Tables

Figure 1. *Design in Study 1: Smartick*

Treatments	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Easy-Difficult	E1	E2	E3	M1	M2	M3	M4	D1	D2	D3
2Easy-1Diffcult	E1	D2	E3	M1	M2	M3	M4	D1	E2	D3
2Difficult-1Easy	D1	E2	D3	M1	M2	M3	M4	E1	D2	E3
Difficult-Easy	D1	D2	D3	M1	M2	M3	M4	E1	E2	E3

Figure 2. *Proportion of No Answer by Treatment and Question*

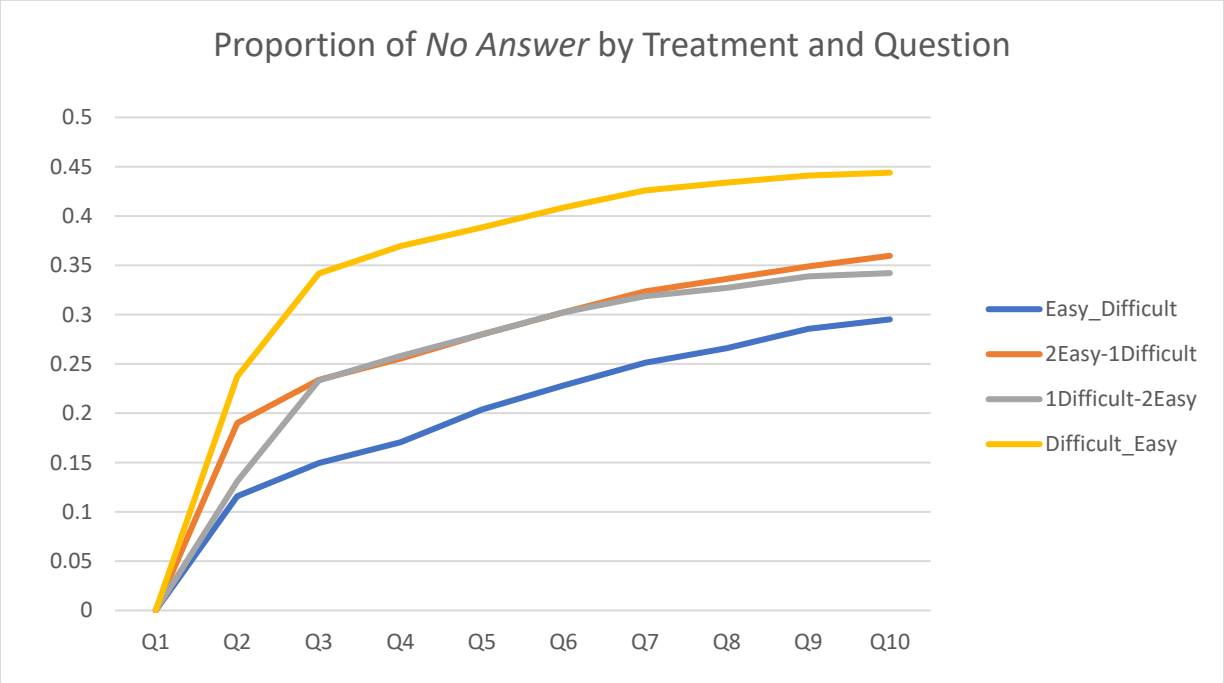
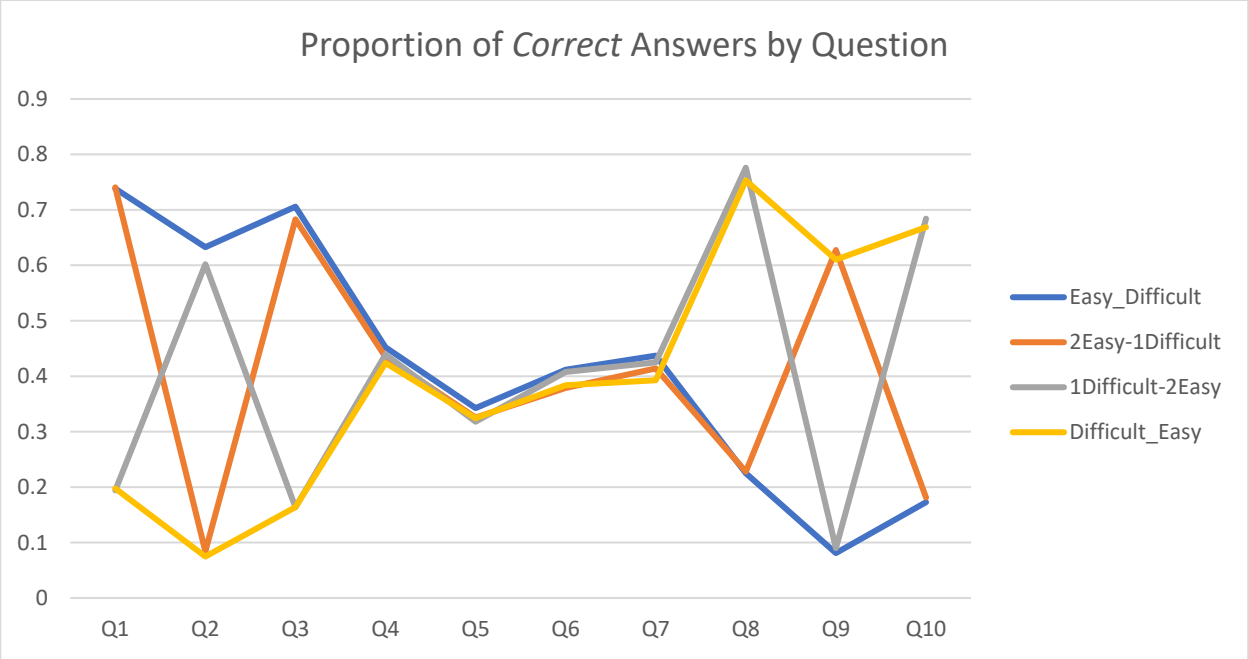


Figure 3. *Proportion of Correct Answers by Question and Treatment*

(a) Proportion of Correct Answers by Question (Overall)



(b) Proportion of Correct Answers in the Second Block of Questions (Q4-Q5-Q6-Q7)

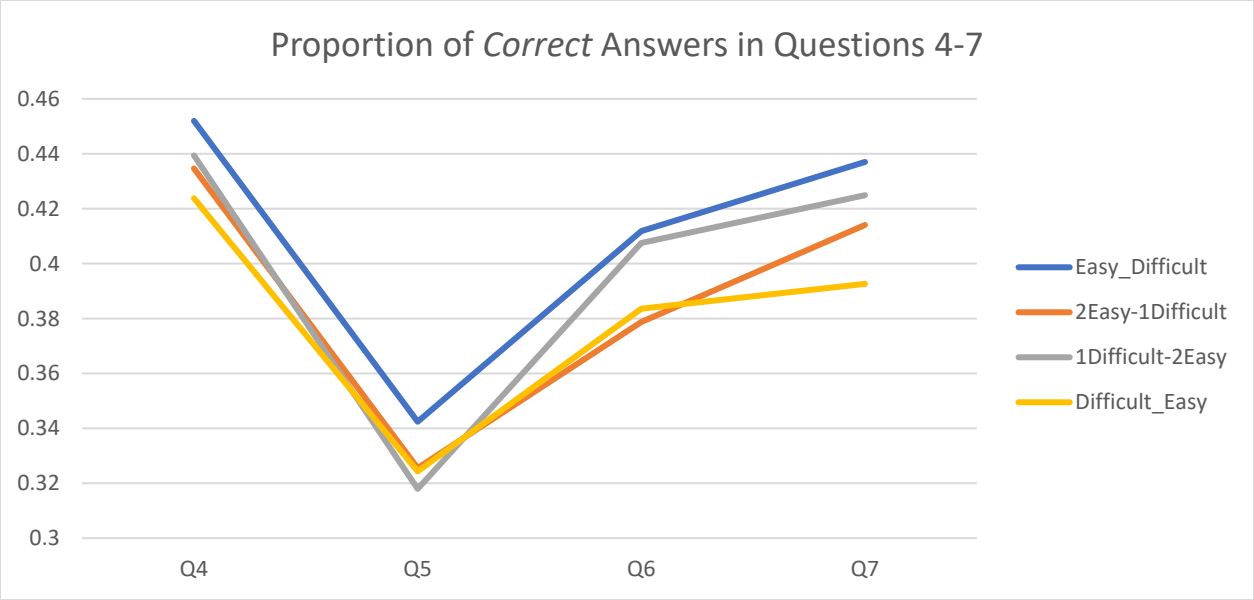
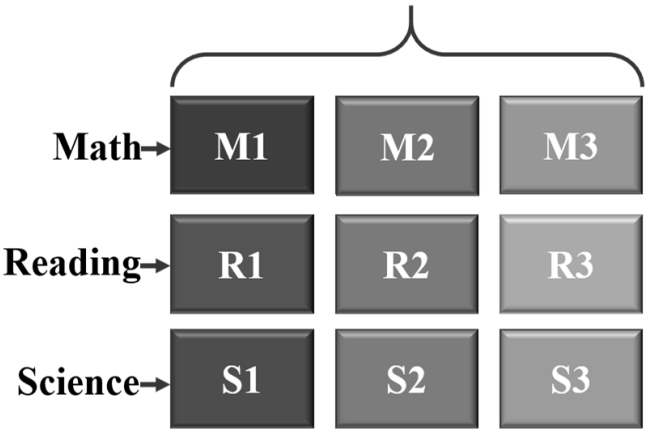


Figure 4: *Structure of PISA*

(a)

Cluster: group of questions about the same subject

Each rectangle represents a cluster of questions in a subject



(b)

Booklets: combination of 4 clusters

Clusters occupy different positions across booklets

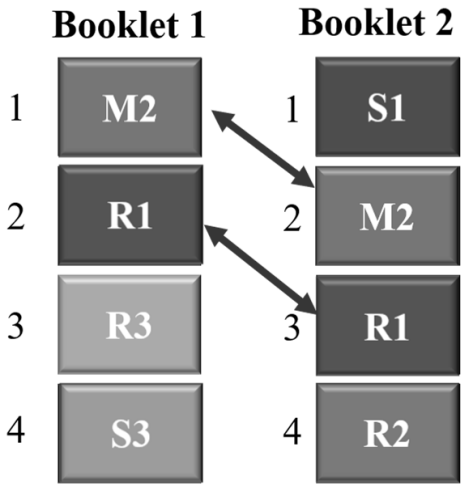
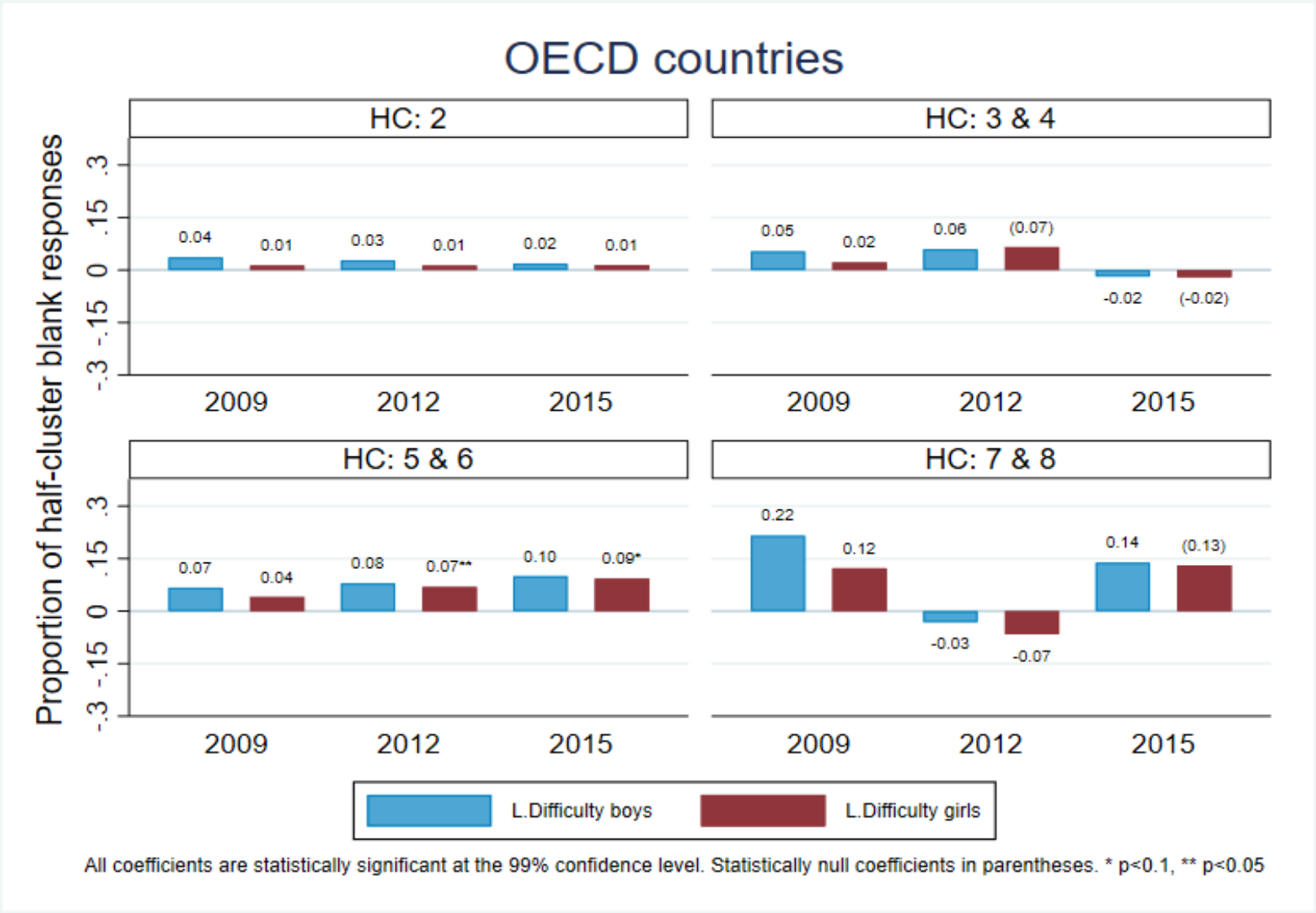


Figure 5: *Effect of previous half-cluster difficulty on the proportion of half-cluster blank responses*



All coefficients are statistically significant at the 99% confidence level. Statistically null coefficients in parentheses. * p<0.1, ** p<0.05

Figure 6: *Effect of prior half-cluster difficulty on half-cluster proportion of correct responses – whole sample*

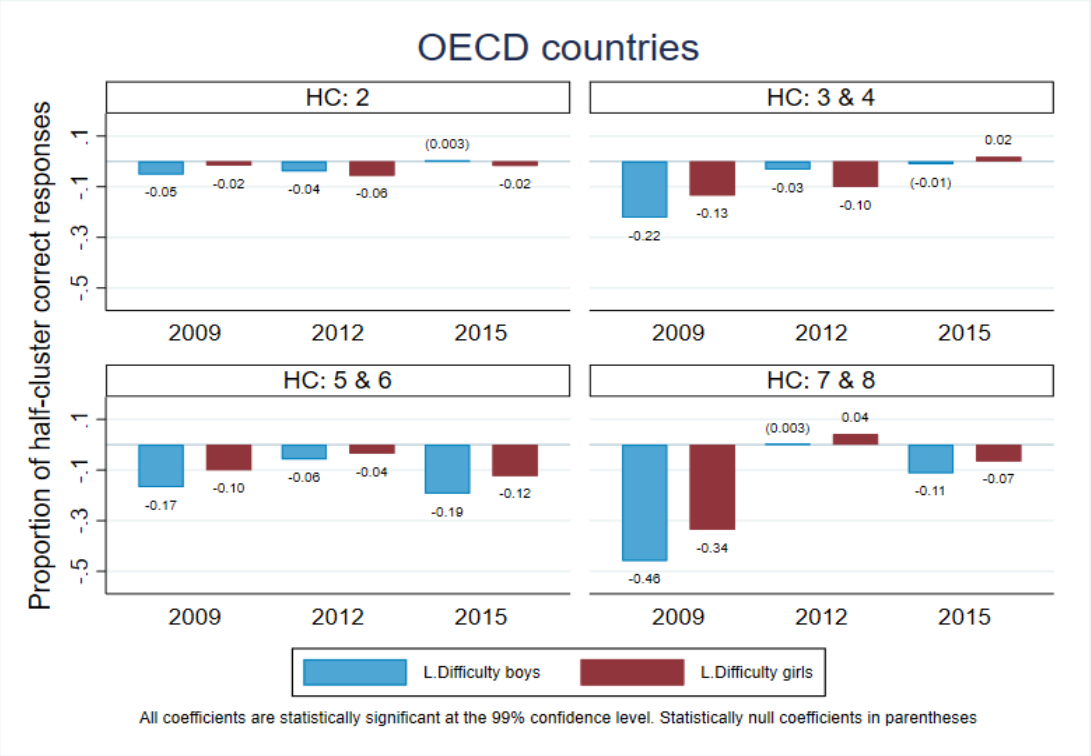


Figure 7: *Effect of prior half-cluster difficulty on half-cluster proportion of correct responses –sample restricted to zero blank responses*

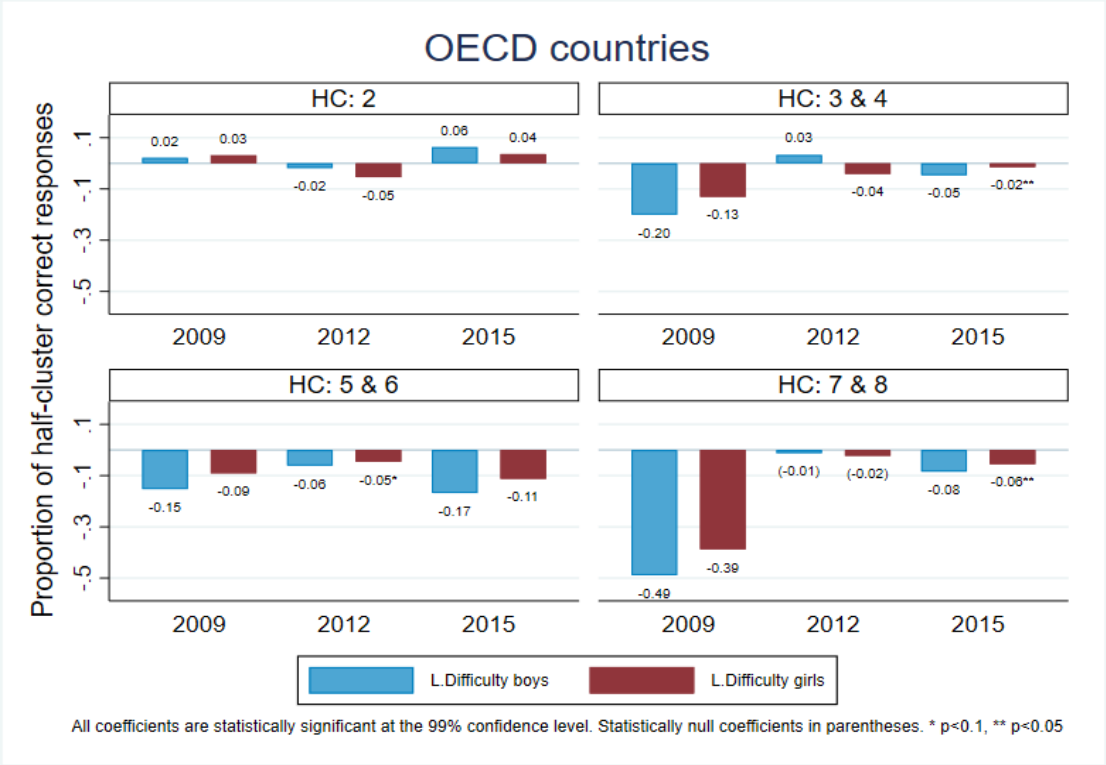


Table 1: Randomization Smartick

	Overall			Easy-Difficult			2Easy-1Difficult			1Easy2-Difficul			Difficult-Easy		
Variable	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
Female	18,952	0.36	0.48	4,799	0.35	0.48	4,604	0.36	0.48	4,663	0.37	0.48	4,886	0.35	0.48
Age:															
Age 4 to 11	18,952	0.07	0.26	4,799	0.08	0.27	4,604	0.07	0.25	4,663	0.06	0.24	4,886	0.07	0.26
Age 12 to 18	18,952	0.32	0.47	4,799	0.31	0.46	4,604	0.33	0.47	4,663	0.32	0.47	4,886	0.31	0.46
Age 19 to 24	18,952	0.15	0.36	4,799	0.15	0.36	4,604	0.15	0.36	4,663	0.15	0.36	4,886	0.15	0.35
Age 25 to 29	18,952	0.09	0.29	4,799	0.09	0.29	4,604	0.09	0.28	4,663	0.09	0.29	4,886	0.09	0.29
Age 30 to 39	18,952	0.13	0.33	4,799	0.13	0.34	4,604	0.12	0.33	4,663	0.12	0.33	4,886	0.12	0.33
Age 40 to 49	18,952	0.13	0.34	4,799	0.12	0.33	4,604	0.13	0.33	4,663	0.14	0.34	4,886	0.13	0.34
Age 50 to 59	18,952	0.07	0.26	4,799	0.07	0.26	4,604	0.07	0.26	4,663	0.08	0.27	4,886	0.07	0.26
Age≥60	18,952	0.04	0.19	4,799	0.04	0.20	4,604	0.04	0.18	4,663	0.04	0.18	4,886	0.04	0.19
Education:															
Primary	18,952	0.11	0.31	4,799	0.11	0.32	4,604	0.11	0.31	4,663	0.10	0.30	4,886	0.11	0.32
Some High School	18,952	0.19	0.39	4,799	0.19	0.39	4,604	0.19	0.40	4,663	0.20	0.40	4,886	0.18	0.39
High School	18,952	0.27	0.45	4,799	0.28	0.45	4,604	0.28	0.45	4,663	0.28	0.45	4,886	0.27	0.44
University	18,952	0.33	0.47	4,799	0.33	0.47	4,604	0.32	0.47	4,663	0.33	0.47	4,886	0.35	0.48
Post-University	18,952	0.09	0.29	4,799	0.10	0.30	4,604	0.09	0.29	4,663	0.09	0.29	4,886	0.09	0.29

Notes: The table shows the mean values of the control variables, overall and by treatment. All variables are dummy variables identifying female participants, participants from particular age groups and participants with a particular education level.

Table 2: Descriptive Statistics Across Treatments of Different Outcome Variables

	Overall Sample (No: 18,952)				
	Easy-Difficult	2Easy-1Difficult	1Easy2-Difficult	Difficult-Easy	
	(4,886)	(4,799)	(4,604)	(4,663)	*p* -values
Dropped	0.30	0.36	0.34	0.44	0.00
Total Answers	8.03	7.37	7.47	6.51	0.00
Total Correct	3.54	3.11	3.00	2.42	0.00
Guessed Correct	6.24	5.93	6.09	5.75	0.00
OverConfidence	1.97	1.77	1.93	1.70	0.00
Total Time (secs)	206.09	187.09	189.27	159.54	0.00
	Selected Sample I: Participants who complete Q1-Q2-Q3 (No: 14,433)				
	Easy-Difficult	2Easy-1Difficult	1Easy2-Difficult	Difficult-Easy	
	(4,156)	(3,677)	(3,530)	(3,070)	*p* -values
Dropped_Q4_Q7	0.12	0.12	0.11	0.13	0.21
Total Answers_Q4_Q7	3.23	3.26	3.30	3.21	0.19
Total Correct_Q4_Q7	1.52	1.44	1.47	1.39	0.00
Total_Time (secs)_Q4_Q7	96.87	94.13	93.74	88.15	0.00
	Selected Sample II: Participants who complete Q1-Q7 (No: 12,719)				
	Easy-Difficult	2Easy-1Difficult	1Easy2-Difficult	Difficult-Easy	
	(3,659)	(3,246)	(3,137)	(2,677)	*p* -values
Dropped_Q8_Q10	0.06	0.05	0.03	0.03	0.00
Total Answers_Q8_Q10	2.88	2.89	2.92	2.93	0.00
Total Correct_Q4_Q7	1.66	1.56	1.60	1.54	0.00
Total_Time (secs)_Q4_Q7	104.43	101.48	101.04	96.44	0.00
	Selected Sample III: Participants who complete all 10 questions (No: 12,139)				
	Easy-Difficult	2Easy-1Difficult	1Easy2-Difficult	Difficult-Easy	
	(3,444)	(3,073)	(3,029)	(2,593)	*p* -values
Total Correct	4.27	4.15	4.16	4.06	0.00
Total Correct_Q4_Q7	1.65	1.57	1.60	1.55	0.01
Total_Time (secs)	253.96	251.48	252.87	243.97	0.00
Total_Time (secs)_Q4_Q7	104.03	101.25	100.87	96.22	0.00

Notes: mean values of outcome variables. *Dropped* takes the value of 1 when the participant abandones the test. *Total Answers* measures the number of provided answers. *Total Correct* measures the number of total answers. *Guessed_Correct* measures the number of correct answers the participant expects. *Overconfidence* takes the difference between the number of correct answers expected by the participant minus the actual number of correct answers. *Total_Time* measures time spent in doing the test in seconds. The last column provides the *p*-value of an *F*-test where the null hypothesis is that the means from all 4 treatment groups are the same.

Table 3: Regression Analysis on the Overall Sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Dropped		Total Answers		Total Correct		Guessed Correct		Overconfidence		Total Time	
Difficult_Easy	0.149***	0.154***	-1.531***	-1.579***	-1.106***	-1.208***	-0.453***	-0.528***	-0.263***	-0.310***	-47.45***	-49.11***
	(0.00973)	(0.0121)	(0.0761)	(0.0952)	(0.0470)	(0.0613)	(0.0668)	(0.0805)	(0.0747)	(0.0904)	(2.412)	(3.052)
2Easy-1Difficult	0.0639***		-0.660***		-0.414***		-0.280***		-0.197***		-18.78***	
	(0.00947)		(0.0721)		(0.0441)		(0.0618)		(0.0683)		(2.362)	
1Easy2-Difficult	0.0465***		-0.564***		-0.512***		-0.139**		-0.0318		-16.83***	
	(0.00948)		(0.0718)		(0.0458)		(0.0618)		(0.0687)		(2.384)	
Female	0.0111	0.0215	-0.0645	-0.121	-0.456***	-0.590***	-0.878***	-0.904***	-0.256***	-0.234**	2.789	-0.580
	(0.00729)	(0.0138)	(0.0569)	(0.100)	(0.0335)	(0.0625)	(0.0498)	(0.0903)	(0.0551)	(0.100)	(1.807)	(3.346)
Female*Difficult_Easy		-0.0136		0.134		0.284***		0.215		0.133		4.645
		(0.0204)		(0.158)		(0.0945)		(0.143)		(0.160)		(4.985)
Observations	18,952	9,549	18,952	9,549	18,952	9,549	12,074	6,009	12,074	6,009	18,952	9,549
R-squared	0.025	0.035	0.032	0.052	0.112	0.132	0.080	0.080	0.008	0.009	0.058	0.075

Notes: OLS regressions for different oucome variables. Columns 1, 3, 5, 7, 9 and 11 show the results for the overall sample and all treatments. Columns 2, 4, 6, 8, 10 and 12 only include treatments *Easy_Difficult* and *Difficult_Easy*. In all regressions the omitted treatment is *Easy_Difficult*, and all regressions include control variables of age groups and completed studies. Robust standard errors in parentheses : *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Summary Statistics of Difficulty, Correct, and Blank Responses at the Half-cluster and Cluster levels for OECD countries

		PISA 2009				PISA 2012				PISA 2015			
Variable		**Mean**	**Std. Dev.**	**Min**	**Max**	**Mean**	**Std. Dev.**	**Min**	**Max**	**Mean**	**Std. Dev.**	**Min**	**Max**
Half-cluster difficulty %	Overall	34.9	9.1	13.2	65.8	38.6	9.0	13.8	72.4	42.6	10.4	14.1	87.1
	Between		5.9	23.9	58.6		5.8	25.3	64.2		5.9	27.7	68.9
	Within		7.7	10.7	55.3		7.5	12.9	66.6		8.5	12.6	80.2
Half-cluster correct responses %	Overall	52.2	27.8	0.0	100	49.3	27.6	0.0	100	48.0	27.6	0.0	100
	Between		20.7	0.0	100		20.5	0.0	100		20.4	0.0	100
	Within		18.7	-29.6	132.5		18.6	-34.4	130.3		18.6	-37.7	133.7
Half-cluster blank responses %	Overall	9.5	19.6	0.0	100	10.3	19.6	0.0	100	5.5	13.1	0.0	100
	Between		13.4	0.0	100		13.8	0.0	100		8.9	0.0	100
	Within		14.2	-76.2	95.3		14.0	-75.4	96.0		9.6	-80.2	91.3
Cluster difficulty %	Overall	34.5	7.3	18.0	61.2	37.9	6.6	21.0	62.8	42.6	8.4	20.5	73.7
	Between		5.9	23.0	61.2		5.2	25.4	57.7		6.1	26.6	69.8
	Within		5.2	20.8	49.1		4.7	21.4	54.5		5.8	16.2	67.8
Cluster correct responses %	Overall	51.7	24.7	0.0	100	49.7	24.1	0.0	100	47.5	24.0	0.0	100
	Between		21.1	0.0	100		20.8	0.0	100		20.7	0.0	100
	Within		12.9	-15.0	113.6		12.3	-15.0	111.6		12.1	-17.1	114.2
Cluster blank responses %	Overall	10.3	18.4	0.0	100	10.7	18.1	0.0	100	5.7	11.5	0.0	100
	Between		14.7	0.0	100		14.6	0.0	100		9.4	0.0	100
	Within		10.9	-56.4	77.0		10.6	-55.9	77.4		6.7	-60.9	72.4
Female %	Overall	50.0	50.0	0.0	100	49.8	50.0	0.0	100	49.7	50.0	0.0	100
	Between		50.0	0.0	100		50.0	0.0	100		50.0	0.0	100
	Within		0.0	50.0	50.0		0.0	49.8	49.8		0.0	49.7	49.7
Observations at the half-cluster level	Overall	N= 1,802,658				N= 1,838,939				N= 1,246,573			
	Between	n= 273,609				n= 272,444				n= 178,199			
	Within	T-bar =6.58845				T-bar =6.74979				T-bar =6.9954			
Observations at the cluster level	Overall	N= 761,172				N= 780,166				N= 534,187			
	Between	n= 266,880				n= 266,238				n= 178,199			
	Within	T-bar =2.85211				T-bar =2.93033				T-bar = 2.9977			

Table 5: Estimates for OECD countries - PISA 2009

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Proportion of half-cluster blank responses				Proportion of half-cluster correct responses - Whole sample				Proportion of half-cluster correct responses - Sample with zero blanks			
	HC: 2	HC: 3 & 4	HC: 5 & 6	HC: 7 & 8	HC: 2	HC: 3 & 4	HC: 5 & 6	HC: 7 & 8	HC: 2	HC: 3 & 4	HC: 5 & 6	HC: 7 & 8
Difficulty Half	0.083***	-0.014***	-0.027***	0.103***	-1.123***	-0.998***	-0.989***	-1.093***	-1.135***	-1.059***	-1.054***	-1.213***
	(0.003)	(0.003)	(0.004)	(0.005)	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.007)	(0.007)
Lag Difficulty Half	0.036***	0.053***	0.067***	0.216***	-0.052***	-0.223***	-0.168***	-0.459***	0.022***	-0.200***	-0.153***	-0.489***
	(0.003)	(0.003)	(0.005)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.008)	(0.009)
Female*Lag Difficulty Half	-0.023***	-0.032***	-0.027***	-0.093***	0.035***	0.088***	0.067***	0.124***	0.010***	0.069***	0.062***	0.101***
	(0.001)	(0.004)	(0.006)	(0.007)	(0.003)	(0.006)	(0.008)	(0.008)	(0.003)	(0.007)	(0.010)	(0.011)
Constant	0.011***	0.060***	0.085***	0.059***	0.993***	0.958***	0.911***	0.983***	1.029***	1.042***	1.015***	1.157***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Observations	273,609	533,760	516,972	478,317	273,609	533,760	516,972	478,317	218,180	393,010	353,420	285,694
R-squared	0.016	0.002	0.002	0.008	0.152	0.140	0.132	0.153	0.176	0.153	0.155	0.195
Number of students		266,880	260,163	240,834		266,880	260,163	240,834		227,424	209,182	172,963
R-squared within model		0.00215	0.00245	0.00820		0.140	0.132	0.153		0.153	0.155	0.195
R-squared overall model		0.00182	0.00225	0.00298		0.0965	0.0670	0.0518		0.139	0.102	0.0814
R-squared between model		0.00173	0.00232	0.00344		0.0855	0.0578	0.0380		0.136	0.100	0.0744

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 6: Estimates for OECD countries - PISA 2012

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Proportion of half-cluster blank responses				Proportion of half-cluster correct responses - Whole sample				Proportion of half-cluster correct responses - Sample with zero blanks			
	HC: 2	HC: 3 & 4	HC: 5 & 6	HC: 7 & 8	HC: 2	HC: 3 & 4	HC: 5 & 6	HC: 7 & 8	HC: 2	HC: 3 & 4	HC: 5 & 6	HC: 7 & 8
Difficulty Half	0.157***	0.178***	0.246***	0.253***	-1.179***	-1.213***	-1.188***	-1.154***	-1.185***	-1.163***	-1.180***	-1.152***
	(0.004)	(0.003)	(0.003)	(0.004)	(0.007)	(0.005)	(0.005)	(0.005)	(0.007)	(0.006)	(0.006)	(0.007)
Lag Difficulty Half	0.028***	0.059***	0.079***	-0.032***	-0.039***	-0.032***	-0.058***	0.003	-0.019**	0.033***	-0.061***	-0.012
	(0.004)	(0.004)	(0.004)	(0.005)	(0.007)	(0.007)	(0.005)	(0.006)	(0.008)	(0.008)	(0.007)	(0.008)
Female*Lag Difficulty Half	-0.015***	0.006	-0.010**	-0.034***	-0.017***	-0.069***	0.023***	0.039***	-0.035***	-0.074***	0.015*	-0.010
	(0.002)	(0.005)	(0.005)	(0.007)	(0.003)	(0.008)	(0.007)	(0.008)	(0.003)	(0.010)	(0.009)	(0.010)
Constant	-0.004*	-0.014***	-0.021***	0.069***	1.006***	1.009***	0.970***	0.891***	1.067***	1.034***	1.052***	1.015***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	272,444	532,476	516,991	517,028	272,444	532,476	516,991	517,028	193,431	375,264	342,437	312,438
R-squared	0.025	0.014	0.022	0.025	0.125	0.260	0.253	0.237	0.156	0.262	0.252	0.245
Number of id		266,238	260,027	260,064		266,238	260,027	260,064		221,310	205,476	189,843
R-squared within model		0.0135	0.0218	0.0248		0.260	0.253	0.237		0.262	0.252	0.245
R-squared overall model		0.000326	0.00244	0.00458		0.114	0.112	0.0903		0.153	0.143	0.135
R-squared between model		0.000193	0.000296	0.00140		0.0680	0.0693	0.0490		0.134	0.126	0.121

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 7: Estimates for OECD countries - PISA 2015

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Proportion of half-cluster blank responses				Proportion of half-cluster correct responses - Whole sample				Proportion of half-cluster correct responses - Sample with zero blanks			
	HC: 2	HC: 3 & 4	HC: 5 & 6	HC: 7 & 8	HC: 2	HC: 3 & 4	HC: 5 & 6	HC: 7 & 8	HC: 2	HC: 3 & 4	HC: 5 & 6	HC: 7 & 8
Difficulty Half	0.168***	-0.024***	0.044***	0.096***	-1.185***	-0.862***	-1.017***	-0.930***	-1.176***	-0.922***	-1.070***	-0.978***
	(0.003)	(0.004)	(0.003)	(0.005)	(0.008)	(0.008)	(0.006)	(0.008)	(0.008)	(0.009)	(0.007)	(0.010)
Lag Difficulty Half	0.018***	-0.018***	0.100***	0.139***	0.003	-0.012	-0.193***	-0.113***	0.063***	-0.047***	-0.168***	-0.084***
	(0.002)	(0.005)	(0.004)	(0.005)	(0.005)	(0.009)	(0.007)	(0.009)	(0.006)	(0.011)	(0.008)	(0.011)
Female*Lag Difficulty Half	-0.004***	-0.002	-0.007*	-0.008	-0.021***	0.030***	0.070***	0.047***	-0.027***	0.032**	0.055***	0.029**
	(0.001)	(0.005)	(0.004)	(0.005)	(0.003)	(0.010)	(0.008)	(0.009)	(0.003)	(0.012)	(0.010)	(0.011)
Constant	-0.038***	0.075***	-0.009***	-0.033***	1.026***	0.834***	0.995***	0.897***	1.040***	0.920***	1.058***	0.967***
	(0.002)	(0.003)	(0.002)	(0.004)	(0.004)	(0.006)	(0.004)	(0.006)	(0.005)	(0.008)	(0.005)	(0.008)
Observations	178,199	356,398	356,398	355,578	178,199	356,398	356,398	355,578	142,550	276,182	278,927	263,862
R-squared	0.024	0.000	0.007	0.006	0.146	0.153	0.204	0.162	0.160	0.158	0.216	0.174
Number of id		178,199	178,199	177,789		178,199	178,199	177,789		158,966	159,783	153,595
R-squared within model		0.000181	0.00715	0.00556		0.153	0.204	0.162		0.158	0.216	0.174
R-squared overall model		6.11e-05	5.56e-06	3.47e-05		0.108	0.114	0.0980		0.130	0.148	0.130
R-squared between model		0.000141	0.000428	3.01e-05		0.0923	0.0895	0.0791		0.125	0.137	0.122

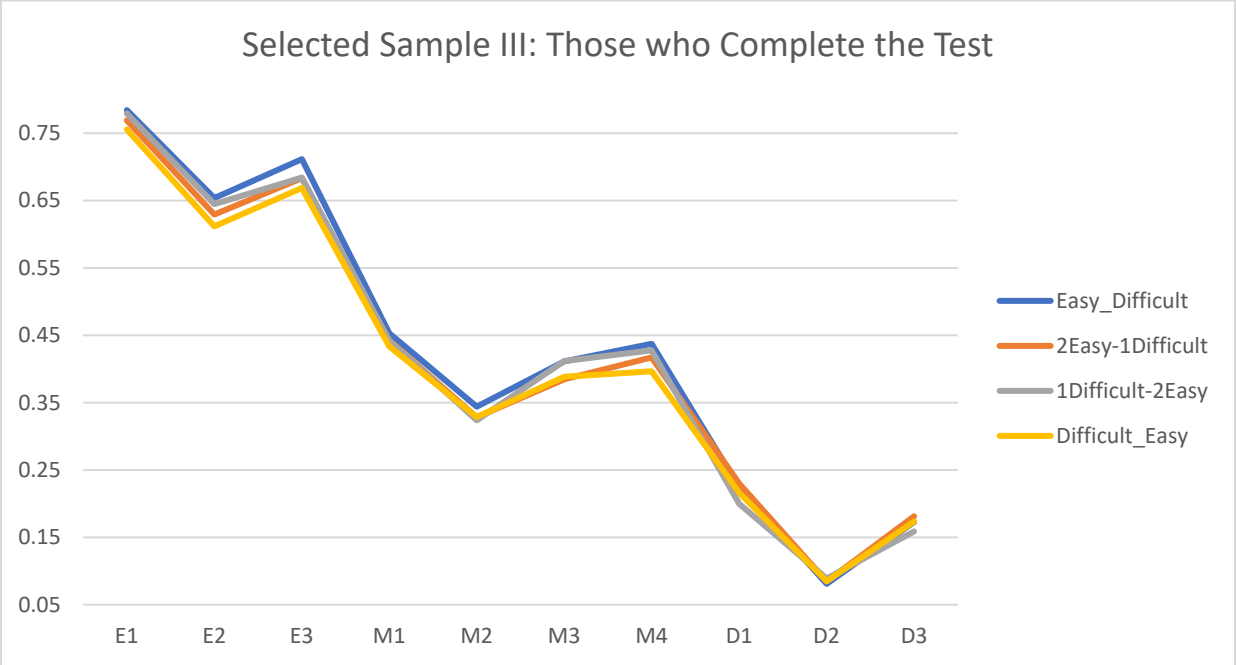
Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

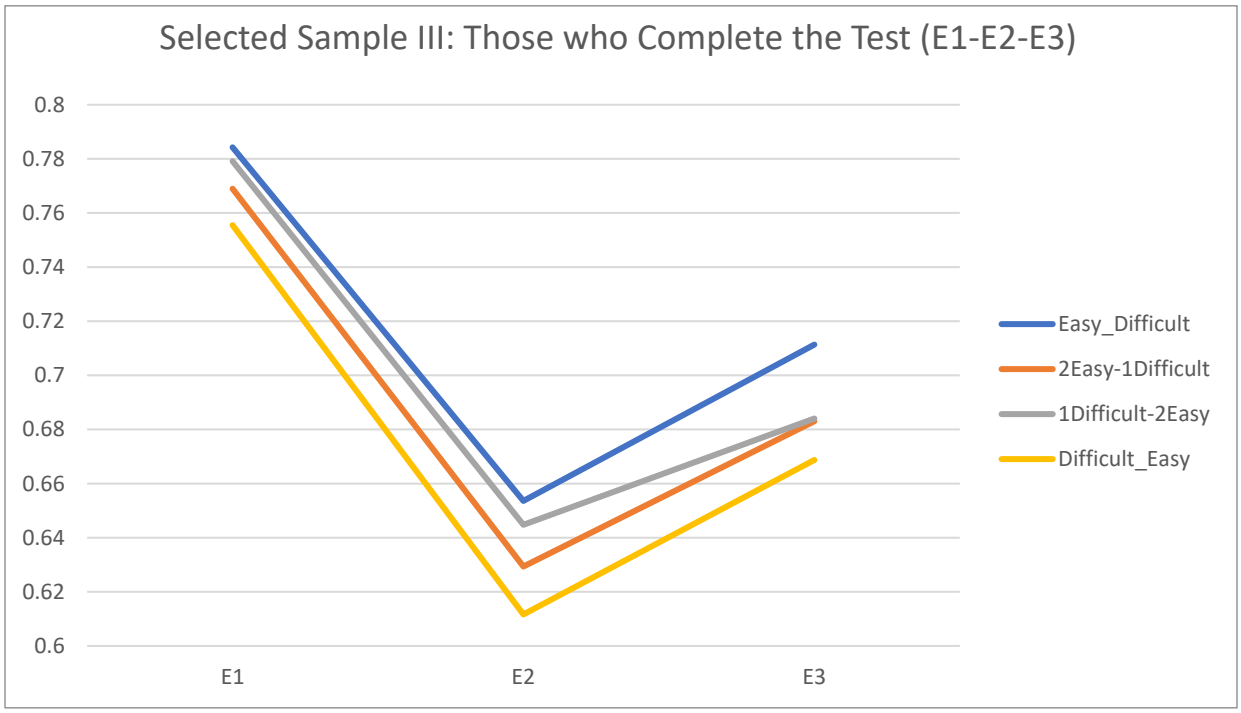
Figures and Tables in the Appendix

Figure A1. Proportion of Correct Answers by Question and Treatment for Selected Sample III: Those who Complete the Test

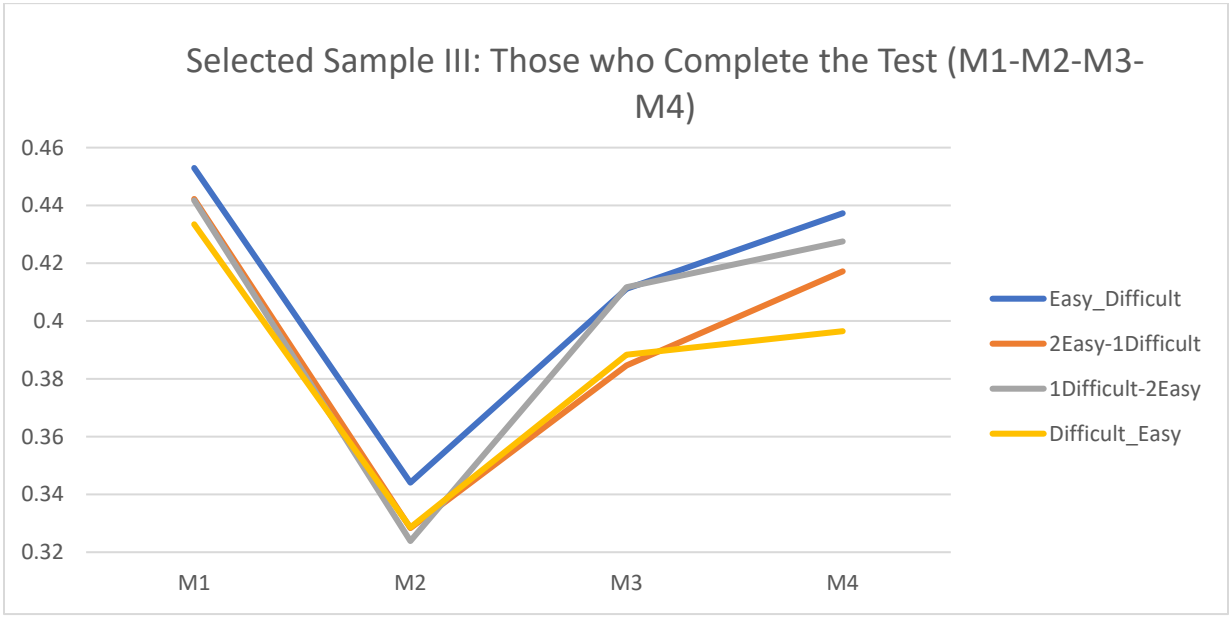
A1a. Proportion of *Correct* Answers for Selected Sample III



A1b. Proportion of *Correct* Answers for Selected Sample III: E1-E2-E3



A1c. Proportion of *Correct* Answers for Selected Sample III: M1-M2-M3-M4



A1d. Proportion of *Correct* Answers for Selected Sample III: D1-D2-D3

Table A1: Mean Correct Responses for Each Question

	Overall				Questions 1-2-3 Only		
Question	Obs.	Mean	Std. Dev.	Question	Obs.	Mean	Std. Dev.
	(1)	(2)	(3)		(4)	(5)	(6)
E1	15,424	0.75	0.43	E1	9,684	0.74	0.44
E2	14,054	0.62	0.49	E2	8,321	0.62	0.49
E3	13,456	0.69	0.46	E3	7,833	0.70	0.46
M1	13,982	0.44	0.50				
M2	13,513	0.33	0.47				
M3	13,089	0.40	0.49				
M4	12,719	0.42	0.49				
D1	16,039	0.21	0.41	D1	9,267	0.20	0.40
D2	13,978	0.08	0.27	D2	7,443	0.08	0.27
D3	13,116	0.17	0.38	D3	6,599	0.16	0.37

Notes: The mean correct for each of the questions in the test, averaged across the four different treatments (columns 1, 2 and 3) and averaged across the treatments but focusing only in the questions 1, 2 and 3 (columns 4, 5 and 6).

Table A2: Ordered Logit for Treatment

	Treatment
Female	0.0162
	(0.0274)
Age 12 to 18	0.0769
	(0.0715)
Age 19 to 24	0.0931
	(0.0788)
Age 25 to 29	0.0314
	(0.0841)
Age 30 to 39	0.0807
	(0.0807)
Age 40 to 49	0.0194
	(0.0805)
Age 50 to 59	0.0663
	(0.0859)
Age≥60	0.0467
	(0.0970)
Some High School	-0.0338
	(0.0627)
High School	-0.0376
	(0.0620)
University	-0.0861
	(0.0644)
Post-University	0.00837
	(0.0742)
Observations	18,952

Notes: the dependent variable takes the four values of the treaments. The control variables are identifyers for female, age groups and education groups, as described in the notes of Table 1. Standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1.

Table A3: Regression Analysis on the Selected Sample I: Participants Who Complete Q1-Q3

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Dropped_Q4_Q7		Total Answers_Q4_Q7		Total Correct_Q4_Q7		Guessed Correct		Overconfidence		Total Time_Q4_Q7	
Difficult_Easy	0.00801	-0.118***	-0.0159	-0.0326	-0.107***	-0.149***	-0.457***	-0.533***	-0.271***	-0.320***	-8.870***	-9.228***
	(0.00782)	(0.0106)	(0.0453)	(0.0547)	(0.0273)	(0.0348)	(0.0668)	(0.0805)	(0.0746)	(0.0903)	(1.149)	(1.451)
2Easy-1Difficult	-0.00363		0.0348		-0.0640**		-0.280***		-0.197***		-2.449**	
	(0.00729)		(0.0416)		(0.0260)		(0.0618)		(0.0683)		(1.081)	
1Easy-2Difficult	-0.00829		0.0742*		-0.0415		-0.139**		-0.0318		-3.066***	
	(0.00726)		(0.0419)		(0.0265)		(0.0618)		(0.0687)		(1.095)	
Female	0.00876	0.0177	-0.0671**	-0.154**	-0.314***	-0.372***	-0.878***	-0.904***	-0.255***	-0.233**	2.029**	0.461
	(0.00575)	(0.0131)	(0.0333)	(0.0627)	(0.0199)	(0.0370)	(0.0498)	(0.0903)	(0.0550)	(0.100)	(0.838)	(1.542)
Female*Difficult_Easy		-0.0103		0.0486		0.120**		0.218		0.139		1.015
		(0.0182)		(0.0969)		(0.0558)		(0.143)		(0.160)		(2.376)
Observations	14,433	7,956	14,433	7,226	14,433	7,226	12,070	6,005	12,070	6,005	14,433	7,226
R-squared	0.012	0.039	0.014	0.013	0.099	0.102	0.081	0.081	0.008	0.009	0.050	0.052

Notes: OLS regressions for different oucome variables. Columns 1, 3, 5, 7, 9 and 11 show the results for the overall sample and all treatments. Columns 2, 4, 6, 8, 10 and 12 only include treatments *Easy Difficult* and *Difficult Easy*. In all regressions the omitted treatment is *Easy Difficult*, and all regressions include control variables of age groups and completed studies. Robust standard errors in parentheses : *** p<0.01, ** p<0.05, * p<0.1.

Table A4: Regression Analysis on the Selected Sample II: Participants who Complete Q1-Q7

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Dropped_Q8_Q10		Total Answers_Q8_Q10		Total Correct_Q4_Q7		Guessed Correct		Overconfidence		Total Time_Q4_Q7	
Difficult_Easy	-0.0280***	-0.257***	0.0546***	0.0660***	-0.0988***	-0.137***	-0.457***	-0.533***	-0.271***	-0.320***	-8.299***	-8.160***
	(0.00513)	(0.00884)	(0.0119)	(0.0136)	(0.0290)	(0.0365)	(0.0668)	(0.0805)	(0.0746)	(0.0903)	(1.139)	(1.449)
2Easy-1Difficult	-0.00596		0.0164		-0.0723***		-0.280***		-0.198***		-2.779***	
	(0.00553)		(0.0123)		(0.0275)		(0.0618)		(0.0683)		(1.067)	
1Easy-2Difficult	-0.0247***		0.0490***		-0.0488*		-0.137**		-0.0299		-3.420***	
	(0.00507)		(0.0115)		(0.0281)		(0.0618)		(0.0687)		(1.082)	
Female	0.00485	0.0231*	-0.0108	-0.00989	-0.336***	-0.388***	-0.876***	-0.904***	-0.254***	-0.235**	3.278***	2.777*
	(0.00396)	(0.0138)	(0.00903)	(0.0188)	(0.0213)	(0.0396)	(0.0498)	(0.0903)	(0.0551)	(0.100)	(0.823)	(1.506)
Female*Difficult_Easy		-0.0125		-0.0342		0.112*		0.219		0.141		-0.407
		(0.0157)		(0.0263)		(0.0599)		(0.143)		(0.160)		(2.340)
Observations	12,719	7,563	12,719	6,336	12,719	6,336	12,066	6,003	12,066	6,003	12,719	6,336
R-squared	0.005	0.110	0.005	0.005	0.100	0.103	0.081	0.081	0.008	0.009	0.046	0.047

Notes: OLS regressions for different oucome variables. Columns 1, 3, 5, 7, 9 and 11 show the results for the overall sample and all treatments. Columns 2, 4, 6, 8, 10 and 12 only include treatments *Easy Difficult* and *Difficult Easy*. In all regressions the omitted treatment is *Easy Difficult,* and all regressions include control variables of age groups and completed studies. Robust standard errors in parentheses : *** p<0.01, ** p<0.05, * p<0.1.

Table A5: Regression Analysis on the Selected Sample III: Participants who Complete All 10 Questions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Total Correct		Total Correct_Q4_Q7		Guessed Correct		Overconfidence		Total Time		Total Time_Q4_Q7	
Difficult_Easy	-0.190***	-0.214***	-0.0871***	-0.121***	-0.458***	-0.531***	-0.273***	-0.320***	-11.02***	-11.30***	-8.257***	-8.068***
	(0.0512)	(0.0643)	(0.0297)	(0.0373)	(0.0668)	(0.0806)	(0.0747)	(0.0903)	(2.412)	(3.081)	(1.168)	(1.485)
2Easy-1Difficult	-0.0863*		-0.0571**		-0.277***		-0.196***		-2.292		-2.713**	
	(0.0478)		(0.0284)		(0.0618)		(0.0683)		(2.230)		(1.101)	
1Easy-2Difficult	-0.110**		-0.0403		-0.136**		-0.0285		-1.519		-3.334***	
	(0.0483)		(0.0288)		(0.0618)		(0.0688)		(2.248)		(1.111)	
Female	-0.623***	-0.663***	-0.337***	-0.376***	-0.877***	-0.902***	-0.254***	-0.232**	6.869***	4.243	3.607***	2.992*
	(0.0370)	(0.0686)	(0.0218)	(0.0409)	(0.0498)	(0.0903)	(0.0551)	(0.100)	(1.715)	(3.105)	(0.845)	(1.558)
Female*Difficult_Easy		0.0697		0.0985		0.211		0.134		0.873		-0.504
		(0.106)		(0.0615)		(0.143)		(0.160)		(4.935)		(2.399)
Observations	12,139	6,037	12,139	6,037	12,060	5,998	12,060	5,998	12,139	6,037	12,139	6,037
R-squared	0.144	0.144	0.099	0.100	0.081	0.081	0.008	0.009	0.061	0.061	0.047	0.048

Notes: OLS regressions for different oucome variables. Columns 1, 3, 5, 7, 9 and 11 show the results for the overall sample and all treatments. Columns 2, 4, 6, 8, 10 and 12 only include treatments *Easy Difficult* and *Difficult Easy*. In all regressions the omitted treatment is *Easy Difficult*, and all regressions include control variables of age groups and completed studies. Robust standard errors in parentheses : *** p<0.01, ** p<0.05, * p<0.1.

Appendix A: Test Questions in the Smartick Math Challenge

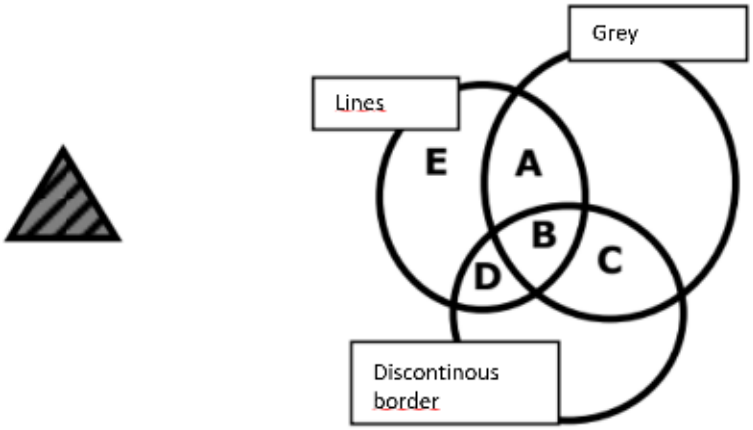
E1: Javier wrote a four digit number, and his brother painted on top of the second digit. Knowing that the number has no repeated digits, that the hidden digit is uneven and that, on top of that, such digit was neither the lowest nor the highest in the whole number, could you tell me which is the hidden digit?

Possible Answers: 1 4 5 7 9
Correct Answer: 7

E2: Luis turns 36 years old today. His age is 9 times bigger than that of his cat Bezout. His dog's age is three halves the age of the cat. The sum of the ages of the cat and the dog is…

Possible Answers: 8 9 10 12 13
Correct Answer: 10

E3: Please, take a look at the following diagram. In which region should the triangle be?



Possible Answers: A B C D E
Correct Answer: A

M1: How many turns does the minute hand of the clock takes in three days?

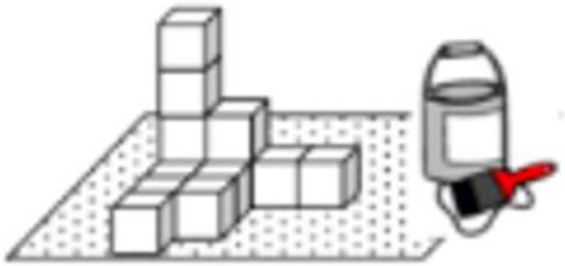
Possible Answers: 72 216 720 2160 4320
Correct Answer: 72

M2: If A and B are positive integers with A<B, which is the biggest fraction?

Possible Answers: (A-1)/(B-1) $(A^2-1)/(B^2-1)$ $(A^3-1)/(B^3-1)$ (A+1)/(B+1) It Depends on A and B

Correct Answer: (A+1)/(B+1)

M3: Ferb has constructed the following figure and, once done, has painted it red without lifting it off the ground. Phineas, has stepped into the figure and the figure has been dismantled. How many of the cubes do have exactly three faces painted red?



Possible Answers: None One Two Three Four

Correct Answer: Three

M4: A jar when full of honey weights 500 grams, and when full of milk weights 350 grams. If we know that the honey weights half what the milk weights, what is the weight of the empty jar?

Possible Answers: 100 150 175 200 225
Correct Answer: 200

D1: If we increase the length of all sides of a square in a fixed percentage, its area increases 96%. In which percentage would the area have decreased if instead of increasing the sides, we would have decreased the sides in the same percentage?

Possible Answers: 4% 64% 94% 48% 36%
Correct Answer: 64%

D2: How many pairs of integers (x,y) with $0 \leq x \leq y$ satisfy the equation $5x^2-4xy+2x+y^2=624$?

Possible Answers: 3 4 5 6 7
Correct Answer:7

D3: Order from lowest to highest the following three numbers, $P=11^{51}$, $Q=1317^{17}$, $R=37^{34}$.

Possible Answers: P<Q<R R<Q<P R<P<Q Q<P<R Q<R<P
Correct Answer: Q<P<R

Appendix B: Announcement of the Test on Social Media

Message shown on Smartick social media announcing the challenge

Translation:

Do you have an engineer's mind?

Where is all the knowledge you gathered at school? Do you know more than simple equations used in your everyday life? Find out in this test of mathematical intuition.

There are 10 questions of differing level of difficulty about logic, calculus, algebra and geometry. Some are easy and some are more difficult, but all could be solved by kids. Do you dare?

Start and find out your level!

Smartick blog, announcing the challenge

Landing Screen of the Math Challenge


Smartick
matemáticas a un clic
¿Tienes mente de ingeniero?
¿Dónde quedó todo lo que aprendiste en el colegio? ¿Van tus conocimientos más allá de las simples ecuaciones del día a día? Averigualo realizando este test de intuición matemática.
Son 10 preguntas de dificultad variable sobre lógica, cálculo, álgebra y geometría. Algunas difíciles y otras más sencillas, pero que podrían ser resueltas por los niños. ¿Te atreves?
¡Comienza y descubre tu nivel!


Algunos datos antes de empezar
¿Eres..?
¿Cuántos años tienes?
Selecciona tu edad
¿Cuál es tu nivel de estudios?
Selecciona
Empezar test