The role of analytical variability in secondary data replications: A replication of Kim et al. (2014)

Christian Bokhovea\*

a Southampton Education School, University of Southampton, Southampton, United Kingdom

Christian Bokhove  
Highfield Campus  
Building 32 Room 2035  
Southampton SO17 1BJ  
C.Bokhove@soton.ac.uk  
Orcid: 0000-0002-4860-8723  
Twitter: @cbokhove

Dr Christian Bokhove is Associate Professor in Mathematics Education with research interests in everything regarding mathematics teaching in secondary schools. He does a lot of research with large-scale assessments like PISA and TIMSS to get insights in mathematics education in an international perspective. In addition, he is involved in open research practices and research transparency.

**THIS IS THE ACCEPTED MANUSCRIPT FOR EDUCATIONAL RESEARCH AND EVALUATION. THE ONLY CHANGE HAS BEEN THAT TABLES HAVE BEEN INSERTED IN THE TEXT. OSF LINKS ARE STILL ANONYMISED.**

The role of analytical variability in secondary data replications: A replication of Kim et al. (2014)

An article from 2014 (Kim et al., 2014) examined individual- and school-level variables affecting the ICT literacy level of Korean elementary school students, finding differential gender effects. In this secondary data replication, we used data from the 2018 International Computer and Information Literacy Study, focussing on data from Korea as main replication. As many characteristics of the study as possible, such as variables and analytical strategy, were modelled in the analysis. Additional analyses included 13 countries and jurisdictions, varied centring techniques for variables, and missing data treatment. The replication and analyses were pre-registered via the Open Science Framework. The main analysis did not replicate the main gender finding. However, it was also clear that despite care taken in a rigorous replication, analytical variability still plays a large role in replications of findings, and with secondary datasets. We discuss the implications of this for secondary data replications.

Keywords: replication, international largescale assessment, analytical variability

# Introduction

In this article we present the findings of a conceptual replication of published research on Information and Communication Technology (ICT) literacy with recent secondary data. The first half of the article describes the replication process in detail. Although it does not replicate the findings, the results are very much influenced by researchers’ analytical decisions. So, the second half of the article discusses such ‘analytical variability’ and recommends that, although such decisions always will be made, transparency in the decisions made, can help to improve science.

Recently, open science practices in education research have received more and more attention, with some emphasis on the practice of *replication*. In a recent special issue on this topic in the journal *Educational Psychologist*, Plucker and Makel (2021) report that replications are still relatively rare. They describe replications as “intentional repetition of previous research to confirm or disconfirm the previous results, serving as a de facto reliability check on previous research” (p. 1). Taking this as starting point, it is useful to focus on influential articles and findings in the field of educational research. In the field of Information and Communication Technology (ICT) literacy, Kim et al. (2014) concluded that, while controlling for other factors, there were differential effects of gender on ICT literacy scores[[1]](#endnote-1). In the context of technology and literacy, such gender differences can be said to be important and so it would be useful and relevant to seek to replicate such a finding. However, the way in which we replicate the findings, for example in the same Korean context, with similar measurements, and other analytical decisions available, is not entirely straightforward. This article reports a replication of Kim et al. (2014) with data from an International Largescale Assessment (ILSA), termed the International Computer and Information Literacy Study (ICILS 2018). The structure of this article is atypical in that we reserve the literature on analytical variability in replications for the final part of the article. With ‘analytical variability’ we refer to the complete analysis ‘pipeline’, from data cleaning, statistical approaches to the final inferences made based on the statistical results. Variability can arise from such decisions like the choice of secondary dataset, the choice of variables, the creation of scales, the statistical methods used, and many other decisions. We take the replication as a given, we run the replication and then report findings and replication experiences. For this purpose, we first describe the original Kim et al. (2014) study and how we replicated the study, within the so-called Systematizing Confidence in Open Research and Evidence (SCORE) project, with pre-registration of the analysis and code. We describe how we followed the original study by Kim et al. (2014) as closely as possible, but that we also include some variations with regard to the sample (focal analysis Korea, as in the original, but we also looked at a sample of multiple countries), centring of variables, and imputation of missing data which were more prevalent in the ICILS 2018 data. We can show that analytical decisions influenced the findings, so, although we have a replication result on the primary outcome for Kim et al. (2014), different analytical decisions yield different results. In addition to these alternative analyses, there are many more analytical decisions made along the way. After reporting the substantive replication result for Kim et al. (2014), we therefore proceed with a discussion of analytical variability in replications with secondary data analyses.

# ICT literacy and gender: A replication

The replication of Kim et al. (2014) took place as part of the Center for Open Science's efforts under the SCORE project.The SCORE project aims to assess and improve the replicability of social-behavioural science.The project as a whole collects 30,000 papers published between 2009 and 2018 from 60+ journals in the social-behavioural sciences that publish primarily empirical, non-simulated research[[2]](#endnote-2) with human participants. This particular project replicates the findings from Kim et al. (2014), through secondary data from ICILS 2018, a study which will be expanded later on in the methodology section. As the study by Kim et al. (2014) has 71 citations at the time of writing, this suggests that it merits replication, as it contributes to the discussion of whether male or female students perform better in terms of Computer and Information Literacy (CIL). The focus of the study is not so much on the construct of CIL itself, but more on the gender findings of Kim et al. (2014). To contextualise this, we review below some of the relevant literature on gender differences in CIL.

Siddiq and Scherer (2019) conducted a meta-analysis of gender differences in ICT literacy. They noted that results were very inconsistent across studies, with numerous studies reporting that female students scored significantly higher than male students[[3]](#endnote-3), but several others showing an opposite direction, or reporting insignificant gender differences. Combining all these studies, they concluded that gender differences in ICT literacy were statistically significant, positive, and favoured girls, with effect sizes larger in primary school than in secondary school.

One of the included studies was that of Kim et al. (2014). Punter et al. (2016) looked in more critical detail at headline messages of ICILS 2013 data, which, in the majority of the participating countries, showed that 14-year-old girls outperformed boys in computer and information literacy (CIL), as this seemed to contrast with the common view of boys having better computer skills. They posit that differences arose because of differences in the type of assessment items, with the dimension ‘applying technical functionality’ being slightly in favour of boys, but no significant differences, and more information-oriented dimensions in favour of girls[[4]](#endnote-4). Gebhardt et al. (2019) reach the same conclusion in reporting that “female students performed relatively better on tasks that involved communication, design, and creativity, and male students generally performed relatively better on more technical tasks.” (p. 69).

With a new wave of ICILS data from 2018, the publisher of the ICILS studies, the International Association for the Evaluation of Educational Achievement (IEA) also used multilevel models in their report, to “review the extent to which different factors at the student and school level are associated with variations in CIL and CT scores” (Fraillon et al., 2020a, p. 217). They found that there were considerable differences in the variance for both CIL (and Computational Thinking) as well as the proportion of variance found between schools across participating countries. They concluded that in most countries being female was a significant, positive predictor of CIL. However, these analyses did not aim to replicate the original models by Kim et al. (2014). Only a few findings can be really compared to the full models in the original study, for example the positive predictors concerning ICT usage, experience with computers, CIL-related tasks at school, and the use of general applications in class. Other predictors, such as education level and socioeconomic background, fall outside the scope of both the original article and this replication. Fraillon et al. (2020a, Chapter 7) also note that they generally found considerable differences across participating countries, i.e. this is another reason why we, in line with the original study by Kim et al. (2014), focus the analysis only on Korea, and conduct an additional analysis for the full sample of countries.

Taken together, these studies present a mixed picture; A replication of the original study by Kim et al. (2014), along with new ICILS 2018 data, could further contribute to understanding gender and CIL. The next section describes the details of the original study and how it was replicated.

# Methodology

This methodology section, reports replication decisions in detail. However, some decisions were predetermined by the original approach in Kim et al. (2014), and therefore some of the more indirect decisions made (following on from the choice of dataset, for example), are discussed later in the article.

## Data

Data come from the ICILS 2018 study. The International Computer and Information Literacy Study 2018 (ICILS 2018) studied the extent to which young people are able to use information and communication technology (ICT) productively in school, home, society, and their future workplaces. The study looked at Computer and Information Literacy (CIL) and Computational Thinking (CT), with this replication focusing on the former. CIL refers to a student’s ability to use computer technologies to collect and

manage information, and to produce and exchange information. Data from different target audiences are collected. In this replication, the final dataset(s)[[5]](#endnote-5) comprised: one dataset including students’ information, one dataset including teacher information, one dataset including school principal and ICT coordinator information. The three datasets include all variables for the main analysis, which is a multilevel model that considers the individual and the school level for Korea alone, and an additional country level for the full sample. The selected variables in the final dataset were matched to the variables used in the original study, however, with some deviations, as not all questions in the ICILS 2018 dataset were the same as those in the original study. Differences are reported in the *variables* section.

## Replication claim

The claim selected for replication here, from Kim et al. (2014) is that gender significantly predicts the ICT literacy of the students. Kim et al. (2014) tested this by using a two-level Hierarchical Linear Model (HLM) to examine the effects of individual- and school-level variables on the ICT literacy of elementary school students, which resulted in the ICT literacy test score of female students being statistically significantly higher than that of male students. We have already alluded to the observation that there seems to be an error in the abstract, such that we take the result from the study to be that males score lower than females. Another element of the original study is that one cannot simply look at the gender effect, but one needs to include additional relevant variables. Therefore, the hypothesis in this replication is that the ICT literacy test score of female students will be higher than the ICT literacy test score of male students, while controlling for the other variables in the original study.

## Study design

The aim of the original study by Kim et al. (2014) was to identify individual- and school-level variables related to the ICT literacy level of Korean elementary school students. Previous studies before Kim et al. (2014) either did not account for the nested structure (students in schools) or they were conducted with different schools and student ages. The original study used stratified samples from a population of elementary schools across the country[[6]](#endnote-6). The data were collected through random sampling from schools within 1% (230 schools) of the stratified sample. In each sample school, one class from each of the 4th, 5th and 6th grades was selected through random clustered sampling. The final analysis subjects comprised 11,767 students in 173 schools who completed the appropriate questionnaires. Schools with 15 or fewer respondents were excluded from the analysis, according to the authors, for ‘stable analysis’. The data source for the replication is the ICILS 2018 study (Fraillon et al., 2020a). It is the second cycle of this study (the first cycle was conducted in 2013) and it was conducted by the International Association for the Evaluation of Educational Achievement (IEA). The ICILS 2018 study gathered data from 46,561 grade 8 (or equivalent) students in more than 2226 schools from 12 countries and 2 benchmarking countries. Representative samples were drawn by means of a systematic random sampling approach that involved multiple sampling stages, clustering, and stratification for two target populations (schools and students). In most participating countries, some 150 schools were sampled and, within schools, 20 students and 15 teachers were sampled. Minimum exclusion and target response rates were determined in order to secure data of sufficient quality[[7]](#endnote-7). Furthermore, the authors of the dataset include predefined indices of important variable constructs which, to a large degree, are comparable to, or even more detailed than, many of the relevant variables in the original study by Kim et al. (2014). The ICILS 2018 target student population comprised students in the grade that represented eight years of schooling, counting from Level 1 of the International Standard Classification of Education (ISCED), provided that the average age of students in this grade was at least 13.5 years at the time of the assessment. This is one main difference from the original study, which focused on students in the 4th, 5th and 6th grade in Korea. However, the original study does not focus too much on the “elementary school children” construct, but more on identifying the important predictors of both levels (individual and school level), as well as differentiating between them by holding constant any of the age groups.

The original study provides few theoretical arguments for the inclusion of certain predictors in the analysis. Some influential variables on which the authors reflect in their theoretical background, e.g., household computer environment, socioeconomic status, parental occupation and educational background, were not included in the original study’s analysis but were included in the ICILS 2018 study. However, the present replication keeps to replicating the original variables. Such lack of transparency with regard to the inclusion of certain predictors and their operationalisation, posed a first hurdle, which was that of vague reporting (e.g., missing variable units or missing provision of the questionnaire for the background variables). As a result, it was sometimes difficult to gauge which variable operationalisations of the potential replication data source (ICILS, 2018) would be appropriate. The analytical decisions and motivations are fully documented in the *variables* section and Table 1.

The international comparative character of the ICILS study is another main difference from the original study. Given that ICT literacy appears to be contextual (e.g. country infrastructure, the curriculum in a country), the main focus of the analysis for the replication here is on Korea. However, additional analyses were also conducted here, with all countries included. In many ways the ICILS 2018 dataset is more detailed than the dataset in the original study.

Regarding statistical power, an estimate of the minimum viable stage1 sample size indicates that the sample size needed to have 90% power to detect 75% of the original effect, for the data analytic replication, was 1,006 students, which meant that both the Korean and ‘all countries’ samples were sufficiently large.

## Complex sampling design

Different from the original study by Kim et al. (2014), the ICILS 2018 study has a complex sampling design. Students are sampled within schools within countries. The complex sampling design of the study is taken into account in this replication (Fraillon et al., 2020b; Rutkowski et al., 2010). This means that three levels – students, schools and countries – are included in the analyses. To cater for different probabilities of units being selected, sampling weights are used. Sampling weights ensure that the choice of sampling design does not have an undesired effect on the analyses of data. In this case, final weights are used at the student and school level. As ICILS 2018 uses an incomplete and rotated-booklet design for testing children on the major outcome variables, five plausible values are used[[8]](#endnote-8). To analyse the data, we combine the five plausible values into a single set of point estimates and standard errors, using Rubin’s rules (Rubin, 1987). For variance estimation, multilevel modelling takes the hierarchical structure of the cluster sample into account. The combination of multilevel modelling, final weights and five plausible values as estimates of students’ CIL, means that the reported multilevel standard errors address both sampling and imputation errors (Fraillon et al., 2020b, p. 230). The choice of the ICILS 2018 dataset introduced particular requirements for the analyses, something that we return to in the discussion section of the article.

## Variables

Table 1 presents the original variables from Kim et al. (2014) and the variables used for the replication.

Table 1 Variables used in the replication.

|  |  |  |
| --- | --- | --- |
| **Kim et al. (2014)** | **Replication based on ICILS 2018** | **Variable** |
| ***Dependent variable*** | |  |
| National-level ICT literacy test scores (see Kim et al, 2014, p.31). | ICT literacy test scores, 5 plausible values (PV1CIL-PV5CIL). Computer and information literacy refers to “an individual’s ability to use computers to investigate, create and communicate in order to participate effectively at home, at school, in the workplace, and in the community” (Fraillon et al., 2020a, p. xvii). | PV1CIL-PV5CIL |
| ***Independent variables for the individual level*** | |  |
| Gender (0=female, 1=male) | Sex (0=boy, 1=girl)[[9]](#endnote-9)[[10]](#endnote-10) | S\_SEX |
| Completion of computer course (Incomplete, complete) | Variable in ICILS 2018: Do you study [computing, computer science, information technology, informatics or similar] in the current school year?  Units in ICILS 2018: Yes (1), No (2). | IS2G30 |
| A variable is created by taking the mean of questions IS2G23A-I. These variables in ICILS 2018 ask: ‘At school, how often do you use ICT during lessons in the following subjects or subject areas?’ (IS2G23A-I).  Units: I don’t study this subject/these subjects (1), Never (2), In some lessons (3), In most lessons (4), In every or almost every lesson (5). So, the range of this variable is [1,5]. | ICT\_INTS |
| Students’ computer usage time for study. Unit not indicated, probably minutes. Log-transformed (unknown why). | Variable in ICILS 2018: Use of ICT for study purposes (Student’s questionnaire, S\_USESTD), pre-defined interval-scaled index.   * Units in ICILS 2018: Higher values indicating more frequent use. * Difference from original study: Even though the original study entails probably usage time in minutes (not indicated, but perhaps due to log transformation), the proposed index measures something similar. | S\_USESTD |
| Satisfaction level of students in classes using ICT. Average value of four 7-point Likert items (interests, fun, understanding of class contents and satisfaction of learning effect) for classes in which the computer was used by teachers. | This variable is not adequately reflected in the ICILS study. However, the aspect of understanding class content in combination with satisfaction of learning effect is reflected in the ICILS study in the index concerning learning of ICT tasks at school. In addition, the aspect of the original study concerning the interest of students in ICT use in classes is taken to be reflected in the ICILS study in the index concerning expectations of future ICT use for work and study.   * Variable in ICILS 2018: Learning of ICT tasks at school scale. | S\_ICTLRN |
| * Variable in ICILS 2018: Expectations of future ICT use for work and study scale. | S\_ICTFUT |
| Student’s computer usage time for purposes other than study. Unit not indicated, probably minutes. Log-transformed. (unknown why). | These variables were not directly included in the original study, but the scale adds to the construct of ‘student’s computer usage time for purposes other than study’.   * Variable in ICILS 2018: Use of ICT for social communication. This index is based on the variables IS2G20A-J, however they only include social communication activities. | S\_USECOM |
| * Variable in ICILS 2018: Use of ICT for social communication. This index is based on the variables IS2G20A-J, however they only include ICT use for exchanging information. | S\_USEINF |
| * Variable in ICILS 2018: How often do you use ICT for the following leisure activities? (IS2G21A-H). The question comprise 8 items, such as “read news stories on the internet” or “play games”, and the students need to self-assess their usage frequency on the scale: Never (1), Less than once a month (2), At least once a month but not every week (3), At least once a week but not every day (4), Every day (5). | ICT\_INTO |
| ***Independent variables for the school level*** | |  |
| School size. Small school with 12 or fewer classes. Medium-sized school with 13–35 classes. Large school with 36 or more classes. | Variable in ICILS 2018: Number of students in school – categorized.   * Units: 1-300 (1), 301-600 (2), 601-900 (3), 901 and above (4). * Difference from original study: The original study differentiates between small, medium and large schools based on the number of classes. This information is not available in the ICILS. However, the total number of students is seen as an adequate substitute, as it will be correlated with number of classes. | P\_NUMSTD\_CAT |
| Location. Islands, isolated and rural areas. Small and medium-sized cities. Major cities | Variable in ICILS 2018: Which of the following best describes where your school is located?   * Units: In a community with fewer than 3,000 people (1), In a town with at least 3,000 but less than 15,000 people (2), In a town with at least 15,000 but less than 100,000 people (3), In a city with at least 100,000 but less than 1,000,000 people (4), In a city with 1,000,000 or more people (5) | IP2G07 |
| Academic achievement level. Average percent of students with normal or excellent level for three subjects including Korean language, mathematics and English in the 2011 National Assessment of Educational Achievement | Difference from original study: This information is not included in the ICILS 2018. | - |
| Number of personal computers per student | Variable in ICILS 2018: Pre-defined ratio of school size and number of computers available for students.   * Difference from original study: The original study includes the number of PCs per student (it is not known if this construct only includes desktop computers, or laptops/notebooks and tablet devices as well). This variable in the ICILS 2018 refers to all ICT devices available for students. | C\_RATSTD |
| Number of PCs per teacher | Here we created the variable by using non-restricted data. C\_RATTEA <- ((II2G07AA1-II2G07AA2)+(II2G07AB1-II2G07AB2)+(II2G07AC1-II2G07AC2))/(12.5+ 25\*(P\_NUMTCH\_CAT-1)) with the first six variables denoting the number of devices available for teachers (all devices minus those for students), divided by a categorical variable denoting the number of teachers. The categorical variable had the categories 1-25 (1), 26-50 (2), 51-75 (3), more than 75 (4), hence midpoints are estimates. | C\_RATTEA |
| Teacher’s computer usage ratio for educational purposes. Mean of teachers’ computer usage ratio for educational purposes for each school | The ICILS 2018 teacher questionnaire variable ‘Use of ICT for teaching practices in class, with a pre-defined interval-scaled index describing usage of different ICTs by the teacher, was used. Measurement direction: Higher values indicating more frequent use. T\_ICTPRAC is the aggregate value of all teachers in a particular school. | T\_ICTPRAC |
| Teacher’s satisfaction level of using computer resources. Mean of teachers’ average values of three items (class preparation and implementation, support for administration tasks of academic affairs and individual learning including training) aggregated for the school. | Variable in ICILS 2018: Teachers agreement (satisfaction) comprising a statement concerning computer resources and opportunities at school, pre-defined interval-scaled index. Measurement direction: Higher values indicating stronger agreement. T\_RESRC is the aggregate value of all teachers in a particular school. | T\_RESRC |
| ***Organisational variables*** | |  |
| School ID | Student ID, School ID[[11]](#endnote-11), Country ID, weight variables. |  |

## Analytical approach

Two-level mixed models were generated with the WeMix R package (similar to Model 5, Table 6 in the original study by Kim et al. (2014)). The present replication study used only the Korean data for the focus of this analysis, as the original article hints at a specific country effect for Korea***[[12]](#endnote-12)***. Two levels are used in the models with Korea alone: student (IDSTUD) and school (schoolid). The dependent variables are five plausible values PV1CIL-PV5CIL. Models are estimated separately and then combined, following Rubin’s rules (Rubin, 1987). As indicated in Table 1, there are 15 predictors in total. Nine predictors are at the student level: S\_SEX, IS2G30, ICT\_INTS, S\_USESTD, ICT\_INTO, S\_ICTLRN, S\_ICTFUT, S\_USECOM, S\_USEINF. As in Kim et al. (2014), these were group-mean centred. Six predictors are at the school level: P\_NUMSTD\_CAT, IP2G07, C\_RATTEA, C\_RATSTD, T\_ICTPRAC, T\_RESRC. As in Kim et al., these were grand-mean centred. Weights at the student level are TOTWGTS and at the school level TOTWGTC. No missing data treatment was performed for the analysis in question here, as in the original study, but the present study included supplemental analyses that did impute missing data, in order to explore what impact missing data might have had. This was especially useful for the analyses with multiple countries, as missing data percentages were quite high. For the supplemental analyses with all countries, a third level was added: CNTRY. To do this, a weight variable TOTWGTN was created with only 1s, in order to ensure equal weights. To test the R code before pre-registration, we used random samples of the data. For Korea, this included 30% of the data, and for the dataset with multiple countries this was 5%. The R code generates estimates, standard errors and p-values. The criterion for a successful replication attempt for the SCORE project, was a statistically significant effect (alpha = .05, two-tailed), with the same pattern as the original study. For this study, this criterion was met by determining the direction and statistical significance of the S\_SEX predictor. No data were intentionally excluded. However, missing data did lead to lower sample sizes.

## Missing data

The original Kim et al. (2014) article did not mention missing data, nor did it describe any missing data treatment. The ICILS 2018 dataset does have some missing data on key variables, so the decisions indicated below were made with regard to missing data. It should be noted, however, that there are many ways to address missing data, and that some arbitrariness will inevitably be introduced. The main decision in this replication was to not use any missing data treatment, except for the default in-built exclusions by the WeMix package used. Note, also, that the original study used HLM software which does not allow missing level 2 data but does allow missing level 1 data. To adhere to an open data analysis approach[[13]](#endnote-13) with available R packages, we can only abide by the in-built missing data rules in the software, in this case the WeMix package. This means that, in addition to completely missing data for some schools, missing data were sizeable, as there were always some variables that had missing data. Nevertheless, it seemed the most trustworthy replication, as no missing data treatment was reported in the original study. To offer an alternative, and to determine the possible effects of the missing data, the analysis also used Expectation-Maximisation simple imputation with both the final student and school files, using the R package simputation[[14]](#endnote-14). The approach used here was also used for numeric categorical values; although there were other methods that yield only the allowable categorical variables, for this analysis, substantive conclusions were not affected. Methods involving ‘Multiple Imputation’ were not chosen, as this would mean that, because of the five plausible values, 5 x 5 datasets would need to be combined. As both school and student files had to be merged, further missing data occurred where there were no school ID data. For example, the Korean student file had a student in school 1127, but the school file did not have any relevant data on this school. If such complete records were missing, these were not imputed. This contributed to a lower sample size, as reported in the results table. As a result, the focus of the present analysis is conducted without missing data treatment, and an additional analysis with missing data treatment. The analysis plan, including code, was pre-registered at ANONYMISED LINK in November 2020[[15]](#endnote-15). The main analysis was run in January, 2021.

## Summary of replication decisions: analytical variability

Several analytical decisions were associated with the choice of data source. ICILS 2018 focuses on a different age from that of the original study. It also samples students from schools, rather than whole classes in the original study. The complex sampling design meant that plausible values, weights and resampling techniques would have to be used, with the latter accounting for using the same analytical strategy as that of the original study, viz., multilevel modelling. As noted earlier, most variables could be included in a way in order to do justice to the original study. Yet, the way in which they were operationalised, constituted another analytical decision. The fact that many variables can be included, either as part of the models or as control variables, constitutes yet another analytical decision. Regarding unreported missing data in the original study, ICILS 2018 *did* have missing data which could have been treated, and such treatments can in come many forms. We did not impute missing data for the focal replication but included imputed datasets as additional analyses, to see if this made a difference to the results. The data included multiple countries, but, for the focal replication, we included only Korea, with the full dataset used for additional analyses. This, of course, also has implications for the number of levels to include in the multilevel models, as, with more countries, an additional level ‘country’ is needed. Finally, the original study used centring, and, in a multilevel context, this can make a real difference to the outcomes (e.g. see Wu & Wooldridge, 2005[[16]](#endnote-16)), so we also included non-centred analyses. To summarise, we included one focus of analysis that was as close as possible to the original study, except for vague reporting and subject to dataset choice, with no missing data treatment, centring and including only Korea. There were seven further analyses, with eight in total (2×2×2, [not\_imputed-imputed]×[Korea-All\_countries]×[centering-no\_centering]), which included multilevel analyses with the variables presented in Table 1. Taken together, many analytical decisions are already introduced by the choice of dataset, with spin-off choices causing more decisions to be taken, for example the missing data treatment decisions (if any, and with what method) or having to include all five plausible values. The implications of these are discussed in the final section of this article.

# Results

The final analytic sample for the focal analysis included 2,513 students from 135 schools, which met the stage1 required sample size of 1,006 (and the stage2 sample size 2,263 threshold), as defined by the power analysis. The full model yielded a small, statistically significant, negative (standardised) coefficient for S\_SEX of -0.002, *p*=.0059[[17]](#endnote-17). This meant that the original result was not replicated: girls performed slightly worse than boys. As described in the methodology section, seven additional analyses were run, i.e. eight analyses in total. The results of the additional analyses show that single imputation generally *does not* appear to influence the outcome, whereas centring and the inclusion of all countries instead of Korea alone *does* seem to influence the outcomes, even reversing the main outcome, see Table 2. The bottom seven rows in Table 2 show the parameters for the eight models.

Table 2 Eight models (one main focus of the analysis and seven alternative models). Rounded to three decimal places. \* *p* < .01, \*\* *p* < 0.001

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Focal model** | | **a2** | | **a3** | | **a4** | | **a5** | | **a6** | | **a7** | | **a8** | |
| INTERCEPT | 533.710 | \*\* | 280.992 | \*\* | 529.338 | \*\* | 226.579 | \*\* | 510.358 | \*\* | 371.876 | \*\* | 497.676 | \*\* | 369.209 | \*\* |
| S\_SEX | -0.002 | \* | 30.017 | \*\* | -0.002 | \* | 30.917 | \*\* | -0.013 |  | 17.149 | \*\* | -0.018 |  | 16.635 | \*\* |
| IS2G30 | -0.001 |  | 44.372 | \*\* | -0.001 |  | 44.055 | \*\* | 0.000 |  | 8.078 |  | -0.008 |  | 10.374 |  |
| S\_USESTD | 0.000 |  | 0.472 |  | 0.000 |  | 0.424 |  | 0.001 |  | -0.286 |  | 0.001 |  | -0.295 |  |
| S\_ICTLRN | 0.000 |  | 0.738 | \*\* | 0.000 |  | 0.690 | \*\* | -0.001 |  | 0.598 | \*\* | -0.002 |  | 0.647 | \*\* |
| S\_ICTFUT | 0.000 | \* | 0.902 | \*\* | 0.000 | \*\* | 0.831 | \*\* | 0.000 |  | 0.469 | \* | 0.000 |  | 0.407 | \* |
| S\_USECOM | 0.000 |  | 0.403 |  | 0.000 |  | 0.383 |  | 0.000 |  | 0.052 |  | -0.001 |  | 0.214 |  |
| ICT\_INTO | 0.000 |  | 13.337 | \*\* | 0.000 |  | 14.847 | \*\* | -0.009 |  | 13.694 | \*\* | -0.008 |  | 14.104 | \*\* |
| ICT\_INTS | 0.000 |  | -3.624 |  | -0.001 |  | -3.775 |  | 0.007 |  | -10.539 | \* | 0.005 |  | -12.362 | \*\* |
| S\_USEINF | 0.000 |  | -0.801 | \* | 0.000 | \* | -0.812 | \* | 0.001 | \* | -0.828 | \*\* | 0.002 | \*\* | -1.087 | \*\* |
| P\_NUMSTD\_CAT | 4.996 |  | 12.145 | \* | 4.791 |  | 11.997 | \* | 5.203 |  | 10.924 |  | 6.252 |  | 8.925 |  |
| IP2G07 | 3.311 |  | -1.801 |  | 3.440 |  | -0.021 |  | 7.115 |  | 7.778 |  | 6.957 |  | 7.540 | \* |
| C\_RATTEA | -2.436 |  | -1.323 |  | -1.743 |  | -0.468 |  | 0.188 |  | 2.959 |  | -0.367 |  | 2.548 |  |
| C\_RATSTD | -0.316 |  | -0.451 |  | -0.259 |  | -0.387 |  | -0.157 |  | -0.111 |  | -0.207 |  | -0.103 |  |
| T\_ICTPRAC | 0.890 |  | 1.490 |  | 0.945 |  | 2.323 |  | -0.451 |  | -0.047 |  | -0.436 |  | -0.040 |  |
| T\_RESRC | -0.277 |  | -0.755 |  | -0.242 |  | -0.637 |  | 0.995 | \*\* | 0.869 |  | 0.884 | \*\* | 0.999 | \* |
| Missing data | No imputation | | No imputation | | With imputation | | With imputation | | No imputation | | No imputation | | With imputation | | With imputation | |
| Countries | Korea | | Korea | | Korea | | Korea | | All countries | | All countries | | All countries | | All countries | |
| Centring | Centring | | No centring | | Centring | | No centring | | Centring | | No centring | | Centring | | No centring | |
| # Missing | 305 missing | | 305 missing | | 0 missing | | 0 missing | | 10593 missing | | 10593 missing | | 0 missing | | 0 missing | |
| # Students | 2513 students | | 2513 students | | 2818 students | | 2818 students | | 24338 students | | 24338 students | | 34931 students | | 34931 students | |
| # Schools | 135 schools | | 135 schools | | 147 schools | | 147 schools | | 1491 schools | | 1491 schools | | 1806 schools | | 1806 schools | |
| # Countries | 1 country | | 1 country | | 1 country | | 1 country | | 13 countries | | 13 countries | | 13 countries | | 13 countries | |

In sum, the results of Kim et al. (2014) were not replicated. However, it appears that analytical variability has a large influence on the main result, as seen in Table 2. We therefore focus the ‘discussion’ section on discussing this variability.

# Discussion

Now that we have reported in-depth on the replication process for Kim et al. (2014), as well as the results, it is pertinent to discuss some of the challenges in secondary data replications. First and foremost, it is clear that the analytical pipeline contains an enormous amount of variability. In this discussion section we discuss many of ingredients of that variability, linking it to prior literature on replication attempts.

Firstly, the choice of secondary dataset already determines many decisions in the analysis pipeline. For example, in this case, the choice of the ICILS 2018 dataset pertained to a different age of participants, and, given that Siddiq and Scherer (2019) had already highlighted the point that effects in primary school were larger than those at secondary age, this might already be a problematic choice. The sampling of ‘whole classes’ or not, is another difference and/or decision point. On the one hand, sampling whole classes might give insight into classroom and teacher processes, but, on the other hand, the individual students will not be statistically independent. The choice of ‘levels’ (i.e., classroom and/or school) in the multilevel models will also be associated with the extent to which students are seen as ‘independent’: students in one class often experience specific classroom effects. Furthermore, the choice of dataset also determines the extent of missing data and choices to be made regarding missing data treatment, as well as with regard to whether a complex sampling design needs to be taken into account. Here, as described previously, plausible values, weights and resampling techniques need to be taken into account, and although data analysis manuals often have one recommended approach, in practice this is less straightforward, with estimation techniques and software packages contributing to numerous different practices concerning missing data.

Then, as indicated earlier and in Table 1, the variables themselves might be worded differently, have different answer options, and be able to be combined in scales in different ways; the operationalisation of variables plays an important role. So, although the positive aspect of this is that data have already been collected, the challenge lies in whether they really can be seen as covering similar constructs. Many of the choices relate to the way we see and interpret replications. On the one hand, replications can never be completely the same as in the original study, as studies are highly contextualised in both time and space. On the other hand, we *are* interested in looking at bodies of knowledge that enable us to make inferences about certain topics, for example, in this case, differences in ICT literacy. In other words, the aims of the replication are important, and we cannot use simplistic typologies for replications.

In the context of educational research, Plucker and Makel (2021) summarise types of replications. In a *direct replication*, researchers attempt to follow the original study’s methods as closely as possible in an effort to arrive at similar results. However, ‘as closely as possible’ is somewhat opaque, though, as is the case in the current replication, the choice of the ICILS dataset alone made it harder to adhere to some of the original study’s features. The present study cannot really be called a conceptual replication, summarised by Plucker and Makel (2021) as focussed on “the theoretical soundness of a particular finding or set of findings, with less focus on repeating exact methods from the original study” (p. 90). That also does not fit the present study very well, as the emphasis *is* on adopting similar methods, but, especially with secondary datasets, there simply might be differences in the way data have been collected, variables operationalised and sampling designs. And what do we do with these differences? Schmidt (2017), for example, posits that a single observation cannot be trusted and that a replication can contribute to an observation transforming into a fact or a piece of knowledge.

Plucker and Makel (2021) give more examples of where direct replications are seen as the only way to verify the reliability of results. But again, can they even be compared if there *are* differences. Are they comparing like with like? Failed replication attempts have seen unproductive tos-and-fros regarding methods being replicated ‘as closely as possible’[[18]](#endnote-18); with every sample uniquely rooted in contexts such as country, age and time, can we really expect similar results? Further, even if we are interested in what Huntington-Klein et al. (2021) call *pure* replication (or “reproduction”) studies (a new study is performed purely to check the results of a prior study using the same data and methods), it certainly is not a given that results are easy to replicate, even with the increasing availability of analysis code and data.

Perhaps the typology by Hüffmeier et al. (2016) provides more nuance, with its distinction between exact replication (direct replication conducted by the same researchers), close replications (also direct but by different researchers), constructive replications (also direct, but being a similar study modified in a small number of ways in order to assess the robustness of the original effect), conceptual replications under laboratory conditions (a conceptual attempt to study theory), and conceptual replications under field conditions (also conceptual and attempting to study the robustness of a theoretical effect). Our example again crosses several categories, indicating that there are many variations across a dimension from ‘completely the same’ (arguably impossible) to ‘measuring the same construct’ (which would need to be evidenced). By necessity, some methods might be able to be replicated, but not all can be. The choice of (secondary) dataset in this case concerns different field conditions.

We agree with Plucker and Makel (2021) in arguing that, regardless of the label, explicit intent to replicate findings is a key factor here, adhering to original methods and yet diverging from them, addressed by providing *transparency* in disclosing every facet of the study. However, in doing so, we might recognise, as in this replication, that the analytical variability is substantial, with one (different) secondary data decision leading to many other decisions to be made, based on the available choices. This is in line with other research on such ‘forking paths’ (Gelman & Loken, 2014).

Huntington-Klein et al. (2021) used a ‘many analysts’ approach to measure the extent and impact of hundreds of decisions about data collection, preparation, and analysis of research in applied microeconomics, and they concluded that there were “large differences in data preparation and analysis decisions, many of which would not likely be reported in a publication” (p. 944). Replicators reported different sample sizes, different standard deviations, varying statistical significance, and, in one case, even a reversed outcome. Huntington-Klein et al. (2021) already highlight this in the initial stages of data cleaning, as they also allowed analysts to make their own data cleaning decisions in that phase. In secondary data analysis, the choice of the dataset already determines much, as we have seen in our own study, e.g., the requirement to take into account the complex sampling design whilst analysing. Data analysis examples from other disciplines also show the extent to which analytical decisions can affect the outcomes of analyses.

In a study on manufacturers’ productivity, White et al. (2018) showed that whether imputation was used, and, if so, the decision on imputation, had a dramatic effect on the outcome measures. In the present study, there were missing data which we did not impute in the focus of the analysis but imputed in others. Although the differences perhaps were not ‘dramatic’, there were substantial differences in some predictors, most notably in the larger (and thus more missing data) ‘all countries’ analyses. The choice of subsamples also is a variable, in this replication represented by ‘Korea’ or ‘all countries’. Clemens and Hunt (2019), for example, showed that results on the effects of immigration were influenced by the racial composition of subsamples. Silberzahn et al. (2018) recruited psychological researchers to analyse whether a given data set of referee calls in soccer games showed evidence of discrimination. Analysts showed differences in, among others, the types of regression, dependency of error terms, and choice of covariates, leading to effect sizes (e.g., odds-ratio) varying from .89 to 2.93, even after peer review. They concluded that “significant variation in the results of analyses of complex data may be difficult to avoid, even by experts with honest intentions” (p. 338). A re-analysis of these data by Auspurg and Brüderl (2021) argues that the main reason for the analytical differences was an unclear research question, and that social science research “needed to be more precise in its ‘estimands’ to become credible” (p. 1).

For ‘functional magnetic resonance imaging’ (fMRI) data for a gambling task, Botvinik-Nezer et al. (2020) noted that replication teams differed in image processing, statistical mapping of activated brain regions, and other parts of the pipeline from data through to analysis, again leading to different conclusions with regard to the hypotheses being tested. For an alternative approach, in which the hypothesis was kept constant but the research design and data differed, Landy et al. (2020) asked several independent research teams to test hypotheses about moral judgements, negotiations and implicit cognition. Again, results varied, with effects of opposite signs, and with the variability being unrelated to the skill of the research teams. In a study by Breznau et al. (2021), 162 researchers in 73 teams tested the same hypothesis with the same data. The outcomes could be explained mostly by variations in researchers’ modelling decisions or prior beliefs: each of the 1,261 test models was a unique combination of data-analytical steps. Highly skilled scientists motivated to come to accurate results varied enormously and such “idiosyncratic researcher variability” is a threat to the reliability of scientific findings (Breznau et al., 2021, p. 2).

In organisational research, Schweinsberg et al. (2021) reported that researchers used radically different analyses and arrived at different outcomes, in some cases with opposite conclusions. In their study, especially with regard to differences in the operationalisation of variables, such differences explained variability above and beyond decisions which statistics to use. In the present replication of this paper, we also saw that the way in which variables were operationalised (i.e., Table 1) involved some arguably arbitrary decisions. Bastiaansen et al. (2020) in psychosomatic research, found variation evident at different analysis stages, from decisions early on, e.g., in pre-processing (variable selection, clustering, missing data treatment) to the types of statistics being used, leading to varying recommendations.

Christensen and Miguel (2018) note that data often are unavailable, even with data-sharing policies. Even if data and scripts are available, results cannot always be reproduced, e.g., because of data cleaning and analysis decisions being either in error, not sufficiently described, or unhelpfully lengthy in the case of a full description.

One might say that, at least with secondary data analysis, there are apparently no data collection issues, as the data are the data; in many cases a simple agreement with terms of use will suffice. Many datasets come with codebooks and information on the specifics of variables. In the present replication, a clear focus of the analysis was decided, even though other analysis paths existed. Seven other paths are included in this article, but this still begs the question of which one is ‘right’ and which one is ‘wrong’, and how to decide and judge here. Even if we simply admit that they are all right but ‘just different decisions’, how does this help us in drawing conclusions? The answer is basically ‘it depends’, and it depends on what research questions you are trying to answer, with often multiple paths leading to a reasonable answer.This matters, because even the decision on the framing and content of a hypothesis can have an impact on the study.

If analytical choices are not wrong, but simply one reasonable option out of many, Simmons et al. (2011), for example, contend that researcher flexibility among accepted options in research design and data analysis allow nearly any hypothesis to be supported. And there are many more decisions to make, as seen in this replication, and from further literature. For example, by adding controls, statistical results could become statistically insignificant (Lenz & Sahn, 2017), which is one reason why, in multilevel modelling, it might be valuable to adopt the recommendations by Dedrick et al. (2009) to report the modelling process transparently and fully. Another example is software use. For open practices, the present replication chose to work with open source packages in R, but many analyses use different software packages. In economics, McCullough and Vinod (2003) found that choice of software package affected studies’ results.

In five multilevel software packages McCoach et al. (2018) found that “noticeable differences among the five packages arose in terms of speed, convergence rates, and the production of standard errors…” (p. 594). Similarly, as in the present replication, how to deal with missing data could vary. Martel Garcia (2014) found different outcomes with different imputation methods, with the most appropriate methods being dependent on the data. Analytical decisions will also be influenced by preferences built up over a career, for example familiarity with a given model, analysists’ ideas about the most appropriate control variables and models, and preferences for parsimony (Huntington-Klein et al., 2021, p. 958).

The discussion has shown that, in previous studies and also in the present secondary data replication, substantial analytical variability exists, including, for example: the choice of secondary dataset; the associated complex sampling design with variation in how to tackle weights and plausible values (e.g., Rutkowksi et al., 2010; Drent et al., 2013 for ILSA data); and the available variables and measurements plus constructed scales. The present replication study raised issues of:

* choosing the subsample (one country or more countries), and it was noted that this influences the choice of analytical strategy (e.g., multilevel models with two or three levels);
* missing data and their treatment (e.g., single imputation, multiple imputation);
* the way in which models are built, including control variables (for a checklist for multilevel models see Dedrick et al., 2009; Lenz & Sahn, 2017);
* centring of variables;
* the choice of software (see McCoach et al., 2018 for multilevel models and the analytical approach in this replication); and
* (ethical) choices regarding availability with or without payment.

Further, issues are raised of changing the context over time (in the present study, for example the Korean school curriculum (e.g., Jeon et al., 2020), with time-lags in educational effects also coming into play). So, substantively, although the main gender result was not replicated in the present study, it could be argued that the replication study included several analytical decisions that could have contributed to having the main result turn out differently.

What can researchers and educationists do if the social science that is conducted can yield different results because of one change? One approach might be to build out many options and possible differences, e.g., by prescribing standard procedures and analyses, a so-called multi-verse analysis (Steegen et al., 2016) or a specification curve (Simonsohn et al., 2020). Although originally perhaps more focussed on data processing (Steegen et al., 2016), this can be extended to data collection (e.g., Harder, 2020) and data analysis (e.g. Mirman et al., 2021), i.e., to start to have more and more prescriptive coverage of the whole research pipeline. In situations in which decisions are unclear, multiverse-style analyses could be conducted in a more exploratory fashion (Del Giudice & Gangestad, 2021). Such analyses could be aided by transparent publication of data and analysis scripts, the latter with suitable documentation and commentary, so that data analytic choices and decisions are clear. Packages to aid researchers’ workflows with this are also appearing[[19]](#endnote-19).

# Conclusion

This article reported the secondary data replication of Kim et al.’s (2014) study of computer literacy. Although the main result was not replicated, it was also observed that that substantial analytical variability exists in the analysis pipeline. Not having to collect primary data can save time, money and having to make some analytical decisions, as some things have already been decided. However, new variability can be introduced by the choice of dataset, and, as the present study finds, such analytical variability matters and can lead to different outcomes. With little hope of a convergence of practices[[20]](#endnote-20), maybe transparency is the best option. For example, although preregistration does not solve the issue of analytical variability (Huntington-Klein et al., 2021, p. 2), it might be that, if the preregistration is suitably detailed and the final reporting is as transparent as possible, this at least shows what paths have been taken when analytical options present themselves. Brandt et al. (2014) argued for a ‘replication recipe’ for close replications, which contains numerous ingredients, such as: carefully defining the effects and methods that the researcher intends to replicate; following the exact methods of the original study; having sufficiently high statistical power; making replication details available; and evaluating replication results and comparing them critically to the results of the original study. The present replication study has striven to accomplish this,e.g., by pre-registering the analytical decisions and providing a documented analysis script, within the context of the SCORE project[[21]](#endnote-21). As is perhaps often the case with social science, researchers might be left with more rather than fewer questions, but the present study hopefully can shine a light on them by emphasising the need for transparency.

# Acknowledgements

The authors approached the ‘data finder’ for this replication, but did not receive a response, so that person cannot be named, though thanked here. The authors thank the SCORE project team at the Center for Open Science and the reviewers of the preregistration.

# References

Auspurg, K., & Brüderl, J. (2021). Has the credibility of the social sciences been credibly destroyed? Reanalyzing the “Many Analysts, One Data Set” project. *Socius*, *7*, 1–4.https://doi.org/10.1177/23780231211024421

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., ... & Bringmann, L. F. (2020). Time to get personal? The impact of researchers’ choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, 110211. https://doi.org/10.1016/j.jpsychores.2020.110211

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R. G., … & Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Breznau, N., Rinke, E. M., Wuttke, A., Adem, M., Adriaans, J., Alvarez-Benjumea, A., ... & van der Linden, M. (2021). Observing many researchers using the same data and hypothesis reveals a hidden universe of data analysis. MetaArXiv. https://doi.org/10.31222/osf.io/cd5j9

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920–980. https:// https://doi.org/10.1257/jel.20171350

Clemens, M.A., & Hunt, J. (2019). The labor market effects of refugee waves: reconciling conflicting results. *ILR Review*, *72*(4), 818–857. https://doi.org/10.1177%2F0019793918824597

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., ... & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*(1), 69–102. https://doi.org/10.3102%2F0034654308325581

Del Giudice, M., & Gangestad, S. W. (2021). A traveler’s guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, *4*(1), 1–15. https://doi.org/10.1177/2515245920954925

Drent, M., Meelissen, M. R., & van der Kleij, F. M. (2013). The contribution of TIMSS to the link between school and classroom factors and student achievement. *Journal of Curriculum Studies*, *45*(2), 198–224. https://doi.org/10.1080/00220272.2012.727872

Fiske, S. T. (2016). How to publish rigorous experiments in the 21st century. *Journal of Experimental Social Psychology*, *66*, 145–147. https://doi.org/10.1016/j.jesp.2016.01.006

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020a). *Preparing for life in a digital world: IEA International computer and information literacy study 2018 international report*. Springer Nature.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020b). *IEA International Computer and Information Literacy Study 2018. Technical Report*. International Association for the Evaluation of Educational Achievement.

Gebhardt, E., Thomson, S., Ainley, J., & Hillman, K. (2019). *Gender differences in computer and information literacy: An in-depth analysis of data from ICILS*. Springer Nature.

Gelman, A., & Loken, E. (2014) The statistical crisis in science: Data‐dependent analysis—“a garden of forking paths”– explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*(6), 460–466. https://www.americanscientist.org/article/the-statistical-crisis-in-science

Harder, J. A. (2020). The Multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, *15*(5), 1158–1177. https://doi.org/10.1177%2F1745691620917678

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology, 66*, 81–92. https://doi.org/10.1016/j.jesp.2015.09.009

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J., Burli, P. H., ... & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*(3), 944–960. https://doi.org/10.1111/ecin.12992

Jeon, S., Son, Y., & Park, S. (2020). Analysis of the ICILS 2018 results by Korean students' educational experience in computer and information literacy and computational thinking. *The Journal of Korean Association of Computer Education*, *23*(3), 1–8. https://doi.org/10.32431/kace.2020.23.3.001

Kim, H. S., Kil, H. J., & Shin, A. (2014). An analysis of variables affecting the ICT literacy level of Korean elementary school students. *Computers & Education*, *77*, 29–38. https://doi.org/10.1016/j.compedu.2014.04.009

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., ... & Crowdsourcing Hypothesis Tests Collaboration. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479. https://doi.org/10.1037/bul0000220

Lenz, G.S., & Sahn, A. (2017). *Achieving statistical significance with covariates and without transparency*. Technical Report. University of California Berkeley.

Martel García, F. (2014). Democracy is good for the poor: A procedural replication of Ross (2006). *Research & Politics*, *1*(3), 1–10. https://doi.org/10.1177/2053168014559094

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., ... & Bellara, A. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics*, *43*(5), 594–627. https://doi.org/10.3102%2F1076998618776348

McCullough, B.D., & Vinod, H.D. (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, *93*(3), 873–892. https://doi.org/10.1257/000282803322157133

Mirman, J. H., Murray, A. L., Mirman, D., & Adams, S. A. (2021). Advancing our understanding of cognitive development and motor vehicle crash risk: a multiverse representation analysis. *Cortex*, *138*, 90–100. https://doi.org/10.1016/j.cortex.2021.01.017

Plucker, J. A., & Makel, M. C. (2021). Replication is important for educational psychology: Recent developments and key issues. *Educational Psychologist*, *56*(2), 90–100. https://doi.org/10.1080/00461520.2021.1895796

Punter, R. A., Meelissen, M. R., & Glas, C. A. (2017). Gender differences in computer and information literacy: An exploration of the performances of girls and boys in ICILS 2013. *European Educational Research Journal*, *16*(6), 762–780. https://doi.org/10.1177%2F1474904116672468

Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. John Wiley.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102%2F0013189X10363170

Schmidt, S. (2017). Replication. In M. C. Makel & J. A. Plucker (Eds.), *Toward a more perfect psychology: Improving trust, accuracy, and transparency in research* (pp. 215–232). American Psychological Association.

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R., Van Assen, M. A., ... & Schulte-Mecklenbeck, M. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249. https://doi.org/10.1016/j.obhdp.2021.02.003

Siddiq, F., & Scherer, R. (2019). Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. *Educational Research Review*, *27*, 205–217. https://doi.org/10.1016/j.edurev.2019.03.007

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … & Nosek, B. A. (2018). Many Analysts, One Data Set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False‐positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177%2F0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9),160384. https://doi.org/10.1098/rsos.160384

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177%2F1745691616658637

White, T.K., Reiter, J.P., & Petrin, A. (2018). Imputation in US manufacturing data and its implications for productivity dispersion. *Review of Economics and Statistics*, *100*(3), 502–509. https://doi.org/10.1162/rest\_a\_00678

Wu, Y. W., & Wooldridge, P. J. (2005). The impact of centering first-level predictors on individual and contextual effects in multilevel data analysis. *Nursing research*, *54*(3), 212–216. https://doi.org/10.1097/00006199-200505000-00009

1. The abstract and the findings in the text contradict one another. The abstract mentions that being a male student is a positive predictor of ICT literacy score, but the text mentions that “the ICT literacy test score of female students was 2.543 points higher than that of male students” (Kim et al., 2014, p. 34), with the coefficient in Table 6 being negative for male students, as gender was coded as female=0, male=1. [↑](#endnote-ref-1)
2. Non-simulated research refers to research that has original data collection. Data that are created by simulations and algorithms are not included. [↑](#endnote-ref-2)
3. Despite the aforementioned discrepancy in the text of the abstract, the meta-analysis included Kim et al. (2014) included studies showing higher achievement for females. [↑](#endnote-ref-3)
4. The researchers used a three-dimensional structure for CIL with dimensions: 1. ‘Applying technical functionality’, 2. ‘Evaluating and reflecting on information’ and 3. ‘Sharing or communicating information’. Girls outperformed boys in most countries on the second dimension, and no statistically significant gender differences were found for Dimension 1. The largest differences in favour of girls were on Dimension 3. [↑](#endnote-ref-4)
5. The R scripts and associated analysis files that are made available take the original ICILS files from the IEA in the raw SPSS format and apply all the necessary data processing and analyses. They are available on the Open Science Framework (OSF). [↑](#endnote-ref-5)
6. According to the authors, strata were based on regional scale (Kim et al., 2014, p. 32). [↑](#endnote-ref-6)
7. Details of this procedure are reported in the technical report (Fraillon et al., 2020, chapter 6). Exclusions, for example, could take place in cases of physical or intellectual disability, or in cases of non-native language speakers without the language proficiency to complete the assessment. [↑](#endnote-ref-7)
8. Students do not complete all assessment items, but only a selection of items from a booklet. However, reducing test length and administering subsets of items, introduces uncertainty in the estimates at the individual level. The problem is addressed by treating a student’s ability estimate as a missing data problem and adopting a ‘plausible values’ methodology that uses all available information from student tests and questionnaires to impute an ability estimate. For more details see the ICILS 2018 technical report (Fraillon et al., 2020b, chapter 11). [↑](#endnote-ref-8)
9. It is important to note that we did not transform the original ICILS 2018 data and that therefore, for the interpretation of the gender effect, we are looking for a ‘reverse’ effect, as the reference categories are different. [↑](#endnote-ref-9)
10. Wordings throughout the table have been kept as similar to the original source as possible. [↑](#endnote-ref-10)
11. The schoolID variable had to be created because in ICILS 2018 data the school ID is not unique across countries. This is a combination of country and schoolid, making a unique schoolid. For the focal analysis with Korea alone, this was not important, but it was necessary for the supplemental analyses with multiple countries. [↑](#endnote-ref-11)
12. The R code is available at ANONYMISED LINK. The code includes (slightly modified) code from the data finder. Most lines have ample documentation. The code includes ‘checks’ (with lines that don’t necessarily analyse anything), for example regarding missing data. [↑](#endnote-ref-12)
13. The use of R packages in itself already is a deviation from the original study which used HLM 6.0 software. [↑](#endnote-ref-13)
14. An R package with imputation methods is available at: <https://cran.r-project.org/web/packages/simputation/index.html> [↑](#endnote-ref-14)
15. Results of these analyses can be replicated by: 1. Obtaining raw ICILS data from the IEA. 2. Putting these, the R script and the helper files (with variable names) into one project.

    3. Running the script ANONYMISED LINK. The script will produce a table of estimates for all eight analyses, including the focal analysis. [↑](#endnote-ref-15)
16. A podcast on this topic is available at <https://www.podbean.com/ew/dir-rbj6c-db707d0> [↑](#endnote-ref-16)
17. See the tab ANALYSIS1\_FOCAL of the output file estimates.xlsx at ANONYMISED LINK [↑](#endnote-ref-17)
18. For example, see Fiske’s (2016) commentary for a special issue on the replication crisis, and the blog by Ulrich Schimmack: <https://replicationindex.com/2020/02/12/fiske-and-the-permanent-crisis-in-social-psychology/> [↑](#endnote-ref-18)
19. For example, the R packages Workflowr and steveproj. However, the goal of having a standardised approach can also lead to a divergence of practices (‘yet another package’). [↑](#endnote-ref-19)
20. Perhaps practices can only improve over generations, across academia, starting at the basis, e.g., Smaldino and McElreath (2016). [↑](#endnote-ref-20)
21. More information on the SCORE project can be found at <https://www.cos.io/score> [↑](#endnote-ref-21)