A Weight Trimming Approach to Achieve a Comparable Increase to Bias across Countries in the Programme for the International Assessment of Adult Competencies

Wendy Van de Kerckhove, Leyla Mohadjer, Thomas Krenzke
Westat, 1600 Research Blvd, Rockville, MD 20815

Abstract
Weight trimming can be used to reduce sampling variance and the impact of influential cases. However, it can also introduce bias into the survey estimates. In surveys with an emphasis on comparing estimates across countries or states, it is important to pay special attention to the amount of trimming carried out in each country. This was the case for the Programme for the International Assessment of Adult Competencies (PIAAC), an international adult literacy survey. To limit the number of cases trimmed and at the same time achieve a comparability across countries, a modification was made to the common method of trimming weights that are over *k* times the median. Rather than using a constant value of *k*, the factor was based on a function of the coefficient of variation (CV) of the country's weights. This paper describes the trimming procedure for PIAAC and reviews some alternative trimming procedures, together with an evaluation of their effects on bias and variance.

Key Words: Weighting, extreme weights, calibration

1. Introduction

Extreme weights can result in inflated variances and be influential, particularly when associated with extreme data values or when producing estimates for subgroups. One common approach for addressing extreme weights is weight trimming, that is trimming weights over a specified cut-point to the cut-point. The goal of trimming is to reduce the mean squared error (MSE) of the estimates. Trimming procedures can introduce some bias. However, as discussed in Lee (1995), the trimming adjustment will generally reduce the sampling error component of the overall MSE more than it increases the bias when the adjustment is applied to only a very small number of weights.

This paper reviews the trimming procedures used in the Programme for the International Assessment of Adult Competencies (PIAAC), and some other cross-country or cross-state surveys. PIAAC is an in-person literacy survey of non-institutionalized adults ages 16 to 65, sponsored by the Organisation for Economic Cooperation and Development (OECD). The survey assesses the proficiency of adults in literacy, numeracy, and problem-solving in technology-rich environments. Twenty-four countries participated in the first round of data collection which occurred between 2011 and 2012, and nine other countries are taking part in the second round in 2014.

Since the PIAAC data are used to make comparisons between countries, it is important to achieve a consistently high level of quality. To this end, the PIAAC Consortium established a series of Technical Standards and Guidelines (OECD, 2014). The document covers all aspects of the survey and is intended to produce data that are reliable and comparable across countries. Any methods developed for trimming should also meet these objectives.

Section 2 provides some background on weight trimming, with a focus on methods used by other surveys that produce weights for multiple countries or multiple states. Reasons for weight variation in PIAAC, along with the trimming method chosen to address such variation, are given in section 3. To take into account the variable dispersion of the weights across countries, it was decided to use a trimming cut-point that incorporates the coefficient of variation (CV) of the weights prior to trimming. Limiting the number of cases trimmed will restrict the magnitude of the potential bias introduced by trimming. An evaluation was performed to compare the PIAAC method to other common trimming methods and results are discussed in section 4. A final discussion is given in section 5.

2. Overview of Trimming Procedures

In weight trimming, weights exceeding a specified cut-point are trimmed to that value. The trimmed weight can be expressed as:

$$w_{jt} = \begin{cases} w_0 \text{ if } w_j > w_0 \\ w_j \text{ otherwise} \end{cases} \quad (1)$$

where w_j is the weight prior to trimming, and w_0 is the trimming cut-point. Often an additional adjustment is then performed to bring the sum of weights back up to the level prior to trimming. If the sample design involves the over- or under-sampling of certain domains (e.g., oversampling minorities), then some weights will be large by design and are not considered outliers. To account for this design feature, trimming is performed separately by sampling domain.

2.1 Examples of trimming procedures

Trimming procedures differ in their choice of cut-point. A wide range of options exists in the literature. Table 1 provides some examples of trimming procedures and the cut-point used by each. One common approach is the *k**median rule. Under this approach, weights greater than a constant (*k*) multiplied by the median weight are trimmed back to the cut-point. The constant is typically chosen to be around 3 or 4. The mean weight or an ideal weight[1] is sometimes used in place of the median. Another approach is the inter-quartile range (IQR) method, which attempts to control the trimming by considering the variation in the weights. When determining a cut-point, this method uses the median of the weights as well as the spread, as measured through the IQR.

Figure 1 provides an example of the difference between the *k**median and IQR rules. The figure shows a SAS® box-and-whisker plot for two sets of weights, each having a median of 14 and a maximum weight of 45. The median is indicated by the center line in

[1] The ideal weight is the weight that would have been assigned if the frame measure-of-size (MOS) for the primary sampling units (PSUs) in a two-stage probability proportional to size (PPS) sample had been accurate and there had been no nonresponse.

Table 1: Examples of Trimming Procedures

Trimming procedure	*Cut-point*	*Definition of terms*
*K**median	k**median*(w_j)	k = constant
(or *k**mean	(or k**mean*(w_j)	w_j = weight prior to trimming
or *k**(ideal weight))	or k*w_{ideal})	(w_{ideal} = the ideal weight)
Interquartile range	*median*(w_j) + k**IQR*(w_j)	*IQR* = inter-quartile range
Contribution to entropy	$\sqrt{k\sum w_j^2/n}$	n = sample size
Weight distribution	w_{0p}, where 1 - $F(w_{0p}) = p$	$F(\ )$ = cumulative distribution function
		p = specified probability of occurrence
Estimated MSE	w_{0m}	w_{0m} = the cut-point resulting in the lowest mean-squared error

the box. The IQR is 21 for the first set of weights and 6.5 for the second set, as indicated by the length of the box. The whiskers of the boxplot extend to the minimum and maximum weights within the "fence". By default, SAS® defines the upper fence boundary as the third quartile plus 1.5 times the IQR, and outlying values are shown as a square dot. The weight of 45 in the second scenario is an outlier. For k = 3, the cut-points for the k*median approach are indicated in orange in the plot, and those for the IQR approach are in green. The k*median procedure would trim the weight of 45 under both scenarios, whereas no trimming would occur for the first set of weights under the IQR procedure.

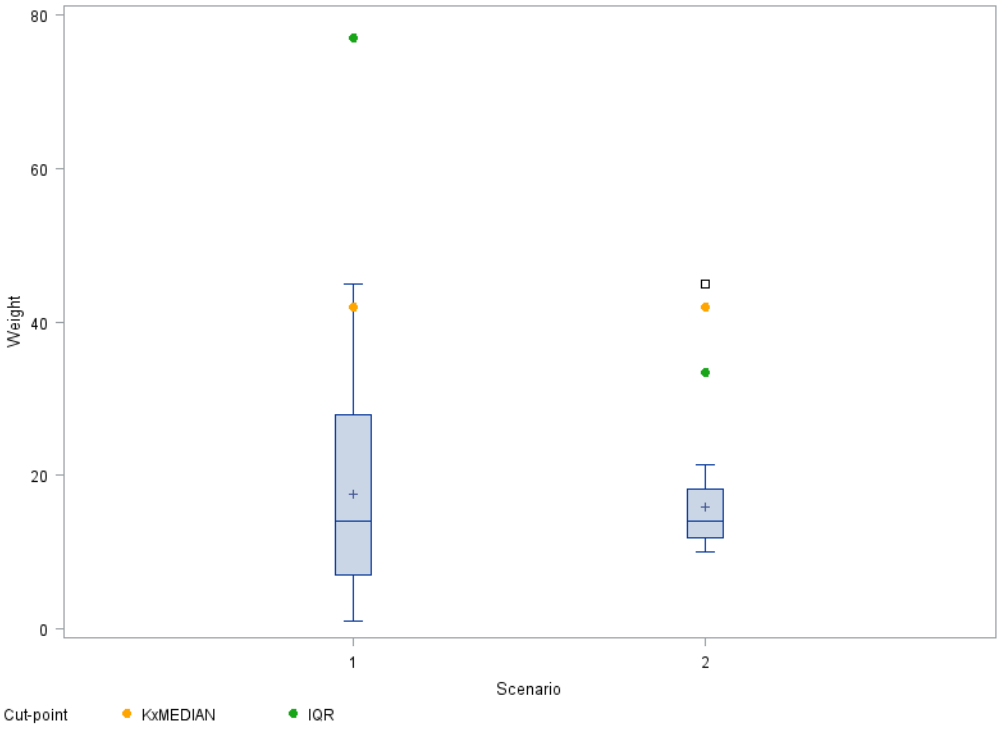

Figure 1: Example of k*median and IQR cut-points for two different weight distributions

Potter (1998, 1990) provides summaries of the other trimming methods in Table 1. One example is contribution-to-entropy which trims by iteratively evaluating the contribution of each weighted observation to the overall variance of the weighted estimate. Another is the weight distribution method, in which the weights are assumed to have a particular probability distribution, such as inverse beta, and large weights are trimmed if they have a low probability of occurrence.

None of the approaches discussed thus far assess the MSE in the outcome estimates. An alternative approach is to try various cut-points and calculate the MSE of the key statistic at each cut-point, choosing the cut-point that results in the lowest MSE. This may not be feasible if the outcome data are not available at the time of weighting, or if there are multiple outcomes of interest with different optimal cut-points. In addition, the bias can be difficult to estimate, which is why the previous approaches focus more on the variance.

In addition to weight trimming, weight modeling (or smoothing) approaches are available to address weight variation. See, for example, Elliott (2008). Alternatively, calibration estimators (Deville and Särndal, 1992) can be used to limit the amount that the weights are adjusted during the weighting process. Such methods are not the focus of this paper.

2.2 Trimming procedures in some cross-country or cross-state surveys

To investigate which trimming procedures are used in practice, a web search was conducted on other surveys that produce weights for multiple countries or multiple states. Documentation was reviewed for the following (accessed August 27, 2014):

- 2011 Trends in International Mathematics and Science Study (TIMSS) (http://timss.bc.edu/methods/pdf/TP_Sampling_Design.pdf
- 2009 Program for International Student Assessment (PISA) (http://www.oecd.org/pisa/pisaproducts/50036771.pdf)
- 2007 National Assessment of Educational Progress (NAEP) (http://nces.ed.gov/nationsreportcard/tdw/weighting/2007/weighting_2007_trimming_adjustments.aspx)
- 2009-2010 National Adult Tobacco Survey (NATS) (http://www.cdc.gov/tobacco/data_statistics/surveys/nats/pdfs/weighting-specs.pdf)
- 2012 Behavioral Risk Factor Surveillance System (BRFSS) (http://www.cdc.gov/brfss/annual_data/2012/pdf/compare_2012.pdf4)
- 2012 European Social Survey (ESS) (http://www.europeansocialsurvey.org/docs/methodology/ESS_post_stratification_weights_documentation.pdf)

This list contains two international school surveys (TIMSS and PISA), one national school survey that produces weights for multiple states (NAEP), two telephone surveys that produce weights for multiple states (NATS and BRFSS), and one international household survey (ESS).

Of the six surveys, TIMSS was the only one that did not perform trimming. In NATS, trimming is based on the IQR procedure described in section 2.1. Trimming is applied to the state-level weights before the weights are poststratified to population totals.

The remainder of the surveys trim based on some variation of *k* times the median, mean, or ideal weight. The two school surveys (PISA and NAEP) sample schools and then students within schools. Weighting adjustments are made at the school and student level to produce the final student weights. In PISA, weight variation is primarily a result of inaccurate measure-of-size (MOS) information on the school frame. Some variation is added through the weighting adjustments to address nonresponse at the school and student levels. The trimming is performed within explicit strata, using a cut-point of *k* times the ideal weight. The value of *k* is typically 3 for school weights and 4 for student weights. Similarly, the trimming in NAEP uses a cut-point of 3 times the ideal weight for schools and *k**median for students, where $k = 3.5$ for students in public schools and 4.5 for students in private schools. In both surveys, only a small percentage of cases are trimmed.

BRFSS applies the Individual and Global Cap Value (IGCV) and Margin Cap Value (MCV) methods described in Izrael, Battaglia, and Frankel (2009). The authors indicate that they primarily apply the IGCV method, but MCV is useful when IGCV does not converge. In IGCV, trimming and raking are carried out as an iterative process. Weights are trimmed if they exceed *k* times the mean weight (global bound), or if an individual weight after raking is larger than some constant times the value before weighting (individual bound). This method also applies lower bounds on the weights. The documentation did not give the value of *k* used for BRFSS. Finally, the ESS survey trims based on 4*mean when producing poststratification weights.

The documentation for PISA, NAEP, NATS, and BRFSS all indicate that the trimming procedures were implemented to identify extreme weights and reduce variance and MSE. However, no reason is provided why a particular trimming method was chosen over another. The following section will explain the motivation and choice of a trimming method for PIAAC.

3. Trimming Procedure in PIAAC

3.1 Motivation

In PIAAC, the sample design for a country depends on the available sampling frame and the size of the country. Countries that are geographically compact and have a high quality population registry can select a one-stage sample of persons from the registry. At the other extreme, large countries with no population registry may require a four-stage area sample. Regardless, the PIAAC Technical Standards and Guidelines require that the core design be an equal probability sample, either of persons for samples from population registries, or of households for countries that have a household stage of selection. Countries have the option to oversample particular domains of interest, such as young adults, but this requires an increase in sample size from the core self-weighting design.

Despite the standard for the core design to be an equal probability sample, there are several reasons why a country's final weights might vary. For countries with a household stage of selection, variation is added through the within-household sampling. To reduce clustering, most countries choose to select one person per household. Under this rule, a person selected from a household with seven eligible persons, for example, would have seven times the weight as someone selected from a household with one eligible person. To reduce this variation, some countries opted to select one person in small households and two in larger households, as recommended in Krenzke, Li, and Rust (2010).

Even if a country's design resulted in an equal probability sample, variation could be introduced through the weighting adjustments. Weights are adjusted for nonresponse and calibrated to population control totals. There is a possibility for larger adjustment factors for some subgroups due to higher nonresponse or noncoverage rates. Rules have been established for collapsing adjustment cells to limit this variation. (See Kalton and Kasprzyk (1986) for more information on adjustment cell criteria.)

There are also various other country-specific reasons for weight variation related to inefficiencies in the sample designs. Examples from the first round of PIAAC included large differences between the expected and actual measure-of-size with a fixed-take sample within PSUs, or having an extra stage of selection to subsample units within multi-unit structures. A full description of each country's sample design can be found in the *Technical Report of the Survey of Adult Skills (PIAAC)* (Mohadjer, Krenzke, and Van de Kerckhove, 2013, section 4).

The method for addressing the weight variation has an impact not just on results for a particular country but also on the ability to make valid comparisons between countries. It was hypothesized that the standard *k**median rule could result in a high percentage of weights trimmed in some countries, putting into question the comparability in the magnitude of the potential bias introduced by trimming. This would not be evident to an analyst when making comparisons.

3.2 Description

To address the above concern, it was decided to modify the standard *k**median rule so that cut-point is a function of the coefficient of variation (CV) of the weights after weighting adjustments. Specifically, the cut-point in formula (1) is defined as:

$$w_0 = 3.5\sqrt{1 + CV^2(w_j)} * median(w_j) \qquad (2)$$

The term $1 + CV^2(w_j)$ represents the design effect due to unequal weighting (Kish 1992). This cut-point is intended to limit the amount of trimming by taking a higher value when there is more variation in the weights. The approach is similar to the IQR trimming method, but with a different measure of the variation in the weights. If the country's design included oversampling of certain sampling domains, then the trimming was carried out within these domains.

After trimming, the weights are re-calibrated to population control totals. The trimming and calibration steps are performed using the Rake-Trim SAS® macro (Rizzo 2014). All adjustments performed on the full sample weights are also performed on the replicates for the purpose of variance estimation using replication methods. The replicate weights were trimmed by the same factor as the corresponding full sample weight.

While serving as a general guideline, the PIAAC trimming rule in (2) was not strictly followed for all countries in Round 1. The rule was applied for 12 of the 15 countries for which the Consortium performed the weighting. The other three countries had extraordinary variation in their weights, so the PIAAC cut-point would have resulted in a high percentage of weights trimmed. Further investigation is needed to determine whether the rule could be modified to work for these countries. In addition, countries had

the option of producing their own weights rather than the Consortium doing so; in this case, they were given the standard rule as a guideline but some deviations were allowed The *Technical Report of the Survey of Adult Skills (PIAAC)* contains more information on the PIAAC weighting process.

Returning to the example in Figure 1, the CV is 0.71 for the first set of weights and 0.43 for the second. By construction, the PIAAC cut-point is higher than that for 3*median. It is lower than the IQR cut-point for the first set of weights but higher for the second. Under the PIAAC method, no trimming would be performed for either set of weights.

4. Evaluation

4.1 Description

To evaluate the PIAAC trimming rule, the weighting process was repeated four times for each of 11 PIAAC countries, using the following trimming procedures:

- The PIAAC rule;
- k*median rule, with $k = 3.5$;
- IQR rule, with $k = 4$; or
- No trimming.

The k*median and IQR rules were chosen for comparison based on their usage in the surveys reviewed in section 2.2. As in the standard PIAAC weighting process, the weights were re-calibrated after trimming.

The level of trimming under the three rules was compared based on three measures. The first was the number of cases trimmed. The next was the maximum amount of trimming for a particular case, which is measured by the trimming factor for the largest weight. The final was a more effective measure that indicated the magnitude of trimming, calculated as the overall percent reduction in the sum of weights prior to re-calibration:

$$\left(1 - \frac{\sum_j w_{jt}}{\sum_j w_j}\right) \times 100 \qquad (3)$$

The effect of trimming on the outcome statistics was also evaluated. The MSE of the mean numeracy score[2] was computed for the three sets of trimmed weights and compared to that without trimming. Since quartiles or quantiles are of interest in some surveys, and the effect of trimming could differ from that on the mean, the MSE of the median numeracy score was also examined.

The MSE was estimated using the standard formula:

$$MSE = Variance + Bias^2,$$

where bias was computed as the difference between the estimate with trimming and the estimate without trimming, and the estimate without trimming was assumed to have no

[2] For simplicity, only the first plausible value was used for this evaluation, so the estimates will differ from other published results. While the evaluation focuses on numeracy, results are expected to be similar for literacy and problem-solving.

bias (although the actual bias is unknown). Variance was calculated using the appropriate replication method for the mean and Taylor series linearization for the median.

The evaluation was performed for the 11 countries for which the Consortium was responsible for weighting, the standard trimming rule was applied, and there was no oversampling. Table 2 provides some basic information on the countries' estimates without trimming, and Figure 2 shows the distribution of the weights prior to trimming. (The cut-points in the figure will be described in the next section.) For the plot, the weights for each country have been scaled to sum to one to enable comparisons. The figure illustrates how the variation in the weights differs between countries, for the reasons stated in section 3.1.

Table 2: Range of numeracy estimates across countries, without trimming

Estimate	*Minimum*	*Maximum*
Sample size	5,010	7,632
Mean score	254	280
Standard error of the mean	0.5	1.1
Median score	256	284
Standard error of the median	0.7	2.7

4.2 Results

The trimming cut-points under the PIAAC method, k*median method, and IQR method are shown in Figure 2, as indicated by the red, orange, and green dots, respectively. Table 3 summarizes the level of trimming for each country under the alternative cut-points. Since $\sqrt{1 + CV^2(w_j)} \geq 1$, the trimming cut-point under the PIAAC rule will always be greater than or equal to that using 3.5*median, so it will result in fewer cases trimmed. With the exception of Ireland, the PIAAC rule also resulted in fewer trimmed cases than the IQR rule. For the United States, England, Ireland, and Cyprus, the k*median rule trimmed more cases than the IQR rule; the opposite was true for the other countries. As shown in Figure 2, the last four countries have larger IQR's than the other countries.

A similar pattern can be seen when looking at the trimming factor for the largest weight and the reduction in the sum of weights. For example, in Austria the largest weight was trimmed by a factor of 0.97 under the PIAAC rule but was reduced by almost a third under the IQR rule. For the Slovak Republic, the number of cases trimmed under the IQR method exceeded that of the other two methods, but the trimming was minimal, with a 0.03 percent reduction in the sum of weights. This is because there were several weights clustered around the cut-point, as can be seen in Figure 2.

The analysis indicated minimal impact of trimming on the bias and standard error of the mean and median numeracy scores for all three methods. The increase in bias was less than 0.1 on an estimate of around 250, and the standard error changed by less than 5 percent for all countries under any rule.

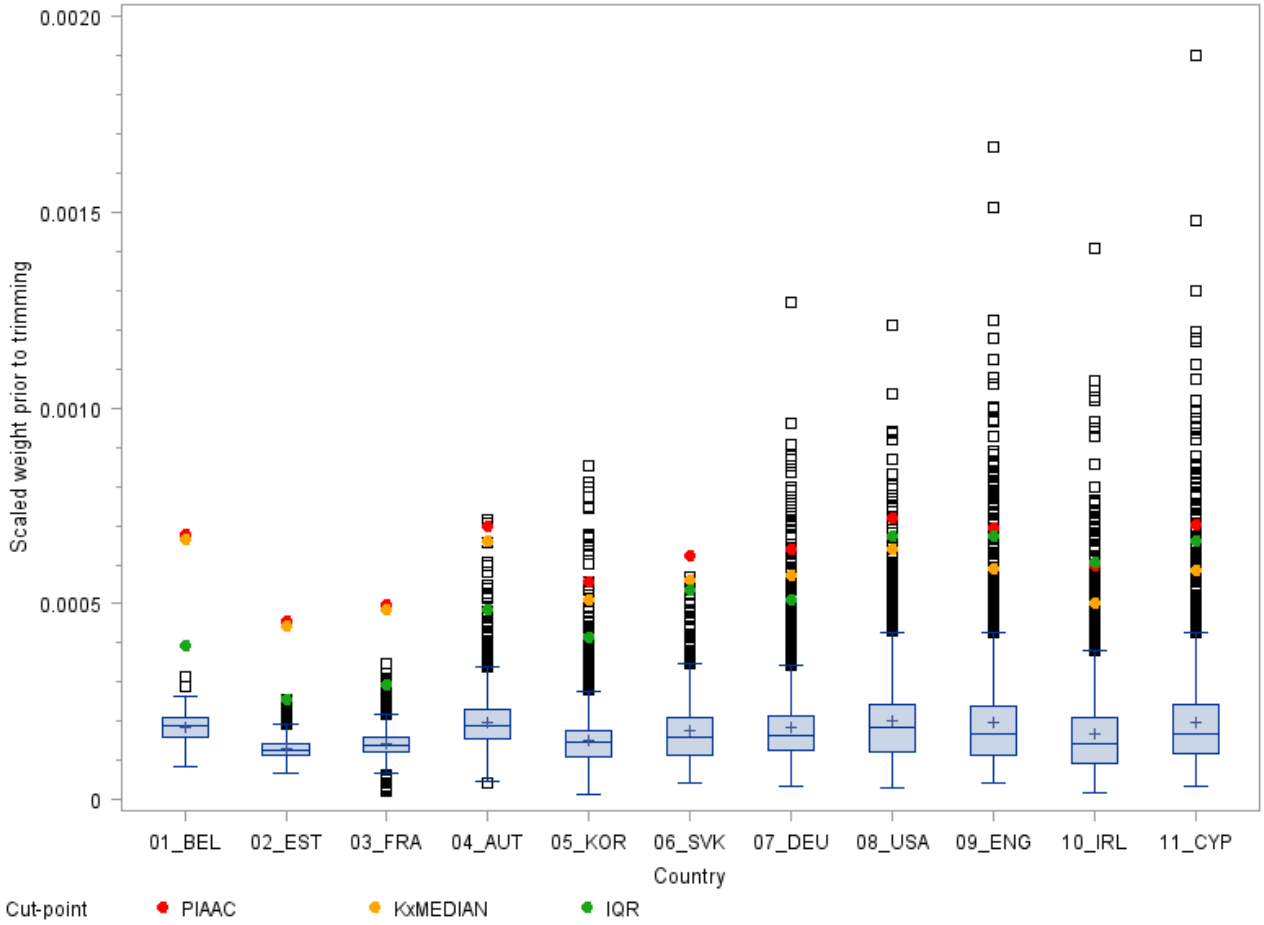

Figure 2. Distribution of the scaled weights prior to trimming, by PIAAC country

Table 3: Level of Trimming by Country and Trimming Procedure
(NA is not applicable.)

	CV of weights before	*Number of cases trimmed*			*Trimming factor for the largest weight*			*Percent reduction in the sum of weights (prior to re-calibration)*		
Country	*trimming*	*PIAAC rule*	*K*median rule*	*IQR rule*	*PIAAC rule*	*K*median rule*	*IQR rule*	*PIAAC rule*	*K*median rule*	*IQR rule*
Flanders (BEL)	20.6	0	0	0	NA	NA	NA	0.00	0.00	0.00
Estonia	20.8	0	0	0	NA	NA	NA	0.00	0.00	0.00
France	22.6	0	0	10	NA	NA	0.84	0.00	0.00	0.02
Austria	32.9	2	2	18	0.97	0.93	0.68	0.00	0.01	0.16
Korea	44.5	13	17	52	0.65	0.60	0.49	0.20	0.27	0.53
Slovak Republic	47.4	0	3	25	NA	0.99	0.94	0.00	0.00	0.03
Germany	48.9	22	41	62	0.50	0.45	0.40	0.34	0.55	0.85
United States	52.4	15	21	17	0.60	0.53	0.56	0.21	0.35	0.29
England (UK)	62.5	39	73	42	0.42	0.35	0.40	0.79	1.34	0.89
Ireland	64.3	42	82	41	0.42	0.36	0.43	0.71	1.30	0.68
Cyprus	65.9	47	85	57	0.37	0.31	0.35	0.94	1.66	1.16

5. Discussion

For the 11 countries in our evaluation, the use of weight trimming and the choice of a trimming method (PIAAC, *k**median, or IQR) had little impact on the resulting estimates of the mean or median numeracy score. In general, the PIAAC rule trimmed fewer cases than the other two procedures. However, the sum of weights after trimming was greater than or equal to 0.98 times the sum of weights before trimming under any rule. The amount of bias introduced by trimming and the change in the standard error of the estimates were negligible for all methods.

While no substantial differences were found in this evaluation, further research would be useful to better determine whether it is beneficial to incorporate the CV of the weights into the trimming cut-point. As noted earlier, the standard errors of the PIAAC estimates were small (less than 0.5% of the estimate), so the analysis could be expanded to include estimates with larger standard errors, in addition to more complex statistics such as the 90^{th} percentile or regression coefficients. One could also look at the impact under circumstances where there are more extreme weights or more extreme values of the outcome variable. The analysis was limited to estimates for the whole target population, but the effect on estimates of subgroups should also be assessed. Finally, this evaluation focused on trimming methods that are commonly used in surveys like PIAAC, but alternative weight trimming methods or weight smoothing could also be considered for comparison.

References

Deville, J.C. and C.E. Särndal. 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.

Elliott, M.R. 2008. Model Averaging Methods for Weight Trimming. *Journal of Official Statistics*, 24:517-540.

Izrael, D., M.P. Battaglia, and M.R. Frankel. 2009. Extreme Survey Weight Adjustment as a Component of Sample Balancing (a.k.a Raking), Paper 247-2009, SAS Global Forum 2009.

Kalton, G. and D. Kasprszyk. 1986. The Treatment of Missing Survey Data. *Survey Methodology*, 12:1-16.

Kish, L. 1992. Weighting for unequal P_i. *Journal of Official Statistics*, vol. 8, pp. 183-200.

Krenzke, T., L. Li, and K. Rust. 2010. Evaluating within-household selection rules under a multi-stage design. *Survey Methodology*, 36(1), 111-119.

Lee, H. 1995. Outliers in business surveys. In B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge & P. Kott (Eds.), *Business survey methods* (pp. 503-526). New York, NY: John Wiley & Sons.

Mohadjer, L., T. Krenzke, and W. Van de Kerckhove. 2013. Sampling and Weighting. In *Technical Report of the Survey of Adult Skills (PIAAC)*, section 4. OECD. Retrieved August 27, 2014 from http://www.oecd.org/site/piaac/_Technical%20Report_17OCT13.pdf.

OECD. 2014. PIAAC Technical Standards and Guidelines. Retrieved August 27, 2014 from http://www.oecd.org/site/piaac/PIAAC-NPM(2014_06)PIAAC_Technical_Standards_and_Guidelines.pdf.

Potter, F. 1988. Survey of procedures to control extreme sampling weights. *ASA Proceedings of the Section on Survey Research Methods*, 446-457.

Potter, F. 1990. A study of procedures to identify and trim extreme sampling weights. *ASA Proceedings of the Section on Survey Research Methods*, 225-230.

Rizzo, L. 2014. A Rake-Trim SAS® Macro and Its Uses at Westat, Paper 1627-2014, SAS Global Forum 2014.