Behavior Research Methods (2022) 54:1051–1062
https://doi.org/10.3758/s13428-021-01553-1

Comparing group means with the total mean in random samples, surveys, and large-scale assessments: A tutorial and software illustration

Sebastian Weirich[1] · Martin Hecht[2,3] · Benjamin Becker[1] · Steffen Zitzmann[3]

Accepted: 21 January 2021 / Published online: 24 September 2021

Abstract

In many disciplines of the social sciences, comparisons between a group mean and the total mean is a common but also challenging task. As one solution to this statistical testing problem, we propose using linear regression with weighted effect coding. For random samples, this procedure is straightforward and easy to implement by means of standard statistical software. However, for complex or clustered samples with imputed or weighted data, which are common in survey analyses, there is a lack of easy-to-use software solutions. In this paper, we discuss scenarios that are commonly encountered in the social sciences such as heterogeneous variances, weighted samples, and clustered samples, and we describe how group means can be compared to the total mean in these situations. We introduce the R package `eatRep`, which is a front end that makes the presented methods easily accessible for researchers. Two empirical examples, one using survey data (MIDUS 1) and the other using large-scale assessment data (PISA 2015), are given for illustration. Annotated R code to run group to total mean comparisons is provided.

Keywords Mean comparison · Linear regression · Weighted effect coding · Survey · Large-scale assessment

In the social sciences, comparing means of two or more groups is one of the most frequently encountered research tasks. A special case is when one of those means is a group mean and the other the total mean. For instance, one could ask whether the mean performance of companies from a specific sector differs from the mean performance of all companies, or whether students in the United States differ from all students worldwide with respect to their mean educational outcome. For such group to total mean comparisons, the classical *t*-test (e.g., Kalpic, Hlupic, & Lovric, 2011) and the *z*-test (e.g., Salkind, 2010) are not appropriate because the groups are not independent from each other (see, for example, OECD, 2005, p. 132).

To construct appropriate group to total mean difference tests, (1) the group to total mean difference $\overline{y}_{group}-\overline{y}$ and (2) the standard error for this difference need to be obtained. Whereas the first calculation is rather trivial, the computation of standard errors requires more effort. A straightforward way is to capitalize on linear regression methods (Searle, 1971). If the contrasts in a linear regression analysis are not explicitly specified, the reference coding which is used by default in most software packages provokes that each group is compared to the reference group ($\overline{y}_{group}-\overline{y}_{ref}$). Changing the contrasts according to weighted effect coding (WEC; te Grotenhuis et al., 2017) in the regression model yields regression parameters which correspond to $\overline{y}_{group}-\overline{y}$, that is, the group to total mean difference. WEC simply requires redefining the contrasts in a linear regression analysis. The intercept then represents the total mean, and the regression coefficients represent deviations of the group means from the total mean. See Appendix 1 for an illustration of the differences between various coding schemes with minimal example data.

The procedure described above yields analytical standard errors when results from linear regression theory are applied. One disadvantage of WEC is that most software packages do not have implemented this procedure or use different coding

Sebastian Weirich and Martin Hecht contributed equally to this work.

✉ Sebastian Weirich
sebastian.weirich@iqb.hu-berlin.de

[1] Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

[2] Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

[3] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

schemes per default. However, we think that using WEC regression comes with several desired features: WEC regression analysis can be adequately adapted when design and/or data characteristics are more complicated. Consider, for example, that the sampling scheme involves unequally weighted cases, because individuals included in the sample are not equally representative of the whole population. In this case, we propose using a coding scheme that we call "weighted effect coding for weighted samples" (WECW; see Appendix 1 for details). WECW simply adjusts the WEC contrasts according to the individual sampling weights. Moreover, by employing the regression approach we can draw and rely on well-studied and established methods and extensions that come into play when design and/or data characteristics are more complicated, for example when clustered or multi-stage sampling is applied, or imputed variables are part of the analyses. For these scenarios, various well-studied extensions building on linear regression exist and are also promising for the estimation of group to total mean differences with WEC or WECW.

Instead of using analytical methods, standard errors for the group to total mean difference can also be estimated by employing bootstrap methods (Davison & Hinkley, 1997; Efron & Tibshirani, 1986). As pointed out by Efron and Tibshirani (1986), bootstrap methods are an alternative when the analytical computation of the standard errors becomes increasingly complicated or in certain situations such as small number of clusters (e.g., Cameron, Gelbach, & Miller, 2008).

Scope and objectives

This article aims to reach practitioners who are confronted with group to total mean comparisons. We assembled and combined trusted statistical techniques to conduct such comparisons for frequently encountered scenarios into one easily accessible software solution. In the present article, we (1) discuss frequently encountered scenarios in social science research, surveys and large-scale assessments, (2) describe how various types of effect coding and statistical routines can be used to target research questions that imply a comparison between a group mean with the total mean within these scenarios, (3) present the R (R Core Team, 2019) package `eatRep` (Weirich, Hecht, & Becker, 2020), which facilitates such comparisons, and (4) give two empirical examples for illustration.

In the following, we present some typical scenarios which come with specific characteristics researchers are confronted within experimental studies and survey analyses and describe how WEC can be adjusted and extended. For the sake of clarity, analytical and implementation details are given in the appendices. All supported scenarios are summarized in Table 1 along with practical guidance on how to use the function `repMean()`

Table 1 Frequently encountered scenarios for group to total mean comparisons, two exemplary combinations, and arguments for function `repMean()`

Scenario						`repMean()` arguments						
No.	Sampling	Clustering	Imputed data	Group variances	Group sizes	Type	wgt	imp	PSU	repWgt	hetero1[a]	Stochastic groupsizes[a]
1	random											
2	weighted						[weight variable]					
3		yes				JK1/JK2/ BRR/Fay			[primary sampling unit variable]	[replicate weights variable]		
4			yes					[imputation variable]				
5				heterogeneous							TRUE	
6					stochastic							TRUE
Exemplary combinations												
2+5+6 (MIDUS 1)	weighted			heterogeneous	stochastic		[weight variable]				TRUE	TRUE
2+3+4+5 (PISA 2015)	weighted	yes	yes	heterogeneous		Fay	[weight variable]	[imputation variable]		[replicate weights variable]	TRUE	

In all scenarios, argument `cross.differences` is TRUE and argument `crossDiffSE` is "wec". [a] default is FALSE

from the `eatRep` package. Annotated R code with runnable examples is provided in the Supplementary Material.

Commonly encountered scenarios

Scenario 1: Random samples

To compare group means with the total mean in random samples, we suggest employing linear regression with *weighted effect coding*, a coding scheme which defines the contrasts in a way that the regression coefficients represent deviations of group means from the total mean (see, e.g., Sweeney & Ulveling, 1972). In contrast to *effect coding* (EC), WEC takes into account that the groups may be of unequal size in the population. The WEC-intercept represents the total mean, whereas the EC-intercept represents the "synthetic" total mean (i.e., the mean of the equally weighted group means). Both approaches yield identical results if the groups are of equal size in the population, which is, however, rarely the case. For an illustration of the differences between EC and WEC, see Table 5 in Appendix 1. The contrasts for group 1 are –1 for EC and $-n_2/n_1$ for WEC, where n_1 is the number of observations in group 1, and n_2 is the number of observations in group 2. As illustrated in Appendix 1, applying linear regression with WEC yields regression estimates which represent point estimates for the differences of the group means from the total mean. Moreover, the corresponding standard errors of parameter estimates represent standard errors for these differences. As shown in Table 1, for scenario 1, the argument `crossDiffSE` of the `repMean()` function needs to be set to "wec" (default).

In the following section, we describe how WEC can be applied to designs and situations that are typically encountered in the context of surveys and large-scale assessments. Samples from such studies differ from common random samples in several respects (for a more detailed description of survey samples, see, e.g., Rutkowski, Gonzalez, Joncas, & Von Davier, 2010).

Scenario 2: Weighted samples

Often, sampling designs include over- and/or undersampled groups. One common reason for this is that groups that are only marginally represented in the population should be represented more strongly in the sample to ensure sufficient power in between-group comparisons (Schofield, 2006). Hence, group-level weights are necessary to ensure that the estimates represent population parameters if the proportion of the groups in the sample does not represent the proportion of the groups in the population. Moreover, individual weights are needed to adjust for nonresponse (Rust, 2014).

To apply weighted effect coding for weighted samples (WECW), the contrasts are defined in a different manner. More specifically, the contrasts now additionally must take into account that the relative group sizes in the sample differ from the relative group sizes in the population. Picking up the example given in Appendix 1, the contrast for group 2 is now calculated as $-\left(\sum_{i=n_1+1}^{n_1+n_2} w_i / \sum_{i=1}^{n_1} w_i\right)$, where w_i are the individual weights, and n_1 and n_2 are the number of examinees in the corresponding group (see Appendix Table 5). Hence, the number of observations in the corresponding group is replaced by the sum of weights for all individuals in the corresponding group. To use WECW instead of WEC, simply supply the name of the weighting variable to the `wgt` argument of the `repMean()` function.

Scenario 3: Clustered samples

If the sampling design is hierarchical, the primary sampling unit is often some kind of higher-level entity, for example, school classes instead of individuals. It is well known that analyzing clustered samples with methods that are based on the assumption of random sampling yields biased standard errors (Lumley, 2004; Wolter, 1985). Alternatively, so-called sandwich estimators (Freedman, 2006; Skinner & Wakefiel, 2017) can provide consistent standard errors even if there is heteroscedasticity or clustered sampling (Rogers, 1993). However, in large-scale assessments, using complex designs which employ clusters with unequal selection probabilities (probability-proportional-to-size (PPS) selection; see Rust, 2014), sandwich estimators are seldom used (Gonzalez, 2014). One possible reason might be that sandwich estimators can yield biased results if the response variable is dichotomous or when cluster sizes are small (Rabe-Hesketh & Skrondal, 2006). Moreover, Efron and Tibshirani (1986) noted that analytical computation of standard errors becomes increasingly complicated for complex sampling designs. Hence, a common approach is to use resampling techniques such as, for example, the bootstrap (Davison & Hinkley, 1997), jackknife (Rust, 2014; Rust & Rao, 1996; Wolter, 1985), or balanced repeated replicates (BRR; Rao & Wu, 1985). Resampling methods like the bootstrap might be superior to analytical methods such as, for example, the sandwich estimator (e.g., Harden, 2011), particularly when the number of clusters is small (e.g., Cameron et al., 2008; Feng, McLerran, & Grizzle, 1996; Sherman & le Cessie, 1997). Resampling techniques are implemented in various software programs (Westat, 2000) as well as in R packages such as `survey` (Lumley, 2019)

or `BIFIEsurvey` (Robitzsch & Oberwimmer, 2019). Which of these methods is appropriate depends on the specific sampling procedure used in the study. For example, when the aim is to re-analyze the PISA 2015 data, the sampling procedure used by PISA 2015 should be taken into account. PISA 2015 used a balanced repeated replication (BRR) variance estimator which is adjusted for sparse population subgroups by Fay's method (Judkins, 1990; OECD, 2017, p. 123). However, when re-analyzing data of TIMSS 2007 (Mullis et al., 2008), the jackknife estimator should be used as described in the TIMSS 2007 technical report (Foy, Galia, & Li, 2008). The R package `eatRep` includes both methods to yield standard errors for group mean comparisons. For technical details, see Appendix 2. In `repMean()`, the replication method can be specified by the `type` argument. Valid options are "JK1", "JK2", "BRR", and "Fay". Depending on the method, some additional arguments need to be specified which are described in the help files of `repMean()` in detail.

Furthermore, a common type of clustered data are repeated measurements (i.e., longitudinal data). Here, the groups (level 2 units) are the persons, and the level 1 units are the observations which are nested within persons. As longitudinal data are just a special case of two-level clustering, the presented methods are suitable for longitudinal data as well.

Scenario 4: Imputed variables

When missing values occur in surveys or large-scale assessments, multiple imputation is a common method to provide complete data for secondary analyses. Also, a special case in which imputed values occur are latent variable models (for an introduction to latent variable modeling, see, e.g., Beaujean, 2014), where individual values on the latent constructs (for example, mathematical or reading literacy) must be inferred from observed indicators, for instance from items of a competence test or from additional background information from a questionnaire. For missing values as well as for latent constructs, imputation techniques (Little & Rubin, 1987; Rubin, 1987; van Buuren, 2007) such as plausible values (PVs) imputation (Mislevy, Beaton, Kaplan, & Sheehan, 1992; von Davier, Gonzalez, & Mislevy, 2009) are often applied. It is not uncommon to replace each single missing value with multiple imputed values, a procedure that results in multiple (imputed) data sets. The analysis of this kind of data requires applying specific routines for pooling the results (Rubin, 1987). These pooling routines are also applicable to linear regression with WEC. Technical details are given in Appendix 3. When using multiple imputed data in `eatRep`, the data needs to be in the long format with a variable indicating the number of the imputation. The name of this variable needs to be passed to the `imp` argument of the `repMean()` function.

Scenario 5: Heterogeneous group variances

Linear regression with weighted effect coding relies on certain distributional assumptions, one of which is homoscedastic residuals. Especially in survey analyses, this assumption is frequently violated, which can also lead to biased standard errors (White, 1980). To compute standard errors which are robust with respect to heteroscedasticity, various methods have been proposed (Bell & McCaffrey, 2002; MacKinnon & White, 1985; Smyth, 2002; Zeileis, 2004). We adopt these methods for comparisons of group means with the total mean to receive unbiased standard errors. Within the R package `eatRep`, the function `lm_robust()` from the `estimatr` package (Blair, Cooper, Coppock, Humphreys, & Sonnet, 2020) is called, which provides a variety of heteroscedasticity-robust variance estimators. In `repMean()`, the argument `hetero` defines whether group variances should be considered as heterogeneous or homogeneous. For heterogeneous variances, just set argument `hetero` to TRUE (default). With the additional argument `se_type` the method to handle heterogeneous variances can be chosen with valid options being "HC3" (default), "HC0", "HC1", "HC2" (which are exactly the same as the labels in the `lm_robust()` function from the `estimatr` package).

Scenario 6: Stochastic group sizes

Mayer and Thoemmes (2019) emphasize the distinction between *fixed* and *stochastic* group sizes. Group sizes are fixed when the researcher determines the number of persons in each group in advance of the sampling. This might be the case, for instance, in experiments where the experimenter determines how many persons are assigned to each experimental group or in surveys/large-scale assessments where the number of sampled units is determined by the sampling design. However, when population group sizes are unknown, these need to be estimated from the group sizes in the sample. As group sizes vary over samples, they are "stochastic" or "random", and estimation is accompanied by uncertainty. This uncertainty should be taken into account to avoid flawed inferences (e.g., Mayer & Thoemmes, 2019). For the estimation of group to total mean differences, we adapted and implemented a multigroup structural equation model with stochastic group sizes as proposed by Mayer, Dietzfelbinger, Rosseel, and Steyer (2016) using the R package `lavaan` (Rosseel, 2012). Thus, the uncertainty associated with

stochastic group sizes enters into the standard errors of the mean differences.

The R package `eatRep`

When group to total mean differences are to be estimated, `eatRep` employs linear regression. If no weights are specified (scenario 1), contrasts are defined according to WEC. If weights are specified (scenario 2), contrasts are defined according to WECW (see Appendix 1 for details). In clustered samples (scenario 3), `eatRep` uses `lm()` in combination with the `withReplicates()` function from the `survey` package (Lumley, 2019) to provide cluster-robust standard errors using replication techniques (see Appendix 2). When imputed variables are part of the analysis (scenario 4), the results of the regression analysis are pooled using the `pool()` function from the `mice` package (van Buuren & Groothuis-Oudshoorn, 2011; see also Appendix 3). Heterogeneous group variances (scenario 5) are taken into account by calling the `lm_robust()` function (instead of the `lm()` function) from the `estimatr` package (Blair et al., 2020). If group sizes should be considered stochastic (scenario 6), a multigroup SEM approach (Mayer & Thoemmes, 2019) is called instead of the `lm()` function, using the R package `lavaan` (Rosseel, 2012). These methods can also be combined. For example, the multigroup SEM approach (scenario 6) can be used with or without imputed data, WECW can be used with or without clustered data, and so on[1].

Empirical examples

The abovementioned scenarios are prototypical. In practice, however, researchers are often confronted with combinations of such scenarios—for example, missing values in weighted clustered samples. The R package `eatRep` (Weirich et al., 2020) offers easy-to-use functionality to compute group to total mean differences for the five presented prototypical scenarios and further combinations. In the following, two empirical examples with annotated R code (see Supplementary Material) are provided to illustrate how these comparisons can be conducted with data from survey and large-scale assessment studies.

Empirical example 1: MIDUS 1

In this example, we investigate whether the mean tobacco usage in several industry sectors differs from the mean tobacco usage in the population. To this end, we use data from the "Midlife in the United States (MIDUS 1), 1995-1996" project (Brim et al., 2019). This example can be seen as a combination of scenarios 2, 5, and 6 because we have a weighted sample and heterogeneous group variances. The group sizes need to be treated as stochastic because the population sizes of the industry sectors are estimated by the sector sizes in the sample. From the total sample of 7108 participants, we chose current smokers from the main sample who completed the phone interview and the self-administered questionnaire with non-missing values on the variables "cigarettes per day" (A1PA44) and "current industry" (A1PINMJ)[2]. Moreover, industries with sample sizes below 30 were discarded. This yielded a sample size for our analysis of 451 participants. As weights, we used the values from the provided weighting variable (A1WGHT2). The analysis was conducted with the `repMean()` function from the R package `eatRep` using linear regression with weighted effect coding for weighted samples (WECW) to test for differences between the group means and the total mean.

Results are presented in Table 2. The estimated total mean was 27.61 cigarettes per day. The group means of the seven industries ranged from 24.30 ("Professional and related services") to 34.25 ("Construction"). The results indicate that in the industries "Construction" ($M = 34.25$) and "Transportation, communications, and public utility" ($M = 31.58$), significantly more cigarettes are smoked each day than in the total population ($p = .003$ and $p = .011$, respectively). In "Professional and related services", the average tobacco use is significantly lower than in the population ($M = 24.30$, $p = .018$). The group means of the other industries do not significantly differ from the total mean. Annotated R code and an example data set which was generated based on these results is provided in the Supplementary Material.[3]

[1] However, not all combinations seem feasible or are implemented yet (see Discussion).

[2] Detailed information on the variables can be found in the MIDUS 1 material of Brim et al. (2019). A1PA44 is a numeric variable in which the quantity of daily smoked cigarettes is encoded. A1PINMJ is a numeric variable in which the respondent's current industry (major group) is represented. These 12 major groups (1=Agriculture, forestry, fishing, and mining; 2=Construction; 3=Manufacturing; 4=Transportation, communications, and public utility; 5=Wholesale trade; 6=Retail trade; 7=Finance, insurance, and real estate; 8=Business and repair services; 9=Personal services; 10=Entertainment and recreational services; 11=Professional and related services; 12=Public administration) were coded from verbatim responses using Census 1980 classification. The weighting variable A1WGHT2 is the "main random-digit-dial (RDD) phone and self-administered questionnaire (SAQ) sampling and post-stratification weight".

[3] The MIDUS data are only available after registration and are therefore not directly downloadable. Also, further distribution is not allowed. Therefore, we cannot provide the original data set. However, in order to offer R code that is executable, we provide a "synthetic" data set which we simulated based on the presented results. This approach yields only approximate results. Therefore, the results presented in Table 2 (based on the original MIDUS data set) differ slightly from the results obtained from our analysis of the "synthetic" data presented in the supplemental R code.

Table 2 Mean number of cigarettes per day by industry (MIDUS 1 data)

Group mean	Industry	n_{group}	M_{group}	Diff.	SE_{diff}	p	SD_{group}
Significantly above total mean	Construction	55	34.25	6.65	2.15	.002	14.24
	Transportation, communications, and public utility	53	31.58	3.97	1.52	.009	11.56
Equal to total mean	Manufacturing	117	27.81	0.20	1.29	.875	14.56
	Business and repair services	35	27.60	0.00	2.61	.999	17.68
	Finance, insurance, and real estate	40	25.36	−2.25	2.05	.273	13.36
	Retail trade	85	25.13	−2.48	1.63	.128	13.51
Significantly below total mean	Professional and related services	91	24.30	−3.31	1.38	.016	12.10

$N_{total} = 475$; $M_{total} = 27.61$; $SD_{total} = 13.90$; $\alpha = .05$; Diff. = Estimated difference of group mean and total mean. All statistics reported are calculated using the weighting variable A1WGHT2

Empirical example 2: PISA 2015

We used data from the 2015 PISA study (OECD, 2016) to compare the OECD countries' performances in science. The purpose was to test which country's mean performance differs from the OECD average[4]. This example can be seen as a combination of Scenarios 2, 3, 4, and 5 as we have a weighted, clustered sample with imputed values (PVs) and heterogeneous group variances. The group sizes, however, are considered to be fixed. The total sample consisted of $N = 248{,}620$ students in 35 countries. As the dependent variable, we used 10 plausible values (variables PV1SCIE to PV10SCIE from the publicly available PISA 2015 data set)[5]. We employed the senate weight variable (SENWT) "to sum up to the target sample size of 5000 within each country" (OECD, 2017, p. 292). Again, the analysis was conducted with the `repMean()` function from the R package `eatRep`.

Results are summarized by Table 3. In line with the results reported in the PISA 2015 report (OECD, 2016, p. 67), means of seven OECD countries (United States, Austria, France, Sweden, Czech Republic, Spain, and Latvia) did not significantly differ from the 2015 OECD average of 493, whereas 18 countries range above the OECD average and 10 below. Annotated R code to reproduce these results using the freely available data from the OECD homepage is provided in the Supplementary Material.

Discussion

Research questions aiming at group to total mean comparisons are frequently encountered in the social sciences. To address such comparison problems analytically, different methods can be used. A straightforward method is linear regression with weighted effect coding. To facilitate and promote usage of this approach, we developed the R package `eatRep`, in which routines for various situations that are typical in survey and large-scale assessment studies (e.g., heterogeneous variances, weighted samples, clustered samples, and multiple imputations) are implemented. To illustrate the usage of `eatRep`, we conducted two empirical example analyses in which we compared mean tobacco consumption in certain industries to the total mean (MIDUS 1 data) and the mean science competence of students in the OECD countries to the total OECD mean (PISA 2015 data).

Several issues and limitations need to be taken into consideration: (1) Weighted effect coding (WEC) presupposes that there is only one single grouping variable. With more than one variable (e.g., country and gender), the crossed groups (e.g., Japanese girls) need to be technically mapped onto one grouping variable. (2) In `eatRep`, the functionality to handle stochastic group sizes is incorporated. To this end, we used the Mayer

[4] The OECD average is calculated as the mean of equally weighted country means, which ensures "an equal contribution by each of the countries" (OECD, 2017, p. 377). Alternatively, the OECD total is calculated as the weighted mean where "each country contributes in proportion to the number of 15-year-olds enrolled in its schools" (OECD, 2016, p. 19). Whether each single country's mean should be compared to the OECD average or the OECD total depends on the specific research question. If the focus lies on comparison of countries, the OECD average might be the appropriate reference criterion. However, if the research question deals with the question whether a randomly selected student from a specific country outperforms a randomly selected student from the entirely of OECD countries, the OECD total might be the appropriate reference criterion. The standard of comparison can be chosen by using the senate weight variable for OECD average and the (adjusted) student weight variable for the OECD total.

[5] Detailed information on the variables can be found in the technical report of the PISA 2015 study (OECD, 2017). The PV1SCIE to PV10SCIE variables refer to 10 plausible values for the latent dependent variable "science literacy". As latent variables are considered to be inherently unobserved, 10 plausible values were generated using the observed item responses as well as background information gathered from student questionnaires (see chapter 9 in OECD, 2017). We use the SENWT variable as the weighting variable. The so-called senate weights assume a population of 5000 in each country. To take the sampling design of PISA into account, we specify Fay's method with the `type` argument in `eatRep`. As the OECD provides the replicate weights within the data, the 80 replicate weighting variables have to be specified with the `repWgt` argument.

Table 3 Mean science performance by country (PISA 2015 data)

Group mean	Country	n_{group}	M_{group}	Diff.	SE_{diff}	p	SD_{group}
Significantly above the OECD average	JPN	6647	538.40	45.19	2.89	<.001	93.48
	EST	5587	534.19	40.99	2.14	<.001	88.91
	FIN	5882	530.66	37.46	2.35	<.001	96.18
	CAN	20,058	527.71	34.50	1.99	<.001	92.37
	KOR	5581	515.81	22.61	2.99	<.001	95.19
	NZL	4520	513.30	20.10	2.37	<.001	104.11
	SVN	6406	512.86	19.66	1.40	<.001	95.20
	AUS	14,530	509.99	16.79	1.51	<.001	102.30
	GBR	14,157	509.22	16.02	2.52	<.001	99.66
	DEU	6504	509.14	15.94	2.68	<.001	99.34
	NLD	5385	508.58	15.37	2.16	<.001	100.95
	CHE	5860	505.51	12.30	2.87	<.001	99.53
	IRL	5741	502.58	9.37	2.30	<.001	88.90
	BEL	9651	502.00	8.80	2.22	<.001	100.19
	DNK	7161	501.94	8.74	2.36	<.001	90.30
	POL	4478	501.44	8.23	2.46	.001	90.80
	PRT	7325	501.10	7.90	2.25	<.001	91.83
	NOR	5456	498.48	5.28	2.19	.016	96.25
Equal to the OECD average	USA	5712	496.24	3.04	3.13	.331	98.63
	AUT	7007	495.04	1.84	2.35	.436	97.35
	FRA	6108	494.98	1.78	2.07	.390	101.97
	SWE	5458	493.42	0.22	3.52	.950	102.49
	CZE	6894	493.20	−0.37	2.18	.865	95.27
	ESP	6736	492.83	−0.42	2.07	.841	88.02
	LVA	4869	492.79	−2.98	1.54	.054	82.22
Significantly below the OECD average	LUX	5299	482.81	−10.40	1.11	<.001	100.41
	ITA	11,583	480.55	−12.66	2.54	<.001	91.44
	HUN	5658	476.75	−16.45	2.41	<.001	96.34
	ISL	3371	473.23	−19.97	1.66	<.001	91.22
	ISR	6598	466.55	−26.65	3.38	<.001	106.37
	SVK	6350	460.78	−32.43	2.54	<.001	98.94
	GRC	5532	454.83	−38.37	3.82	<.001	91.93
	CHL	7053	446.96	−46.25	2.38	<.001	86.02
	TUR	5895	425.49	−67.71	3.84	<.001	79.27
	MEX	7568	415.71	−77.49	2.12	<.001	71.41

N_{total} = 248,620 (unweighted); OECD average = 493.20; OECD SD = 98.55; α = .05; Diff. = Estimated difference of group mean and OECD average. All statistics reported (except for the unweighted N_{total} and n_{group}) are calculated using the weighting variable SENWT. The weighted group size is 5000 for all countries, and the weighted total size is 175,000

et al. (2016) approach, which takes the additional uncertainty due to the stochasticity of the group sizes into account. Mayer and Thoemmes (2019) note that alternative model-based approaches are also feasible, for example, a multinomial model which could be estimated using the KNOWNCLASS option in M*plus* (Muthén & Muthén, 1998–2017). (3) To date, the multigroup SEM implemented in `eatRep` does not account for clustered data. Hence, for complex samples, the standard errors of the group to total mean differences are determined using resampling approaches. To appropriately account for stochastic group sizes in clustered and/or complex data, we believe that an appropriate resampling procedure needs to be chosen. For example, we assume that resampling approaches are needed in which the group sizes vary over replicates (e.g., classic bootstrap with case-wise resampling). However, as this is a topic for future research, to date, `eatRep` treats group sizes as fixed

when resampling methods are applied. (4) The implementation of weighted effect coding for clustered samples and/or for imputed data is based on replication methods. In contrast to alternative methods for cluster-robust standard errors like sandwich estimators, replication methods like BRR or jackknife are also suitable for nonlinear statistics (Krewski & Rao, 1981; Rao & Wu, 1985) and therefore more flexible. They come, however, with substantially more computational effort. Following the PISA example, 80 replication analyses are conducted according to 80 replicate weights, and afterwards, the whole procedure is replicated 10 times according to 10 plausible values. Overall, $80 \times 10 = 800$ replications are necessary which is computationally very demanding. In the future, the currently implemented routines in `eatRep` might possibly be improved, for instance, by employing computationally more efficient C++ routines or computational optimizations, for example, suitable time-saving shortcuts for replication methods (e.g., Magnussen, McRoberts, & Tomppo, 2010; Westfall, 2011). (5) Most data in the context of large-scale assessments provided by institutions such as the OECD is presented in the wide format; that is, each line in the data set represents one discrete person. Imputed variables, if present, occur in different columns. However, the package `eatRep` requires that data are in long format. Thus, as illustrated in the supplementary R code, the user needs to reshape the data manually. Amongst others, the R packages `reshape2` (Wickham, 2007) or `tidyr` (Wickham & Henry, 2020) provide convenient and efficient functionality for this task. (6) Although we used trusted statistical approaches and routines, the complexity encountered in survey and large-scale assessment studies calls for further validation of the proposed methods. In future research, simulation studies should examine their performance and estimation quality.

In conclusion, we have compiled trusted methods into a versatile software solution that can be used to solve the common problem of comparing group means to the total mean, and we hope that this will help researchers conducting such mean comparisons in the future.

Open practice statement

We did not preregister our presented work because we do not test substantive hypotheses. The data which we have used are already publicity available (see MIDUS 1 and PISA 2015 citations). We provide annotated R code to reproduce the reported analyses as supplementary material.

Appendix 1: Examples of linear regression with four coding schemes

Here, we illustrate with an example that effect coding (EC), weighted effect coding (WEC), and weighted effect coding for weighted samples (WECW) yields the desired target statistics. Additionally, we contrast effect coding with the popular dummy coding (DC).

For the purpose of illustration, consider the hypothetical data in Appendix Table 4. Let us assume $N = 5$ persons, of which $n_1 = 3$ persons are in group 1 (with values y_{11}, y_{12}, and y_{13}) and $n_2 = 2$ persons are in group 2 (with values y_{21} and y_{22}). The weighting variable consists of group-level components and individual components—the average weights differ between groups, and the individual weights differ between examinees within each group. Based on the data provided in Table 4, Table 5 shows the contrasts according to EC, WEC, and WECW.

Table 4 Exemplary data for $N = 5$ persons

Person i	Group j	Value y_{ji}	Weight w_i
1	1	y_{11}	$^2/_3$
2	1	y_{12}	$^4/_3$
3	1	y_{13}	2
4	2	y_{21}	$^3/_8$
5	2	y_{22}	$^5/_8$

Table 5 Codes for three coding schemes

	Coding			
Group	EC	WEC	WECW	DC
1	–1	$-n_2/n_1$	$-\left(\sum_{i=n_1+1}^{n_1+n_2} w_i / \sum_{i=1}^{n_1} w_i\right)$	0
2	1	1	1	1

EC effect coding, *WEC* weighted effect coding, *WECW* weighted effect coding for weighted samples, n_1 number of observations in group 1, n_2 number of observations in group 2, w_i individual weights

(1) EC

The value column vector is $\mathbf{y} = (y_{11}, y_{12}, y_{13}, y_{21}, y_{22})'$.

The predictor matrix with effect codes is:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

With **y** and **X** the regression coefficients can be calculated with the well-known equation:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

This yields:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \dfrac{2y_{11} + 2y_{12} + 2y_{13} + 3y_{21} + 3y_{22}}{12} \\ \dfrac{-2y_{11} - 2y_{12} - 2y_{13} + 3y_{21} + 3y_{22}}{12} \end{pmatrix}.$$

The regression intercept $\widehat{\beta}_0$ can be rewritten as $\frac{1}{2}\overline{y}_1 + \frac{1}{2}\overline{y}_2$, thus, it is the total mean of equally weighted group means ($\overline{y}_{ew}$).

$\widehat{\beta}_1$ can be transformed:

$$\begin{aligned} \widehat{\beta}_1 &= \frac{-2\left(3\overline{y}_1\right) + 3\left(2\overline{y}_2\right)}{12} \\ &= \frac{1}{2}\overline{y}_2 - \frac{1}{2}\overline{y}_1 \\ &= \frac{1}{2}\overline{y}_2 + \frac{1}{2}\overline{y}_2 - \frac{1}{2}\overline{y}_2 - \frac{1}{2}\overline{y}_1 \frac{1}{2} - \overline{y}_1 + \frac{1}{2}\overline{y}_1 \\ &= \frac{1}{2}\overline{y}_2 + \frac{1}{2}\overline{y}_2 - \left(\frac{1}{2}\overline{y}_1 + \frac{1}{2}\overline{y}_2\right) \\ &= \overline{y}_2 - \overline{y}_{ew}. \end{aligned}$$

Thus, $\widehat{\beta}_1$ is the difference of the mean of group 2 and the total mean of equally weighted group means.

(2) WEC

The predictor matrix with weighted effect codes is:

$$\mathbf{X} = \begin{pmatrix} 1 & -2/3 \\ 1 & -2/3 \\ 1 & -2/3 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

This yields:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \dfrac{y_{11} + y_{12} + y_{13} + y_{21} + y_{22}}{5} \\ \dfrac{-2y_{11} - 2y_{12} - 2y_{13} + 3y_{21} + 3y_{22}}{10} \end{pmatrix}.$$

We see that the regression intercept $\widehat{\beta}_0$ is the total mean ($\overline{y}$) or, rewritten, the weighted mean of the group means: $\widehat{\beta}_0 = (3\overline{y}_1 + 2\overline{y}_2)/5$, where $\overline{y}_1$ is the mean of group 1 and $\overline{y}_2$ is the mean of group 2.

$\widehat{\beta}_1$ can be transformed:

$$\begin{aligned} \widehat{\beta}_1 &= \frac{-2\left(3\overline{y}_1\right) + 3\left(2\overline{y}_2\right)}{10} \\ &= \frac{-6\left(\dfrac{5\overline{y} - 2\overline{y}_2}{3}\right) + 6\overline{y}_2}{10} \\ &= \frac{10\overline{y}_2}{10} - \frac{10\overline{y}}{10} \\ &= \overline{y}_2 - \overline{y}. \end{aligned}$$

Thus, $\widehat{\beta}_1$ is the difference of the mean of group 2 and the total mean.

(3) WECW

The predictor matrix with weighted effect codes adjusted with weights is:

$$\mathbf{X} = \begin{pmatrix} 1 & -1/4 \\ 1 & -1/4 \\ 1 & -1/4 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The weight matrix is:

$$\mathbf{W} = \begin{pmatrix} 2/3 & 0 & 0 & 0 & 0 \\ 0 & 4/3 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3/8 & 0 \\ 0 & 0 & 0 & 0 & 5/8 \end{pmatrix}.$$

The regression parameters can be obtained by:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}.$$

This yields:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{16y_{11} + 32y_{12} + 48y_{13} + 9y_{21} + 15y_{22}}{120} \\ \dfrac{-4y_{11} - 8y_{12} - 12y_{13} + 9y_{21} + 15y_{22}}{30} \end{pmatrix}. \end{aligned}$$

The regression intercept $\widehat{\beta}_0$ can be rewritten as $\frac{\frac{2}{5}y_{11} + \frac{4}{5}y_{12} + 2y_{13} + \frac{3}{5}y_{21} + \frac{5}{5}y_{22}}{5}$. Thus, it is the total mean of weighted individuals ($\overline{y}_w$).

$\widehat{\beta}_1$ can be transformed:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{-4y_{11}-8y_{12}-12y_{13}+9y_{21}+9y_{21}-9y_{21}+15y_{22}+15y_{22}-15y_{22}}{5\cdot 6}\\ &= \frac{-4y_{11}-8y_{12}-12y_{13}-\frac{9}{4}y_{21}-\frac{15}{4}y_{22}}{5\cdot 6}+\frac{9y_{21}+9y_{21}-\frac{27}{4}y_{21}+15y_{22}+15y_{22}-\frac{45}{4}y_{22}}{30}\\ &= -\frac{\frac{2}{3}y_{11}+\frac{4}{3}y_{12}+2y_{13}+\frac{3}{8}y_{21}+\frac{5}{8}y_{22}}{5}+\frac{\frac{3}{8}y_{21}+\frac{5}{8}y_{22}}{1}\\ &= \bar{y}_{2w}-\bar{y}_w\end{aligned}$$

Thus, $\widehat{\beta}_1$ is the difference of the weighted mean of group 2 and the total mean of weighted individuals.

(4) DC

The predictor matrix for the dummy codes (with group 1 being the reference group) is:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

This yields the following regression parameters:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \frac{y_{11}+y_{12}+y_{13}}{3} \\ \frac{-2y_{11}-2y_{12}-2y_{13}+3y_{21}+3y_{22}}{6} \end{pmatrix}.$$

We see that the regression intercept $\widehat{\beta}_0$ is the mean of the reference group, $\bar{y}_1$. $\widehat{\beta}_1$ can be rewritten as:

$$\begin{aligned}\widehat{\beta}_1 &= \left(\frac{1}{2}y_{21}+\frac{1}{2}y_{22}\right)-\left(\frac{1}{3}y_{11}+\frac{1}{3}y_{12}+\frac{1}{3}y_{13}\right)\\ &= \bar{y}_2-\bar{y}_1.\end{aligned}$$

Thus, $\widehat{\beta}_1$ is the difference between the mean of group 2 and the mean of the reference group 1. Notice that this interpretation differs from the interpretation of $\widehat{\beta}_1$ under effect coding, where $\widehat{\beta}_1$ is the difference of the mean of group 2 and the total mean. Therefore, when group to total mean comparisons are of interest, effect coding schemes should be chosen. In contrast, dummy coding is suitable for group to group mean comparisons.

Appendix 2: Resampling methods for clustered samples

Assume a hierarchical sampling design which corresponds to the jackknife procedure with n unique jackknife replicates. Then, linear regression with a certain coding scheme is applied to each of the n jackknife replicate samples, using one of the coding schemes given in Appendix Table 5. Let $\widehat{\theta}_{(1)}$ be the estimator of the group mean difference in the ith jackknife replicate sample. The empirical average of the jackknife replicates then is $\widehat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(i)}$, and the standard error is defined as $SE_{jack}\left(\widehat{\theta}\right) = \left\{\frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{(i)}-\widehat{\theta}_{(\cdot)}\right)^2\right\}^{1/2}$ (Efron & Tibshirani, 1986; McIntosh, 2016). If the hierarchical sampling design is according to the balanced repeated replication method with R replicates, the standard error can be estimated by the square root of the covariance matrix of replication estimates, i.e. $SE_{BRR}\left(\widehat{\theta}\right) = \left\{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_{(r)}-\widehat{\theta}_{(\cdot)}\right)^2\right\}^{1/2}$ (SAS Institute Inc., 2018, p. 9941; Wolter, 1985). When applying Fay's method to BRR, the standard error can be estimated via $SE_{BRR(Fay)}\left(\widehat{\theta}\right) = \left\{\frac{1}{R\cdot(1-k)}\sum_{r=1}^{R}\left(\widehat{\theta}_{(r)}-\widehat{\theta}_{(\cdot)}\right)^2\right\}^{1/2}$ (Judkins, 1990, p. 225; SAS Institute Inc., 2018, p. 9942), where $0\leq k<1$ is referred to the Fay factor. In PISA, as originally suggested, $k = 0.5$ (OECD, 2017, p. 125).

Appendix 3: Pooling rules for imputed variables

When analyzing multiple imputed data sets, the combining rules of Rubin (1987) have to be applied to the estimates obtained from linear regression with weighted effect codes. Assume a data set with missing values or latent constructs for which M imputed data sets (or M plausible values) have been generated. The quantities of interest (i.e., the regression coefficients) have to be estimated for each imputed data set separately. The regression coefficients will slightly vary between imputed data sets. Let $\widehat{\theta}^{(m)}$ be the quantity of interest in the mth imputed data set. In random samples with missing data, $\widehat{\theta}^{(m)}$ is simply the regression coefficient for one group in the mth data set; in clustered samples with missing data, $\widehat{\theta}^{(m)}$ is the jackknife or BRR estimate of the regression coefficient for one group in the mth data set. According to Rubin (1987), the pooled estimate $\overline{\theta}$ is simply the average of the individual estimates: $\overline{\theta} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}^{(m)}$ To pool the standard errors, we need the overall average of the associated variance estimates, $\overline{U}$:

$$\overline{U} = \frac{1}{M}\sum_{m=1}^{M}\left(SE\left(\widehat{\theta}\right)^{(m)}\right)^2.$$

$SE\left(\widehat{\theta}\right)^{(m)}$ is the standard error of the regression coefficient in the mth imputed data set. Again, in clustered samples, $SE\left(\widehat{\theta}\right)^{(m)}$ may result from jackknife or BRR procedures. Hence, the combining rules of Rubin (1987) are applicable to random and clustered samples as well. The between-imputation variance is $B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\theta}^{(m)} - \overline{\theta}\right)^2$. The estimated total variance is $B = T = \overline{U} + \left(1 + \frac{1}{M}\right)$. To estimate the pooled standard error, we take the square root of T. For nested or two-stage multiple imputation (Harel & Schafer, 2003; Reiter & Raghunathan, 2007; Weirich et al., 2014), the combining rules of (Rubin, 2003) can be used. For the sake of brevity, we omit the formulas and refer the interested reader to the appendix of Weirich et al. (2014).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01553-1.

References

Beaujean, A. A. (2014). *Latent variable modeling using R: A Step-by-Step guide*. Routledge

Bell, R. M., & McCaffrey, D. F. (2002). Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples. *Survey Methodology*, 28(2), 169-182

Blair, G., Cooper, J., Coppock, A., Humphreys, M., & Sonnet, L. (2020). estimatr: Fast Estimators for Design-Based Inference (Version R package version 0.22.0). Retrieved from https://cran.r-project.org/web/packages/estimatr/index.html

Brim, O. G., Baltes, P. B., Bumpass, L. L., Cleary, P. D., Featherman, D. L., Hazzard, W. R., ... Shweder, R. A. (2019). *Midlife in the United States (MIDUS 1), 1995-1996 [Data file, documentation, and code book]*. Inter-university Consortium for Political and Social Research

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3), 414-427

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and other Measures of Statistical Accuracy. *Statistical Science*, 1(1), 54-77

Feng, Z., McLerran, D., & Grizzle, J. (1996). A comparison of statistical Methods for clustered data analysis with Gaussian error. *Statistics in Medicine*, 15, 1793-1806

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science asssessment. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 Technical Report* (pp. 225-280). TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College

Freedman, D. A. (2006). On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, 60(4), 299-302

Gonzalez, E. (2014). Calculating standard errors of sample statistics when using international large-scale assessment data. In R. Striethold, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments*. Waxmann

Harden, J. J. (2011). A Bootstrap Method for Conducting Statistical Inference with Clustered Data. *State Politics & Policy Quaterly*, 11(2), 223-246

Harel, O., & Schafer, J. L. (2003). *Multiple imputation in two stages*. Paper presented at the Proceedings of Federal Committee on Statistical Methodology Research Conference, Washington DC

Judkins, D. R. (1990). Fay's Method for Variance Estimation. *Journal of Statistics*, 6, 223-229

Kalpic, D., Hlupic, N., & Lovric, M. (2011). Student's t-tests. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 1559-1563). Springer

Krewski, D., & Rao, J. N. K. (1981). Inference From Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Method. *Annals of Statistics*, 9, 1010-1019

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analyses with Missing Data*. Wiley

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1), 1-19

Lumley, T. (2019). survey: analysis of complex survey samples (Version 3.35-1)

MacKinnon, J., & White, H. (1985). Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics*, 29(3), 305-325

Magnussen, S., McRoberts, R. E., & Tomppo, E. O. (2010). A resampling variance estimator for the k nearest neighbours technique. *Canadian Journal of Forest Research*, 40(4), 648-658

Mayer, A., Dietzfelbinger, L., Rosseel, Y., & Steyer, R. (2016). The EffectLiteR approach for analyzing average and conditional effects. *Multivariate Behavioral Research*, 51, 374-391

Mayer, A., & Thoemmes, F. (2019). Analysis of Variance Models with Stochastic Group Weights. *Multivariate Behavioral Research*, 54(4), 542-554

McIntosh, A. (2016). *The Jackknife Estimation Method*. Retrieved from https://arxiv.org/abs/1606.00497v1

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29, 133-161. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1992-41607-001&site=ehost-live

Mullis, I. V. S., Martin, M. O., Foy, P., Olson, J. F., Preuschoff, C., Erberber, E., ... Galia, J. (2008). *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. TIMSS & PIRLS International Study Center

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén

OECD. (2005). *PISA 2003 Data Analysis Manual: SAS Users*. OECD

OECD. (2016). *PISA 2015 Results (Volume I). Excellence and equity in education*. OECD Publishing

OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.1). R Foundation for Statistical Computing

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of Royal Statistical Society*, 169(4), 805-827

Rao, J. N. K., & Wu, C. F. J. (1985). Inference From Stratified Samples: Second-Order Analysis of Three Methods for Nonlinear Statistics. *Journal of the American Statistical Association*, 80(391), 620-630

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple Adaptions of Multiple Imputation. *Journal of the American Statistical Association*, 102, 1462-1471

Robitzsch, A., & Oberwimmer, K. (2019). BIFIEsurvey: Tools for survey statistics in educational assessment (Version 3.3-12)

Rogers, W. H. (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin*, 13, 19-23. Retrieved from https://www.stata.com/support/faqs/statistics/stb13_rogers.pdf

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1-36

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

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57(1), 3-18

Rust, K. (2014). Sampling, weighting, and variance estimation. In L. Rutkowski, M. Von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment*. CRC Press

Rust, K., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310

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

Salkind, N. (2010). *Encyclopedia of research design*. Sage

SAS Institute Inc. (2018). *SAS/STAT 15.1 User's Guide*. SAS Institute Inc

Schofield, W. (2006). Survey Sampling. In R. Sapsford & V. Jupp (Eds.), *Data Collection and Analysis* (pp. 26-56). Sage

Searle, S. R. (1971). *Linear Models*. Wiley

Sherman, M., & le Cessie, S. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics - Simulation and Computation*, 26(3), 901-925

Skinner, C. J., & Wakefiel, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, 32(2), 165-175

Smyth, G. K. (2002). An efficient algorithm for REML in heteroscedastic regression. *Journal of Computational and Graphical Statistics*, 11, 836-847

Sweeney, R. E., & Ulveling, E. F. (1972). A Transformation for Simplifying the Interpretation of Coefficients of Binary Variablesin Regression Analysis. *The American Statistician*, 26(5), 30-32

te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., & Konig, R. (2017). When size matters: advantages of weighted effect coding in observational studies. *International Journal of Public Health*, 62, 163-167

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16, 219-242

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments*, 2, 9-36

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education*, 2(9), 1-18

Weirich, S., Hecht, M., & Becker, B. (2020). Educational Assessment Tools for Replication Methods (Version R package version 0.13.4). Retrieved from https://cran.r-project.org/web//packages/eatRep/index.html

Westat. (2000). *WesVar*. Westat

Westfall, P. H. (2011). On Using the Bootstrap for Multiple Comparisons. *Journal of Biopharmaceutical Statistics*, 21(6), 1187-1205

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48, 817-838.

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, 21(12), 1-20

Wickham, H., & Henry, L. (2020). tidyr: Tidy Messy Data (Version R package version 1.1.0). Retrieved from https://cran.r-project.org/web/packages/tidyr/index.html

Wolter, K. M. (1985). *Introduction to variance estimation*. Springer

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11, 1-17