Educational and Psychological Measurement
2018, Vol. 78(6) 1021–1055
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164417736092
journals.sagepub.com/home/epm
SAGE

Have Standard Formulas Correcting Correlations for Range Restriction Been Adequately Tested? Minor Sampling Distribution Quirks Distort Them

Wendy Johnson[1], Ian J. Deary[1], and Thomas J. Bouchard Jr.[2]

Abstract

Most study samples show less variability in key variables than do their source populations due most often to indirect selection into study participation associated with a wide range of personal and circumstantial characteristics. Formulas exist to correct the distortions of population-level correlations created. Formula accuracy has been tested using simulated normally distributed data, but empirical data are rarely available for testing. We did so in a rare data set in which it was possible: the 6-Day Sample, a representative subsample of 1,208 from the Scottish Mental Survey 1947 of cognitive ability in 1936-born Scottish schoolchildren (70,805). 6-Day Sample participants completed a follow-up assessment in childhood and were re-recruited for study at age 77 years. We compared full 6-Day Sample correlations of early-life variables with those of the range-restricted correlations in the later-participating subsample, before and after adjustment for direct and indirect range restriction. Results differed, especially for two highly correlated cognitive tests; neither reproduced full-sample correlations well due to small deviations from normal distribution in skew and kurtosis. Maximum likelihood estimates did little better. To assess these results' typicality, we simulated sample selection and made similar comparisons using the 42 cognitive ability tests administered to the Minnesota Study of Twins Reared Apart, with very similar results. We discuss problems in developing further adjustments to offset range-restriction distortions and possible approaches to solutions.

[1] University of Edinburgh, Edinburgh, UK
[2] Steamboat Springs, CO, USA

Corresponding Author:
Wendy Johnson, University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, UK.
Email: wendy.johnson@ed.ac.uk

Keywords

study participation, range restriction, skew, adjustment formulas, distortion, statistical bias

Despite often extensive and broad recruitment efforts, those willing to participate in scientific studies tend not to represent their full populations well. Among other characteristics, they tend to be a little healthier, wealthier, more likely female, more conscientious, lower in neuroticism, better educated, and to score higher on cognitive ability measures than general populations (Lonnqvist et al., 2007; Nishiwaki, Clark, Morton, & Leon, 2005; Volken, 2013). It is a positive development in social science research that there is increasing awareness of this among those interested in associations between psychological, lifestyle, and demographic characteristics and physical and mental health outcomes. There is also increasing awareness among researchers that these deviations from full-population representativeness introduce distortions in estimates of the very associations in which they are most interested, sometimes even misrepresenting their direction (Ree, Carretta, Earles, & Albert, 1994).

This growing awareness takes two forms in the literature. First, there has been a steady stream of articles introducing new methods to adjust results to offset the distortions in estimates created by sampling selectivity, both intended and not (Alexander, 1990; Bobko, Roth, & Bobko, 2001; Botella, Suero, & Gambara, 2010; Hagglund & Larsson, 2006; Hunter, Schmidt, & Le, 2006; Mendoza & Mumford, 1987; Taylor, 2004). Second, it is increasingly common for researchers to report their estimates after ‘‘correction’’ using these adjustment formulas (e.g., Berry & Sackett, 2007; Berry & Zhao, 2015; Deary, Pattie, & Starr, 2013; Johnson, Corley, Starr, & Deary, 2011; Tarescavage, Fischler, Cappo, Hill, Corey, & Ben-Porath, 2015). Emergence of the latter attests to successful communication of the former, but it can only improve the accuracy of the literature to the extent that the adjustment formulas produce accurate corrections to the estimates to which they are applied. Much of the impetus for this work has come from meta-analysis. Gene Glass (1976), who perhaps did the most to originate the technique, realized from the beginning that accuracy of any summary of study results depends intrinsically on quality of the individual studies being summarized and actively worked to create statistical procedures to compensate ex post facto for study limitations during the meta-analytic process (e.g., Glass, 1982). As with most areas of statistics, techniques have improved dramatically in recent years, often driven to large degrees by improvements in computing power and ease of use. Schmidt and Hunter (2014), two of the major contributors to this effort, even feature need for it in the subheading of the title of the third edition of their classic textbook on meta-analysis, but their concern for its need goes back to the first edition.

Researchers developing adjustment formulas do of course attempt to test their accuracy, and their tests have generally supported the formulas' accuracy at least within bounded circumstances outlined by the researchers (e.g., Alexander, 1990; Alexander, Carson, Alliger, & Cronshaw, 1989; Bobko et al., 2001; Chernyshenko &

Ones, 1999; Greener & Osburn, 1980; Gross & Fleischman, 1983; Hoffman, 1995; Lawley, 1943; Le & Schmidt, 2006; Ree et al., 1994; Sackett, Laczo, & Arvey, 2002; for examples of exceptions pointing out remaining biases, see Gross & Fleischman, 1983; Mendoza & Mumford, 1987; Roth, Bobko, Switzer, & Dean, 2001). Yet the circumstances outlined have generally focused on the kinds of contexts faced by organizations that make selection decisions such as which job applicants to hire or which applications to educational programs to accept and then later follow up to understand sources of relative levels of performance that can assist the next round of applicant review. In these situations, awareness of need to consider sample selectivity tends to be particularly high because overt efforts have been made to select the ''best,'' however defined, from some broader pool. As the studies cited exemplify, the tests are usually carried out by simulating very large random samples drawn from variables with properties considered similar to those encountered in practice in these kinds of settings, imposing various forms of selection on the randomly generated samples to mimic the levels of association and sample sizes often encountered. Then, the sample properties of interest can be calculated in the selected samples, the adjustment formulas applied, and results compared with those in the full random samples.

This is a sound approach. Simulation can reveal how well a method can recover the true underlying model or data properties in a way that empirical data never can. But simulation only works well if the assumptions on which the random samples and variable distributions on which they are based are realistic representations of those encountered in the real world. In tests of sample-selection adjustment formulas, this means that the variable properties assumed in the random sample generation and the full random samples and the processes used to impose selection on them represent those encountered in practice thoroughly. If these properties do not, simulation not only can fail to reveal ''truth,'' but it can provide an inappropriate sense of security that methods function as intended. Most researchers carrying out tests of sample-selection adjustment formulas have assumed selection to be direct. Selection is strictly direct when selectors simply will not select candidates with scores on some variable above or below some preset cutoff score. This takes place in practice occasionally in some application settings, for example, when some passing score is required on a job-knowledge test. ''Directness'' of selection is often relative or dimensional, however, even in these settings, with scores or evaluations on several different measures being considered probabilistic predictive indicators of in-role performance, without strictly applied cutoff scores on any of them. Many universities (and other organizations) even have overtly compensatory selection processes in which, for example, grades can offset lower Scholastic Assessment Test or other scores, or low socioeconomic status can offset both.

In many other selection settings, especially research sample recruitment settings, selection on variables with relevance to research questions of interest is almost always quite indirect. Considered dichotomously, selection is indirect when there is only a tendency for people with some levels of a variable of interest to be more likely to be selected than others rather than some cutoff requirement. For example, if a job ad specifies that the employer is looking for applicants with university degrees, there will be direct selection for university degree, but the applicant pool will, as a by-product, indirectly also tend to overrepresent the above-average ranges of the population distribution of IQ and underrepresent the below-average ranges because people who attain university degrees tend to have higher IQs than those who do not. But people from the lower ranges of the IQ distribution have not been precluded from applying. Some who do have university degrees will usually apply, and may even be hired, though usually indirect selection is considerably stronger (or, practically if not conceptually equivalent, less indirect, reflecting dimensionality in ''directness'') in the group actually hired than it was in the applicant pool. Researchers' participant-selection settings are far less formal. Researchers are usually looking for samples as broadly representative of their target populations as possible, and are rarely intentionally picking and choosing among possible participants on the basis of the variables on which indirect selection into their samples tends to take place. But the same kinds of individual self-selection processes involving personality and cognitive characteristics that get people to apply for some jobs and not others are often if not usually involved in decisions to participate in research studies, especially when researchers issue calls for volunteers rather than recruiting from population rosters (e.g., Johnson, Brett, Calvin, & Deary, 2016).

Unfortunately, the challenges involved in adjusting correlations for direct selection are straightforward compared with those for adjusting for indirect selection. Since most studies evaluating adjustment formula accuracy have focused on direct-selection adjustment formulas and tested their accuracy by simulating direct-selection situations, the relevance of their conclusions to indirect-selection situations is not clear. Moreover, these formula-evaluation studies have tended to be based on assumption that the variables on which the focal range restriction takes place are normally distributed, at least in the full population (e.g., Hoffman, 1995; Sackett, Laczo, & Arvey, 2002). Formally, normality of these variables is not required for formula accuracy (Lawley, 1943); all that is required is that eliminated data be formally ''missing at random'' (MAR), regression of range-restricted variable on the other variable be linear, and involved error variances be homogeneous (Greener & Osburn, 1980). The latter two, however, generally depend on population-level normality of at least one of the relevant distributions, but range-restriction formula-evaluation studies have not generally considered this. Even Greener and Osburn's (1980) study, that established this by testing sampling-selectivity correction formula accuracy when the variable on which range was restricted substantively deviated from normality in six different ways, assumed that the other variable, or outcome, in the correlation was normally distributed in the underlying population.

The formal term *MAR* has found a place in the statistical literature, but this is somewhat unfortunate, as many assume that it means that the probability of missingness is unrelated to the data distribution in any way, formally ''missing completely at random'' (MCAR; Little & Rubin, 1987). This situation is clearly not the case when range is restricted, but the term *MAR* does not have this meaning. Instead, it has the less restrictive meaning that the *conditional* probability of missingness is unrelated to the data distribution. In other words, the probability of missingness can depend on the data distribution as it exists in the full-population sample, but *not* on the data distribution as it exists in the selected sample. For example, the common situation that older participants who tend to score lower on cognitive tests are less likely to participate in research studies that test cognitive ability makes the data not MCAR. It would take something such as reluctance to participate explicitly *because of* participant awareness of their likely low scores to make the data not MAR. As this is much less common, it is often reasonable to consider range restriction to be an example of randomly missing data absent evidence that it is not.

For many constructs, at least population-level normal distribution and MAR may generally be reasonable assumptions, perhaps especially for the kinds of cognitive ability and achievement-related measures often the focus of studies of situations such as evaluation of job and educational program applicants where awareness of need to consider sample selection has been particularly high. But even if the constructs we intend to measure are completely normally distributed in the population, none of our measures of them ever is in practice, even at the population level (Micceri, 1989). This is of course at least partly due to the fact that we never manage to obtain completely population-representative samples, but it is importantly also due to the fact that our measures are never adequate to generate completely normally distributed samples even if we *were* able to obtain a completely population-representative sample of a completely normally distributed construct. When samples are close to population-representative and relations among constructs reasonably linear (or their deviations from linearity reasonably captured by straightforward transformations such as squaring or taking logarithms), many estimation procedures are quite robust to small deviations from normality, and the boundaries beyond which they are not, and what to do about it, have been quite well articulated (e.g., Wilcox, 1996, 1997). But sample selection, by its very nature, usually alters not just the variance of a distribution but its degrees of skew and kurtosis because it eliminates or ''thins out'' representation of people who earn scores at one end of the distribution more than the other. These are exactly the kinds of circumstances that introduce violations of linearity and homogeneity of variance in selected samples where they do not exist in the full population. Moreover, in practice, sample range is often restricted not just on one variable, but on several, including the variable considered the outcome, all of which inevitably show at least small deviations from normality at the population level. Formula-evaluation studies have not generally considered the impact of this. Accuracy of estimates adjusted for deviations from underlying assumptions is more sensitive to appropriateness of underlying assumptions than is accuracy of the estimates themselves. This is because any adjustment formula must accurately reflect both how far off the estimate is and counteract that inaccuracy, but the estimation process only has to remain relatively stable in the presence of the deviation from assumption. Perhaps, even more important, even norming samples of most psychological measures are still subject to considerable selection, so we can never be sure that what we observe to be ‘‘the’’ score distribution from a sample intended to represent the population reflects test properties rather than sample selection.

This article is an outgrowth of work with a sample for which there was evidence of unusual population representativeness (the Scottish Mental Survey, 1947 [SMS1947]; Scottish Council for Research in Education [SCRE], 1949), a subsample that resulted from a commonly used and thoroughly applied recruitment process (the 6-Day Sample Follow-Up Study; Brett & Deary, 2014; Deary & Brett, 2015), and two well-known and highly correlated cognitive tests considered to have high construct, concurrent, and predictive validities and test–retest reliabilities that showed similar evidence of indirect selection effects in the subsample relative to the full population from which the subsample was recruited. In the full sample, both tests also showed very typical distributional properties that would not generally arouse any suspicion of deviation from normality, though the particular small deviations they did show differed. Usually, the underlying population correlation that is the target of procedures adjusting subsample correlations for selection is unknown, so it must be assumed that any such procedures applied have ‘‘done the job’’ as there is no practical way to evaluate their actual accuracy. In our case, however, this was possible.

In the process of work focusing on the follow-up subsample, we noted that application of the most commonly used approaches to adjusting correlation coefficients involving these cognitive tests for range restriction in the subsample did not reproduce full-sample correlations involving these tests at all well, and that the kinds and magnitudes of deviations the two tests produced differed. This led to exploration of the reasons for the failures of the adjustment procedures and evaluation of the degrees to which such failures are typical. This kind of exploration must be supplemented by empirical data: Simulation can help test whether potential reasons for any failures are accurate, but it cannot reveal the extent to which failures may be typical in practice as it relies on aggregations of large numbers of samples and variables that must be generated artificially based on ultimately arbitrary assumptions that can never reflect all real-world possibilities. Of course, no single empirical study can do this either, because, like simulation, empirical data have much more power to refute arguments or theories than they do to confirm them. Still, empirical data can reveal common conditions that simulators would never conceive of modelling. This article thus reports the results of empirical explorations in typical real-world data. The 6-Day follow-up sample was a small subset of even the surviving members of the original sample, so of course one of the primary possible reasons for failure was violation of the required assumption that the data were MAR (Little & Rubin, 1987). Based on available measures, there was no evidence that this was the case, but the data were clearly not missing completely at random, and there was no way to verify completely that they were in fact MAR. To accomplish our evaluations of the degrees to which such failures are typical, we thus turned to a separate study in which a larger, more population-representative group of participants had completed 42 different typical cognitive ability tests comprising three separate test batteries and in which we could ensure MAR subsamples.

Method

Source of Our Original Observations: The 6-Day Sample and Older-Age Follow-Up Study

Participants. SCRE conducted the SMS1947 on June 4, 1947, of almost all children born in 1936 who were attending schools in Scotland (*n* = 70,805 of the full birth cohort of 75,211; those not tested were not present that day at school; SCRE, 1949). The purpose of the survey was to examine to what degree the population distribution of cognitive ability might have changed since SCRE's first such survey in 1932 of children born in 1921 (SCRE, 1933). These two surveys are among the most completely population-representative samples ever (they tested almost the entire population year-of-birth cohorts), as extensive efforts were made to test all schoolchildren born in the targeted years, even those in remedial or special education programs or who suffered other disabilities. Following the 1947 Survey of those children who had been born in 1936, a subsample of survey-eligible children born on the first day of any even-numbered month (thus effectively randomly) received an additional cognitive assessment and their families and teachers completed a Sociological Survey at age 11 years and were reassessed almost annually on a number of factors from ages 15 to 27 years. These 1,208 (618 female) participants were called the 6-Day Sample (MacPherson, 1958; Maxwell, 1969). Teachers, educational psychologists, and survey administrators representing SCRE visited and surveyed participants' homes and interviewed sample members and their parents on measures of personality, details of family socioeconomic circumstances, school attendance, and, after participants left secondary school, details of further schooling, employment, marriage, and family to administer the follow-up surveys, and head teachers completed assessments of participating students at their schools at age 14 years. Table 1 shows data comparing the 6-Day Sample with the full SMS participant group on demographic variables available in the full SMS. All differences were of very small effect size, and the largest were in age, consistent with the 6-Day Sample participants' births on the first days of months.

In 2012, a follow-up study was launched to recontact the 6-Day Sample participants in older age. The original 6-Day Sample participants were traced through United Kingdom and Scottish population records, locating as many of those surviving as possible and recording deaths and their causes (Brett & Deary, 2014; Deary & Brett, 2015). Located participants residing in Scotland, England, or Wales were invited by mail to participate in a study to explore associations between their early-life demographic circumstances and psychological characteristics and their current demographic circumstances, psychological characteristics, well-being, health status, and attitudes toward life in late 2012 and 2013, when they were about age 77 years. The invitation included the assessment package, with instructions for self-administration. Invitees were requested to return a one-page form indicating willingness (or not) to participate, and to return the assessment package by mail on completion.

Of the original 1,208 participants, 636 (including 1 earlier emigrant; 370 females) were located and invited to the follow-up. About a third (417; 164 females) were

Table 1. The 6-Day Sample and the Scottish Mental Survey.

	6-Day Sample, *n* = 1,208			Scottish Mental Survey, *n* = 70,805			Effect size of mean difference
	Mean	*SD*	Skew	Mean	*SD*	Skew	
Full sample							
Age at MHT	10.98	0.33	.27	10.93	0.29	–.08	.17
Percent female	51.16	—	—	49.45	—	—	1.71
MHT score	37.4	15.8	–0.31	36.6	15.8	–.34	.05
Females							
Age at MHT	10.96	0.32	.19	10.93	0.29	–.07	.10
MHT score	37.8	14.9	–.27	37.5	15.1	–.36	.02
Males							
Age at MHT	11.00	0.35	.31	10.93	0.29	–.08	.24
MHT score	37.1	16.7	–.32	35.8	16.4	–.30	.08

Note. SD = standard deviation; MHT = Moray House test. Effect size is (6-Day mean-SMS mean)/SMS *SD*, except for sex, which is simple percentage difference.

deceased; 66 could not be located; and 89 had emigrated from the United Kingdom. Among the 636 invited, since being located, 1 had emigrated, 2 were deceased, and 20 were deemed not capable by English/Welsh law. Despite follow-up mailing, no replies were received from 268; another 139 refused participation. The primary reason for refusal was lack of interest. The remaining 205 indicated willingness to participate, either by completing the one-page form or telephoning the study office. Completed assessments were received from 171 (90 females), for participation rates of 27% of those invited and 83% of those who had indicated willingness. These participation rates may sound low, but the number-invited denominator represents the total cohort potentially available much more clearly and likely fully than those from many studies. This is because of the extensive efforts involved in locating or accounting for all the original participants of the unusually population-representative 6-Day Sample, including matching to National Health Service medical records on which almost everyone in the United Kingdom is recorded.

Measures. We examined accuracy of the standard range-restriction adjustment formulas to reflect actual associations among several variables from the early-life survey whose content would commonly be of interest today. The particular associations we report in the 6-Day Sample follow-up participants would not generally be of interest in themselves as the measures were taken at age 11 years, and the selection processes examined reflected survival and study participation status at age 77 years. They offered, however, an unusual opportunity to assess the accuracy of adjustment formulas in real-world test data under a naturally occurring selection process, using correlations of magnitudes similar to those often explored in recent studies. Previous studies of adjustment-formula accuracy have, among other potential limitations, relied on modelled selection processes. Any model requires assumptions about appropriate representation of the process being modelled, and it is not uncommon that later emerging empirical data indicate that the assumptions used in simulation studies generated results that misled researchers who relied on them. There are, for example, several instances of this in genetic research in the past 50 years or so, including too-long prevailing ideas that genetic stratification in populations, gene-environment interaction and correlation, and intergenerational transmission of gene expression patterns (epigenetics) could be disregarded in understanding interrelations of genetic and environmental influences on behavioral traits and clinically related phenotypes. The measures we considered were the following.

Moray House Test 12 (MHT). Most (n = 1,112) of the 1,208 6-Day Sample participants had been present at school on the day of the Scottish Mental Survey of 1947, and so completed the MHT. This test (SCRE, 1933) requires 45 minutes to administer and is a valid, group-administered test of cognitive ability consisting of 71 items (SCRE, 1949, 1953), with a maximum possible score of 76. It features many types of items, though verbal reasoning items predominate. Specific test–retest reliability over short time periods has not been assessed, but given its stability over long time periods and comparability with other IQ and cognitive ability tests, Deary, Whiteman, Starr, Whalley, and Fox (2004) estimated it at .90.

Terman–Merrill IQ Test: 1937 Revision (TMIQ). Participants in the 6-Day Sample also completed the individually administered L form of the TMIQ, one of the best-validated and most often used IQ tests at the time (SCRE, 1949). One of the purposes of administering this test was to corroborate validity of the MHT (SCRE, 1949); the correlation between the two tests' scores was .80. Weiner (2003) listed the test–retest reliability of the related Stanford–Binet test at .90.

Personality. Teachers assessed six areas of 6-Day Sample participants' personalities as part of the First School Schedule in 1950. They rated self-confidence, perseverance, mood stability, conscientiousness, originality, and desire to excel, each on a scale that ranged from *marked lack* (1) to *very* (5). All these personality characteristics tend to be related conceptually and empirically to school achievement regardless of specific measure used. As cognitive ability scores are also related to school achievement, the question of association between cognitive ability scores and ratings of personality is often of interest. We selected two of the rated characteristics, Self-confidence and Originality, to use as example correlations in this study. In the full sample, Originality's correlations with the cognitive tests were the highest of the personality measures; Self-confidence's correlations were typical of the personality characteristics rated. All were higher than commonly observed for personality-cognitive ability associations, likely due to their having been rated by the children's teachers. Test–retest reliability of single items can vary considerably, is not often directly measured, and was not assessed for these teachers' ratings of child personality. Littman, White, Satia, Bowen, and Kristal (2006) found retest-reliabilities of .66 and .74 for two single items of psychosocial stress. We considered .70 a reasonable estimate for our personality items.

Height was measured in inches at age 11 years. We considered height because it is commonly considered in epidemiological studies to reflect social class and/or general constitutional robustness, and samples are often healthier than their underlying populations. As height was measured at school and thus with considerable consistency, we considered .99 a good estimate of the measure's reliability.

Sample Used to Evaluate "Typicality": The Minnesota Study of Twins Reared Apart (MISTRA)

Participants. The MISTRA participants were gathered over a period extending from about 1979 to 2000, as rarely occurring pairs of twins who were separated early in life, reared in adoptive families, and not reunited until adulthood came to the attention of the researchers and were recruited for study (). The purpose of the study was to quantify the degree to which personal, medical, anthropomorphic, and psychological characteristics that show individual differences can be considered genetically influenced in a sample of people for which genetic and familial environment influences were not inherently confounded. To enhance incentive to participate, twins were encouraged to invite their spouses and adoptive and other biological relatives and friends as available to participate as well. The assessment covered a very wide range of individual differences. The researchers were particularly interested in the various ways in which cognitive abilities can be manifested and administered a very comprehensive set of 42 individual cognitive ability tests that spanned three established test batteries (Johnson & Bouchard, 2011). The 436 (188 males, 248 females) participants that contributed data for these analyses came from a broad range of occupations and socioeconomic backgrounds and several different countries, though most were from the United States or Great Britain. They varied in age from 18 to 79 years (mean = 42.7 years), with education levels that ranged from less than high school to postgraduate experience. Details of recruitment and assessment are reported by Segal (2000).

Measures

The three test batteries administered were the following.

Comprehensive Ability Battery (CAB). Developed by Hakstian and Cattell (1975), the CAB consists of 20 specific ability tests intended to span the range considered relevant to human intelligence. Each test is short, requiring only 5 to 6 minutes to keep administration varied and manageable. To avoid task duplication in the extensive MISTRA assessment, six of the tests in the CAB were not administered to the participants. In addition, because we judged it not directly relevant to cognitive ability, we eliminated the Esthetic Judgment test. As the Verbal Ability test consists of two completely separable tasks, we considered the scores on the two parts separately. Thus, we had a total of 14 test scores from this battery. Hakstian and Bennett (1977) reported split-half and retest reliabilities for its subtests ranging from .64 for Perceptual Speed and Accuracy to .96 for Memory Span.

Hawaii Battery, Including Raven's Progressive Matrices. The Hawaii Battery consists of 15 tests of primary abilities. It was developed to assess familial resemblance in cognitive ability in the Hawaii Family Study of Cognition (DeFries et al., 1974). Again, each test is short, requiring 3 to 10 minutes. for administration. To avoid duplication of tasks and more fully identify likely ability factors, 2 tests were not administered and the battery was supplemented with 4 tests from the Educational Testing Services, so there were 17 tests in this battery. In the validation sample, internal consistency and retest reliabilities for the tests ranged from .58 for Immediate Visual Memory to .96 for Vocabulary (Kuse, 1977). We referred to Desai's (1952) estimated reliability of .77 for the Raven, and Watkins (1979) for reliabilities for the supplemented Educational Testing Services tests.

Wechsler Adult Intelligence Scale. The 11 subtests of the 1955 version of the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955) WAIS were administered. According to the manual, internal consistency reliabilities range from .79 for Comprehension to .94 for Vocabulary. For this sample, scores normed at the 1955 level ranged from 79 to 140, with mean of 118.5, standard deviation (*SD*) of 19.8, and skew of –.08. Hanson, Hunsley, and Parker (1988) reported test–retest reliabilities for the WAIS subtests, and we made use of their estimates. Adjusted to time of administration for the Flynn effect, the IQ scores ranged from 61 to 140, with mean of 101.2, *SD* of 14.8, skew of .02, and kurtosis of –.23. Given the long time-span between publication of the WAIS version used and even the earliest MISTRA assessments, the sample was quite representative of English-speaking countries and it was also quite normally distributed. For this study, we used the subtest scores as individual cognitive tests, and the Flynn-adjusted IQ scores as the test variable with which to correlate them. Because of application of the formula for calculating WAIS IQ scores from the raw subtest scores, and removal of age and sex effects from this followed by adjustment for the Flynn effect, there was no direct dependency between any subtest score and the IQ score. We considered its reliability to be .90.

Personality. The MISTRA sample completed the full 300-item version of the Multidimensional Personality Questionnaire (Tellegen & Waller, 2008). From the 11 scales it includes, we selected Well-being, Absorption, and Alienation to use as examples of personality correlations with cognitive abilities due to the variety of distributional properties they encompassed. High scorers on the Wellbeing Scale have cheerful dispositions, feel good about themselves and their lives, are optimistic, and enjoy their activities. The scale tends to be somewhat negatively skewed in most samples, including MISTRA (–1.29, with kurtosis of 1.18). High scorers on the Absorption scale are responsive to sensory stimulation, tend to think in images and experience vivid imaginings, and often have cross-modal experiences and become "lost in thought." The scale is generally quite normally distributed, and was so in MISTRA (–.10, with kurtosis of –.70). High scorers on the Alienation scale feel "used" and "pushed around" by friends, and believe that they have been victims of bad luck and others wish them harm and have betrayed and deceived them. The scale tends to be somewhat positively skewed in most samples, including MISTRA (1.28, with kurtosis of 1.07). Tellegen and Waller (2008) reported 30-day test–retest reliabilities of .90 for Well-being, .91 for Absorption, and .87 for Alienation.

Analytical Approach

Five methods for adjusting observed correlations for range restriction can be considered broadly recognized and in common or recommended use. Four have been tested for accuracy using simulated population samples assumed to be normally distributed; the fifth has been tested only in a form relevant to meta-analysis. It was in applying these approaches to adjust observed correlations in the early-life measures in the Follow-Up Study sample that we observed their inaccuracies, particularly for the two cognitive tests, relative to those observed in the full original 6-Day Sample. To assess the extent to which such problems are common, we generated two completely random subsamples of the full MISTRA sample, mimicking the kinds of indirect selection on general cognitive ability, as measured by Flynn-adjusted IQ (Johnson et al., 2007; important because IQ was assessed using the 1955 version of the Wechsler in this sample recruited throughout the period from the late 1970s until 2000), that tend to occur within research studies. Of course, like other studies that have assessed adjustment-formula accuracy, we simulated these processes. However, we had two advantages that other studies of formula accuracy have lacked: (1) Other studies have had to rely on guesses about how populations stratify into study samples, but we could use the naturally-selected 6-Day Sample follow-up participation patterns as general guides in modelling the processes to generate the two MISTRA subsamples while assuring random, if not completely random (Little & Rubin, 1987) missingness. That is, we divided the full 6-Day Sample into quintiles, noted the proportions of these quintiles participating in the Follow-Up Study, ratioed these up to two different degrees because the participation rate in the 6-Day Follow-Up Study was so low and we wanted overall higher "participation" rates in our samples, applied these ratios of the quintiles of the MISTRA Flynn-adjusted IQ scores, and generated completely random Bernoulli-distribution variables to indicate including or not each MISTRA participant in our two "selected" MISTRA samples. (2) Prior studies of formula accuracy have had to rely on assumed distributions of the full-population scores used to assess formula accuracy, and have assumed at least the outcome distributions to be formally normal; in contrast, we had 42 available examples of naturally occurring "population-level" score distributions to which to apply our simulated selection processes to both variables in the correlations, all of which showed very typical small deviations from formally normal distribution. We modelled extent of indirect selection on IQ as somewhat stronger in the second selected MISTRA subsample than in the first. We applied each of the five adjustment methods to the correlations in the two selected subsamples between each of the 42 MISTRA cognitive ability tests and Wechsler IQ, and each of the three Multidimensional Personality Questionnaire scales.

Selection Processes Generating Range Restriction and the Commonly Used Adjustment Methods. Direct selection often (but far from always) reduces the sample standard deviation relative to that in the full population. If the relation between the two variables is linear and homoscedastic in the population, extent of attenuation is directly affected by the direct-selection cutoff score. Thorndike (1949) laid out the relations among the distribution properties involved. Formally, the degree of attenuation is

$$\frac{\frac{s}{\sigma}}{\sqrt{1+\left(\left(\frac{s^2}{\sigma^2}\right)-1\right)\rho^2}} \tag{1}$$

where σ is the population standard deviation, s the sample standard deviation, and ρ the population correlation (Hunter, Schmidt, & Le, 2006) (Hunter et al., 2006). Theoretically, the correlation in a directly selected sample can be corrected to the full-population–level correlation by multiplying it by the reciprocal of Equation (1). But this requires knowledge of both the full-population correlation and the full-population standard deviation. Often neither of these is known, and the exercise would not even be necessary if we knew the former. In his Case II, Thorndike (1949) handled this by 'reversing the nonlinear algebra of the attenuation formula' (Hunter et al., 2006, p. 596) to approximate the full-population correlation ρ as

$$r\cdot\frac{\frac{\sigma}{s}}{\sqrt{1+\left(\left(\frac{\sigma^2}{s^2}\right)-1\right)r^2}} \tag{2}$$

where r is the sample correlation, and this has become the standard treatment. In practice, the correction adjustment is often approximated even further by dropping the denominator in Equation (2) completely because the population standard deviation is unknown and so must be assumed. This is apparently done under the further assumption that, given the approximation involved in assuming the population standard deviation, the adjustment offered by the denominator is small. Terming the adjustment ''small'' is a matter of judgment, of course. For example, with a rather low ratio of population-to-selected sample *SD* of 1.2, and a moderately strong observed sample correlation of .40, the full-population correlation would be estimated at .46 using Equation (2), and .48 using Equation (2) without the denominator. Higher ratios of population-to-selected sample standard deviation and higher correlations, common in psychology study variables and samples, generate greater distortions. For example, with a ratio of population-to-selected ratio of 1.4, the analogous estimates would be .52 and .56. With a sample correlation of .5, they would be .57 and .60.

Thorndike's (1949) Case III formula has long been considered standard for adjusting for indirect range restriction. This is

$$\rho_{XY} = \frac{r_{XY} + r_{XZ} r_{YZ}(U_Z^2 - 1)}{\sqrt{1 + (U_Z^2 - 1) r_{XZ}^2} \sqrt{1 + (U_Z^2 - 1) r_{XZ}^2}} \tag{3}$$

where U_Z is the ratio of the full-population standard deviation to the sample standard deviation of the unmeasured variable on which direct selection has actually taken place that is correlated with at least X. In research participant-recruitment settings, Z would be willingness to or interest in participating in research studies. Of course, the distributional properties of this variable are effectively never known, in either the sample or the population. This is no doubt part of the reason that the formula for direct range restriction has been applied in practice and tested for accuracy much more often than Equation (3), even in situations where selection is clearly indirect. The tests comparing accuracy of the direct- and indirect-selection formulas that have been applied have generally involved substantial additional volumes of calculations and work to apply the indirect formula, as is apparent in inspecting the formulas. Their indications that the direct formula is relatively accurate have justified continuing its use.

Questions about accuracy, however, continue to surface, especially as meta-analysis has become more widely used and thorough, because there researchers must cope with many different samples and specific measures, inevitably gathered under different conditions, while relying only on reports of those conditions from the primary researchers. Recently, Hunter et al. (2006) developed a procedure they termed *Case IV* to estimate the needed Z distributions and to incorporate recognition of imperfect test reliability in both direct and indirect range-restriction situations. Under this procedure, the researcher first estimates the reliability ρ_{XX} of X in the full population as

$$1 - u_X^2(1 - r_{XX}) \tag{4}$$

where u_X is the reciprocal of U_X, the ratio of the sample standard deviation of the test to its full-population standard deviation, and r_{xx} is the reliability of the test in the sample population. The next step is to estimate $u_Z = 1/U_Z$, or the ratio of the sample standard deviation of the unmeasured variable on which direct selection into the sample (some kind of willingness to participate in the study) has taken place to its full-population standard deviation, as

$$\left[\frac{u_X^2 - (1 - \rho_{XX})}{\rho_{XX}}\right]^2 \tag{5}$$

Then the researcher corrects the observed correlation for population-level unreliability in the two measures by dividing it by

$$\sqrt{(r_{XX}r_{YY})} \tag{6}$$

to obtain r_C. The final step is to apply Thorndike's (1949) Case II formula for direct range restriction to r_C, using the estimated U_Z. That is,

$$\rho = \frac{U_Z r_C}{\sqrt{U_z^2 r_C^2 - r_C^2 + 1}} \tag{7}$$

The fourth method we applied has not been as widely used or discussed, but it has been mentioned occasionally as a possibility for more than 20 years (e.g., Mendoza, 1993) and has started to receive increased attention as a way to adjust correlations in range-restricted samples to population levels. This is maximum likelihood estimation (MLE) of expected population-level statistics given only a subset of the full population. Use of this method relies on the assumptions that the restricted range of the data in the examined sample has arisen because some data are MAR (Little & Rubin, 1987), and independent and identically (basically) normally distributed, though the method can be used assuming any distribution that seems relevant. The normal distribution has usually been the distributional form considered most relevant in studies of psychological variables involved in range-restricted samples. Here, considerable work has been done to evaluate MLE's robustness to violations of normality in general, and to offer alternatives where important. Enders (e.g., 2011) and Savalei (e.g., Savalei & Rhemtulla, 2012) have been particularly active, though their work has not focused specifically on selection-based range restriction.

Given the assumed properties and the observed data, the goal of MLE is to estimate parameters of the distributional function underlying the data. In range-restricted samples, this would mean, for example, estimating the full-population-level mean and variance and correlations of the variable with other variables of interest. The first step in doing this is to consider the joint probability density function that must exist to generate all these data, given its parameters. The data can then be considered the parameters of this joint density function, and this function itself considered the likelihood of these particular parameters (data) having arisen. The next task is then to maximize this likelihood function. Parenthetically, it is often more computationally tractable to work with the logarithm of this function rather than the function itself, giving rise to the commonly used term *maximum log-likelihood.* The result is the same either way (given appropriate back-transformation) as the logarithmic function increases monotonically. Sometimes this process leads to a directly computable solution, but often it does not, and numerical optimization methods must be applied. Other potential practical problems in using this method are that it is not uncommon that there are many very similarly maximally likely solutions, the likelihood just keeps increasing indefinitely, and/or the indicated maximum likelihood varies with the values used to start needed numerical optimization methods, making it difficult to ascertain that any indicated solution is in fact the true maximally likely one. Many programs operationalizing this method, however, have built-in features that address these complications.

Finally, noting concerns over inaccuracies that can arise in the Le and Schmidt (2006) ''Case IV'' adjustment method, Le, Oh, Schmidt, and Wooldridge (2016) recently developed a ''Case V'' adjustment method, based on a previously little-known formula developed by Bryant and Gokhale (1972). They adapted Bryant and Gokhale's basic formula

$$\rho = r_c u_x u_y + \sqrt{\left(1 - u_x^2\right)\left(1 - u_y^2\right)} \tag{8}$$

where u_y is the ratio of the sample outcome variable standard deviation to its full-population standard deviation, to reflect unreliability of measurement in the two variables involved in the correlation. This means adjusting for unreliability of measurement and indirect range restriction in the two variables in the correlation, using $r_{xp} = r_c/\sqrt{r_{yy}}$ and $r_{tp} = r_{xp}/\sqrt{r_{xx}}$ to correct the two variables' unreliability, and

$$u_t = \sqrt{r_{xx}r_{xp}^2/\left(1 + r_{xx}r_{xp}^2 - r_{xp}^2\right)} \tag{9}$$

and

$$u_p = \sqrt{r_{yy}r_{tp}^2/\left(1 + r_{yy}t_{xp}^2 - r_{tp}^2\right)} \tag{10}$$

to account for indirect selection. With these adjustments in place, the adapted Bryant and Gokhale (1972) formula becomes

$$\rho = r_{tp} u_t u_p + \sqrt{\left(1 - u_t^2\right)\left(1 - u_p^2\right)} \tag{11}$$

They noted that sometimes the two variables involved in the correlation are correlated with the indirectly selected variable in opposite directions. When this is the case, the plus sign before the radical should be changed to a minus sign. As Le et al. (2016) noted, the correlation between the two variables of interest will also generally be negative unless it is rather low. If it is substantially negative, one variable could be reverse-scored and formula (11) applied. Le et al. evaluated the accuracy of their adapted formula for meta-analytic purposes, but not for use with individual sample correlations. In doing so, they considered individual study situations with considerably more regularity and smaller ranges of variation in correlation size than tend to occur in practice, and, in particular simulated distributions of true full-population construct correlation and observed sample measure correlation pairs from varying numbers of studies with varying numbers of sample sizes. In the process, they did not address the possibility that observed standard deviation was greater than population standard deviation. Their formula generally becomes undefined when this is the case for one of the involved variables. We evaluated their formula for use in individual studies. Along with even the basic Bryant and Gokhale (1972) formula (8)'s reliance on knowing the full-population standard deviations for both variables, this is likely one of the reasons their formula has never received much attention.

Results and Discussion

6-Day Sample Results Motivating Further Study

To indicate the extent to which the 6-Day Sample actually represented the full birth cohort form which it was drawn, Table 1 compares its relevant statistics with those in the much larger SMS1947. Inevitably, given two separate sampling procedures, there were some mean differences. All were also inevitably significant, given the large size of the SMS1947. The effect sizes of the MHT-score differences were, however, trivial, and even standard deviations and skews of all the variable distributions were highly similar. The 6-Day Sample was slightly older than the SMS1947 sample, but the difference corresponded to 18 days. This is slightly longer than the average difference between the first and last day of any month, and thus almost exactly what would be expected given selection on birthdate the first of even-numbered months, and could also possibly explain their slightly higher MHT scores. The 6-Day Sample did have a slightly higher proportion of females than did SMS1947 (less than 2% difference). Given higher rates of infant mortality and incapacitating disabilities in males than females (National Records of Scotland, 2013), however, it is very possible that the 6-Day Sample was the more population-representative. The MHT scores were consistently slightly negatively skewed in females and males, This appears to be a general property of the test rather than a product of sample selectivity, as it has been observed in all samples studied to date and other cognitive tests in these samples did not show negative skew (Johnson et al., 2016). Skews of similar magnitudes in either direction are generally found in all psychological measures. They can arise through item-‘‘difficulty’’ properties of the measures as well as through population characteristics, and most analytical methods in common usage are robust to such minor violations of the normality assumption commonly underlying them.

Table 2 compares the youth scores of the 791 original participants surviving to be potentially eligible to participate in the age-77 Follow-Up study and those who actually agreed to participate in it with the full original sample. The first thing to note is the distributional properties of the measures in the full sample, particularly the skews in parentheses following the standard deviations, because they provide the best indications of the measures’ psychometric properties. The TMIQ distribution was similar to the properties usually noted for it, and generally claimed for it. Though not shown in the table, the relative magnitudes (ratios) of standard deviations to means are important in evaluating the shapes of distributions, and such ratios can be compared in measures on very different scales, as was the case here. The two personality measures had similar ratios (.27 and .31), with the TMIQ’s being lower (.20) and the MHT’s higher (.42). Height’s ratio was very small (.05). These ratios indicate that MHT scores varied much more within their possible range than did TMIQ scores, which can reflect more detail in measurement scaling, but also lower reliability when

Table 2. 6-Day Sample Youth Descriptives of Age-77 Participation Status Groups.

	Mean (*SD*, skew)		
	Full 6-Day Sample *n* = 1,208	Alive at age 77 *n* = 791	Participating at age 77 *n* = 171
Moray House Test	37.4 (15.8, –.31)	39.3 (15.3, –.38)*^{1}	48.1 (11.3, –.78)*^{2}
Terman–Merrill IQ	102.6 (20.1, .50)	105.0 (20.6, .49)*^{3}	115.6 (19.7, .18)*^{4}
Self-confidence	3.0 (0.80. 06)	3.0 (0.80, .03)	3.2 (0.76, .05)*^{5}
Originality	2.6 (0.80, –.16)	2.7 (0.79, –.15)*^{6}	2.9 (0.75. 10)*^{7}
Height	54.0 (2.8, .05)	54.1 (2.9, .00)	54.7 (2.8, .55)*^{8}

Note. All measures in youth. Superscripts denote effect sizes of differences, more restricted to less restricted: 1 = .35, 2 = .74, 3 = .37, 4 = .65, 5 = .26, 6 = .19, 7 = .43, 8 = .23.

*Difference in means was significant between survivors and deaths (middle column), or participants and nonparticipants (right column), at $p < .01$ to offset multiple testing.

the scores are intended to measure the same construct, as these two are. When these ratios are not similar, the dissimilarity can also primarily reflect degree of population clustering around the mean, which was probably the primary reason for the low height ratio.

None of the measures was skewed to a degree that would typically generate concern about material deviation from normal (maximum magnitude –.78 in the Follow-Up Sample MHT), but the two cognitive tests were more skewed than the other three, and the TMIQ was somewhat positively skewed (.50), while the MHT was somewhat negatively skewed (–.31). In addition, the TMIQ became less positively skewed as degree of selection increased, while the MHT became more negatively skewed. This is typical when the lower scores of such distributions are dropped in greater proportions than higher scores. Survivors to age 77 had moderately higher cognitive test scores in youth than those who had passed away before that age (effect sizes of .35 and .37 for MHT and TMIQ, respectively). They also had slightly higher Originality scores as rated by their teachers (effect size .19), but did not differ significantly in either Self-confidence or height. Survival sets the ultimate boundary for participation, so selection of at least these magnitudes would be expected among the participants in the Follow-Up study relative to the original full sample, but there is no predominating a priori reason for it to be greater *within* the surviving group. In fact, however, actual selection among participants was considerably larger: compared with surviving nonparticipants, participants had considerably higher cognitive test scores in youth (effect sizes of .74 and .65 for MHT and TMIQ, respectively), and somewhat higher self-confidence and moderately higher originality as rated by their teachers (effect sizes of .26 and .43, respectively). They were also taller in youth (effect size .23). This indicated considerable selection on characteristics that are usually very highly correlated (stable in population-level rank ordering) over long periods of the lifespan in the age-77 Follow-Up Sample.

Table 3. 6-Day Sample Early-Life Correlations.

		Adjusted for		
	Observed	Direct range restriction	Indirect range restriction	Maximum likelihood estimate
Moray House Test				
Participating at 77				
Originality	.323	.423	.538	.409
Self-confidence	.152	.203	.269	.248
Height	.199	.263	.532	.268
Alive at 77				
Originality	.437	.450	.567	.441
Self-confidence	.255	.263	.333	.268
Height	.315	.325	.367	.298
Difference/ratio between actual and estimated in participating sample				
Full sample				
Originality	.383	−.040/.906	−.155/.712	−.026/.936
Self-confidence	.246	.043/1.212	−.023/.914	−.002/.992
Height	.314	.051/1.194	−.218/.590	.046/1.172
Terman–Merrill IQ				
Participating at 77				
Originality	.288	.293	.370	.428
Self-confidence	.144	.147	.185	.248
Height	.210	.215	.241	.259
Alive at 77				
Originality	.474	.473	.576	.470
Self-confidence	.283	.282	.340	.278
Height	.306	.305	.327	.289
Difference/ratio between actual and estimated in participating sample				
Full sample				
Originality	.432	.139/1.474	.062/1.168	.004/1.009
Self-confidence	.272	.125/1.850	.087/1.470	.024/1.097
Height	.296	.081/1.376	.055/1,228	.037/1.143

Note. See text for reliabilities used in calculating the estimated correlation adjustments.

Table 3 gives the correlations that motivated this article. For purposes related to an article on another topic completely, we happened to calculate the correlations between the 6-Day Sample MHT and TMIQ and personality and height measures (all taken in youth) in the age-77 Follow-Up Sample and in those who had survived to be potentially eligible for the Follow-Up Study. These correlations were not of intrinsic interest in and of themselves, and generally would not be. But the opportunity to make such calculations in naturally selected samples for which the full-sample correlations can also be made is rare. As is common, the Follow-Up–participant correlations were considerably lower than the full-sample correlations. This was not the case for the survivor-sample correlations: They were much closer to the full-sample correlations, and some were higher but others lower, though none reproduced the full-population correlation exactly. We applied the methods outlined above (omitting the approximate adjustment for direct selection often made in the absence of knowledge of the full-population standard deviation) to adjust the sample correlations for direct and indirect selection and made maximum likelihood estimates of the full-sample correlations based on both selected-sample correlations. For the adjustments for indirect selection, we had to assume test–retest reliabilities for some of the measures, and those noted in the Method section for Originality and Self-confidence were of necessity particularly arbitrary. All else being equal, lower reliabilities generate higher adjusted correlations and vice versa, so we were able to assess the impacts of likely differences between our assumed reliabilities and actuals. As the table shows, all the adjusted estimates differed from the actual full-population correlations, some of them considerably. The magnitudes and directions of deviation also differed somewhat systematically for the two cognitive tests and the adjustment methods applied. Though selection on cognitive ability appeared to be involved in both Follow-Up Study participation and survival, the specific processes involved appeared to differ, as the patterns of deviation in the two samples differed as well.

For example, for the MHT, some of the adjustments overstated the full-population correlations, while others understated them, and over- and understatement were consistent in the two samples for one correlation, but not the other two. In contrast, for the TMIQ, all the adjustments understated the full-population correlations, and most of the understatements were quite a bit larger in absolute magnitude than those with the MHT. The TMIQ in this sample was rather unusual in that its standard deviation remained effectively the same no matter the degree of naturally occurring selection that took place, and the arithmetic of the direct- and indirect-selection adjustment formulas generated these understatements. The maximum likelihood estimates tended to be most accurate in both samples, and it appeared that participation selection was much more directly on MHT score than on TMIQ score, given that its correlations adjusted for direct were more accurate than those for indirect range restriction, but the opposite was the case for the TMIQ. This was especially puzzling given the consistency of the biasing of estimates of the population correlation by all methods for this test, which would tend to suggest more overt sample selection based on it. In contrast, if anything, this situation was reversed for survival selection, though less consistently so. For the MHT, the adjustment for indirect selection that relied on our assumed test–retest reliabilities overestimated all the full-sample correlations, but it underestimated them for the TMIQ. This meant that, had we assumed lower reliabilities, the formula would have generated greater errors for the MHT correlations, but smaller errors for the TMIQ correlations, but would not have altered the overall extent of formula misestimation.

It was this consistent presence of differences between the actual population-level correlations and their estimates using commonly used or considered methods to offset the clearly present range restriction (especially in the follow-up participant sample) that led us to evaluate how common such deviations might be and whether, if common, there might be consistent patterns in them that could be more clearly revealed.

To evaluate this, we needed many more test scores in a single sample, preferably ranging in size of correlation in systematic ways. To accomplish this, we turned to the MISTRA sample with its 42 cognitive ability tests and personality measures as described above.

MISTRA Sample Results

Basic Statistics. Table 4 shows the descriptive statistics for each of the 42 cognitive ability tests administered in MISTRA in the full and two selected samples. In the full sample, all the tests would be considered very reasonably normally distributed: All had means and standard deviations close to those expected for standardized variables, with the deviations reflecting having winsorized outliers that occurred in removing effects of age and sex from the variables. None of the skews exceeded the commonly applied rule-of-thumb level of 1.00 in absolute value, and most were substantially below that. Still, some of the tests generated small negative skews, while others generated small positive skews. Kurtosis levels also were small, but some tests generated scores more concentrated around 1 *SD* in absolute value than would be expected in a strictly normal distribution (Moors, 1986) (Moors, 1986), while others did the opposite. This variety without extremity was a good feature of these data for our purpose, as would be expected if there were no systematic biases in either the participant or test samples.

Table 5 shows the correlations between each of the 42 MISTRA cognitive ability tests and Flynn-adjusted IQ and Multidimensional Personality Questionnaire Well-being, Absorption, and Alienation in the full and two selected samples. The magnitudes of these correlations reflected both individual measure reliability and content. As would be expected, all the correlations with Flynn-adjusted IQ were positive and generally at least moderate, ranging from .201 for CAB Immediate Visual Memory to .650 for WAIS Vocabulary. Those with Well-being were also almost all positive, though generally small. Absorption correlations were also generally positive, but even smaller, and several were negative. Alienation correlations were negative, and small to moderate. About a third of the correlations were weaker in the selected than in the full sample, as is often assumed to be the case when range is known to be restricted but no further information is available, but often the difference was tiny. And about two thirds of the correlations were actually stronger in the selected samples; some also took the opposite direction. This suggests quite strongly that the common assumption that correlations in range-restricted samples are weaker than the full-population correlations is at best tenuous.

Accuracy of Formulas to Adjust for Range Restriction. The degree to which the adjustment formulas intended to account for direct selection could recover the full-sample correlations in the two selected samples is recounted in Table 6. At the top of the table, we show statistics about the differences between the full-sample correlations and the subsample correlations, followed by the same kinds of comparisons for the estimates of the full correlations based on application of the adjustment formulas. The means of

Table 4. MISTRA Full and Selected Samples—Descriptive Statistics for Age-Sex–Adjusted Cognitive Test Scores.

Cognitive test	Full sample				First selected sample				Second selected sample			
	Mean	*SD*	Skew	Kurt.	Mean	*SD*	Skew	Kurt.	Mean	*SD*	Skew	Kurt.
Hawaii Battery												
Vocabulary	.010	.991	–.318	–.716	.074	.923	–.372	–.522	.181	.962	–.461	–.645
Immed Vis Mem	.016	.933	–.685	.353	.045	.940	–.654	.273	.113	.870	–.822	.471
Subtract/Multipl	–.001	.993	.393	–.057	.044	.980	.373	.026	–.002	.948	.257	–.297
Line Dots	–.011	.988	.170	.064	.029	1.013	.124	.345	.050	.976	.074	–.094
Word Begin/End	.003	.995	.227	–.191	.034	.920	.130	–.214	.178	1.039	.053	–.045
Card Rotation	–.010	.956	.443	.476	.010	.493	.726	.908	–.076	.989	.584	.459
Delay Vis Mem	.014	.990	–.253	–.146	.089	1.005	–.331	–.157	.154	.941	–.210	–.006
Pedigrees	.011	.994	–.328	–.016	.036	.964	–.484	.075	.143	.995	–.199	–.460
Mental Rotation	–.001	.985	.113	–.203	.047	.975	.112	.253	.109	.940	.205	–.051
Identical Pictures	.001	1.004	.225	–.256	.046	1.010	.147	–.267	.116	1.005	.279	–.188
Paper Form Bd	.005	1.001	.460	–.082	.068	.977	.589	.227	.160	1.091	.467	–.343
Hidden Patterns	–.002	1.000	–.161	–.297	.123	1.027	–.293	–.359	.117	.998	–.262	–.250
Things	–.003	.994	.262	–.124	.002	1.016	.418	–.007	.160	1.039	.108	–.389
Different Uses	–.003	.998	.083	–.333	–.017	.973	–.005	–.266	.180	1.037	.060	–.439
Cubes	–.001	.993	.334	.186	–.126	1.050	.343	–.208	.164	.980	.453	.299
Paper Folding	.002	.999	.053	–.311	.038	1.046	.043	–.467	.113	1.038	.163	–.319
CAB												
Vocabulary	.010	.982	–.738	–.124	.086	.915	–.661	–.240	.182	.998	–.974	.317
Proverbs	–.009	.995	–.886	.223	.032	.946	–1.096	.767	.134	.984	–1.057	.661
Number	.005	.996	.248	.017	.009	.935	.278	–.147	.094	1.079	.156	–.333
Spatial	.009	.998	.047	.286	–.056	1.014	–.072	.304	.087	.962	.310	.178
Speed of Closure	–.001	1.000	.552	.330	–.117	.986	.729	.511	–.037	.993	.559	.265
Perceptual Speed	–.001	.995	.346	–.022	.011	1.061	.335	–.323	.057	.951	.278	–.164
Induction	.001	.994	.151	–.322	.122	.978	–.242	–.183	.120	1.019	.203	–.355
Flex of Closure	–.004	.998	.054	–.763	.051	1.009	.025	–.792	.144	.991	–.025	–.802
Assoc Memory	.010	1.002	.517	–.573	.140	1.007	.398	–.543	.115	.973	.418	–.659
Mechanical Reas	.011	.994	–.249	–.330	.091	.950	–.116	–.296	.246	.955	–.226	–.513
Memory Span	.004	.986	–.023	–.426	–.015	.981	–.070	–.138	.076	1.004	–.094	–.407
Meaningful Mem	.000	1.002	–.362	–.648	.116	.981	–.380	–.602	.108	.987	–.444	–.511
Spelling	.006	.994	–.177	–.975	–.022	.951	–.268	–.826	.145	1.013	–.392	–.815
Word Fluency	.013	.999	.133	.286	.042	.923	.220	.928	.168	1.041	–.107	.093
WAIS												
Information	–.023	.992	–.229	–.444	.066	.965	–.380	–.475	.115	1.009	–.246	–.522
Comprehension	–.018	.994	–.516	–.134	.052	.967	–.301	–.186	.196	.912	–.612	.263
Arithmetic	.003	.997	–.129	–.733	.114	.951	–.059	–.796	.124	1.006	–.321	–.652
Similarities	–.008	.997	–.661	.160	.025	.944	–.648	.382	.080	.992	–.609	–.154
Digit Span	–.006	.991	.313	–.559	–.002	.954	.353	–.556	.040	.983	.199	–.501
Vocabulary	–.015	.994	–.481	–.191	.060	.879	–.779	.572	.212	.979	–.598	.022
Digit Symbol	.007	.999	–.144	–.302	.037	.940	–.263	–.280	.097	.982	–.135	–.182
Picture Completion	–.007	.989	–.354	–.025	.053	1.036	–.555	.164	.075	.958	–.315	–.178
Block Design	.004	.990	–.280	–.113	.026	1.035	–.399	–.023	.114	1.035	–.492	.239
Pattern Arrangement	–.002	.999	–.221	–.355	.106	.973	–.218	–.512	.097	.997	–.277	–.587
Object Assembly	–.004	.985	–.659	.093	–.020	.997	–.723	.206	.066	1.008	–.644	–.199
Raven	–.002	.998	–.668	–.090	.059	.977	–.487	–.352	.111	.999	–.694	–.131
Mean	.002	.992	–.081	–.176	.037	.968	–.087	–.089	.115	.992	–.128	–.189

Note. MISTRA = The Minnesota Study of Twins Reared Apart; CAB = Comprehensive Ability Battery; WAIS = Wechsler Adult Intelligence Scale.

Table 5. Full- and Selected-Sample Correlations by Cognitive Test.

Cognitive test	Full sample				First selected sample				Second selected sample			
		MPQ				MPQ				MPQ		
	Flynn-adjusted FSIQ	Well-being	Absorption	Alienation	Flynn-adjusted FSIQ	Well-being	Absorption	Alienation	Flynn-adjusted FSIQ	Well-being	Absorption	Alien ation
Hawaii Battery												
Vocabulary	.597	.068	.151	–.298	.600	.175	.157	–.345	.590	.018	.229	–.316
Immed Vis Mem	.201	.047	–.040	–.142	.165	.052	–.016	–.174	.173	.120	–.014	–.131
Subtract/Multipl	.384	.098	–.044	–.264	.461	.180	.004	–.316	.551	.055	.040	–.211
Line Dots	.264	.055	.066	–.142	.251	.141	.033	–.113	.322	.069	.196	–.220
Word Begin/End	.526	.181	.124	–.263	.557	.249	.161	–.297	.573	.208	.139	–.273
Card Rotation	.286	.212	.001	–.158	.255	.216	–.050	–.221	.355	.211	.179	–.248
Delay Vis Mem	.269	.073	.059	–.156	.261	.139	.030	–.229	.340	.099	.098	–.217
Pedigrees	.569	.190	.108	–.301	.648	.255	.096	–.416	.609	.239	.237	–.346
Mental Rotation	.314	.116	.079	–.113	.403	.159	.068	–.255	.323	.081	.248	–.197
Identical Pictures	.417	.127	.106	–.202	.479	.175	.106	–.297	.401	.117	.237	–.283
Paper Form Bd	.490	.215	.122	–.172	.492	.224	.096	–.199	.482	.225	.309	–.335
Hidden Patterns	.484	.164	.098	–.212	.542	.229	.097	–.314	.549	.131	.250	–.212
Things	.453	.134	.172	–.114	.499	.127	.191	–.176	.425	.202	.272	–.107
Different Uses	.483	.179	.232	–.165	.554	.194	.252	–.243	.471	.282	.306	–.221
Cubes	.440	.199	.057	–.241	.398	.218	.008	–.281	.449	.220	.190	–.329
Paper Folding	.460	.232	.036	–.303	.477	.276	.004	–.400	.463	.109	.191	–.362
CAB												
Vocabulary	.597	.090	.162	–.317	.580	.165	.140	–.418	.633	.147	.229	–.338
Proverbs	.545	.066	.206	–.232	.595	.115	.170	–.267	.504	.025	.255	–.156
Number	.511	.135	.039	–.270	.610	.212	–.009	–.357	.584	.161	.173	–.260
Spatial	.321	.192	.019	–.184	.306	.202	–.005	–.211	.343	.242	.171	–.350
Speed of Closure	.339	.148	.104	–.205	.482	.166	.042	–.313	.393	.056	.210	–.226
Perceptual Speed	.351	.088	–.003	–.207	.432	.146	.004	–.276	.396	.079	.206	–.221
Induction	.463	.130	.021	–.218	.464	.180	–.006	–.273	.500	.099	.101	–.167
Flex of Closure	.440	.131	.053	–.210	.503	.236	–.042	–.299	.463	.204	.208	–.216

(continued)

Table 5. (continued)

Cognitive test	Full sample				First selected sample				Second selected sample			
		MPQ				MPQ				MPQ		
	Flynn-adjusted FSIQ	Well-being	Absorption	Alienation	Flynn-adjusted FSIQ	Well-being	Absorption	Alienation	Flynn-adjusted FSIQ	Well-being	Absorption	Alien ation
Assoc Memory	.324	.142	.044	–.244	.377	.169	.023	–.329	.312	.165	.152	–.286
Mechanical Reas	.419	.143	.040	–.142	.452	.190	.009	–.232	.433	.096	.123	–.206
Memory Span	.415	.093	–.005	–.256	.502	.121	.029	–.296	.420	–.002	.128	–.129
Meaningful Mem	.472	.118	.078	–.285	.498	.242	.143	–.343	.464	.088	.149	–.279
Spelling	.540	.126	.067	–.314	.585	.208	.137	–.350	.559	.102	.151	–.277
Word Fluency	.515	.146	.071	–.256	.543	.220	.119	–.403	.574	.162	.117	–.282
WAIS												
Information	.634	.149	.184	–.256	.655	.151	.160	–.326	.535	.133	.294	–.269
Comprehension	.579	.093	.233	–.173	.563	.120	.241	–.246	.517	.145	.311	–.244
Arithmetic	.587	.156	.041	–.273	.649	.235	.020	–.336	.559	.172	.100	–.265
Similarities	.571	.145	.127	–.255	.569	.161	.092	–.307	.461	.133	.239	–.218
Digit Span	.405	.122	.013	–.182	.486	.202	.020	–.190	.409	.058	–.026	–.121
Vocabulary	.650	.100	.160	–.290	.676	.175	.129	–.360	.618	.155	.219	–.288
Digit Symbol	.464	.119	–.015	–.278	.547	.183	.013	–.360	.457	.054	.052	–.258
Picture Completion	.480	.107	.085	–.198	.490	.057	.043	–.269	.477	.019	.147	–.265
Block Design	.498	.204	.095	–.219	.513	.244	.082	–.307	.543	.172	.323	–.274
Pattern Arrangement	.436	.103	.170	–.087	.473	.161	.158	–.154	.437	.028	.256	–.042
Object Assembly	.389	.096	.101	–.106	.398	.026	–.001	–.163	.464	.008	.223	–.212
Raven	.538	.136	.052	–.255	.531	.152	.023	–.318	.602	.029	.170	–.262
Mean	.455	.133	.083	–.218	.489	.177	.071	–.285	.470	.122	.185	–.241

Note. CAB = Comprehensive Ability Battery; WAIS = Wechsler Adult Intelligence Scale; MPQ = MPQ = Multidimensional Personality Questionnaire.

Table 6. Means and Ranges of Effectiveness of Range-Restriction Adjustments to Selected-Sample Correlations for Direct Selection.

	First selected sample				Second selected sample			
		MPQ				MPQ		
	Flynn-adjusted FSIQ	Well-being	Absorption	Alienation	Flynn-adjusted FSIQ	Well-being	Absorption	Alienation
Mean full-sample correlation	.455	.133	.083	–.218	.455	.133	.083	–.218
Range of correlations	(.321, .650)	(.047, .232)	(–.044, .233)	(–.314, –.106)	(.321, .650)	(.047, .232)	(–.044, .233)	(–.314, –.106)
Mean correlation difference	–.033	–.045	.004	.067	–.015	.011	–.106	.023
Range of differences	(–.143, .042)	(–.124, .070)	(–.089, .095)	(–.030, .147)	(–.167, .110)	(–.103, .123)	(–.294, .039)	(–.127, .163)
Mean, absolute values of differences	.042	.051	.029	.069	.038	.047	.108	.049
Mean correlation ratio	.941	.829	3.038	.768	.970	.728	.369	.965
Range of differences	(.703, 1.218)	(.389, 3.743)	(.020, 70.0)	(.664, 1.266)	(.791, 1.238)	(.388, –41.1)	(.006, 2.019)	(.501, 2.083)
Mean absolute correlation ratio	1.090	1.346	4.379	1.244	1.083	2.942	1.669	1.237
Common approximation								
Mean range-restriction adjustment	1.003	1.003	1.003	1.003	1.018	1.018	1.018	1.018
Range of adjustments	(.917, 1.090)	(1.090, .917)	(.917, 1.090)	(.917, 1.090)	(.938, 1.131)	(.938, 1.131)	(.938, 1.131)	(.938, 1.082)
Mean-adjusted correlation	.488	.177	.079	–.285	.480	.124	.081	–.245
Range of adjusted correlations	(.177, .687)	(.025, .234)	(–.051, .288)	(–.416, –.144)	(.172, .698)	(–.002, .289)	(–.052, .296)	(–.362, –.043)
Mean-adjusted correlation difference	–.033	–.044	.004	.067	–.025	.008	.001	.027
Range of differences	(–.146, .039)	(–.125, .071)	(–.084, .094)	(–.029, .154)	(–.174, .084)	(–.110, .128)	(–.087, .096)	(–.126, .171)
Mean, absolute value of differences	.042	.051	.029	.069	.043	.049	.030	.050
Mean-adjusted correlation ratio	.939	.830	3.083	.767	.954	.714	3.171	1.257
Range of adjusted correlation ratios	(.698, 1.136)	(.378, 3.830)	(–.020, 72.7)	(.423, 1.251)	(.688, 1.172)	(.391, –40.9)	(–.019, 76.2)	(.523, 3.593)
Mean absolute correlation ratio	1.088	1.352	4.465	1.245	1.090	2.934	4.536	1.340

(continued)

Table 6. (continued)

	First selected sample				Second selected sample			
	Flynn-adjusted FSIQ	MPQ			Flynn-adjusted FSIQ	MPQ		
		Well-being	Absorption	Alienation		Well-being	Absorption	Alienation
Full adjustment for direct selection								
Mean further adjustment	1.001	1.000	1.000	1.000	.994	1.000	.999	.999
Range of further adjustment	(.971, 1.029)	(.998, 1.004)	(.993, 1.003)	(.994, 1.010)	(.951, 1.013)	(.996, 1.001)	(.996, 1.005)	(.991, 1.006)
Mean further-adjusted correlation	.489	.178	.079	–.286	.476	.124	.081	–.245
Range of further-adjusted correlations	(.177, .682)	(.025, .267)	(–.051, .286)	(–.416, –.114)	(.172, .664)	(–.002, .289)	(–.052, .295)	(–.359, –.043)
Mean-adjusted correlation difference	–.034	–.044	.004	.067	–.021	.009	.001	.027
Range of differences	(–.146, –.038)	(–.112, .071)	(–.084, .094)	(–.029, .153)	(–.172, .090)	(–.075, .128)	(–.087, .093)	(–.068, .170)
Mean, absolute value of differences	.041	.051	.029	.069	.040	.048	.030	.050
Mean-adjusted correlation ratio	.936	.829	3.083	.767	.960	.714	3.167	.948
Range of adjusted correlation ratios	(.699, 1.139)	(.378, 3.830)	(–.020, 72.7)	(.424, 1.251)	(.691, 1.170)	(.391, –40.9)	(–.019, 76.0)	(.457, 2.028)
Mean absolute correlation ratio	1.085	1.351	4.465	1.245	1.086	2.934	4.532	1.233

Note. MPQ = Multidimensional Personality Questionnaire. Differences are full (actual)-sample correlation less selected (sub)sample correlation, after adjustment as labeled. Absolute correlations ratios are $|(1 - \text{ratio}) + 1|$. For the ranges of ratios, the extremes are expressed as the greatest absolute ratios to 1; this does not always note the largest sign shift. FSIQ is (WAIS) full-scale IQ.

the differences between the full- and subsample correlations look rather small, but this obscures that for some tests the subsample correlations were higher than the full-sample correlations, and the reverse was true for others. Reflecting this, in all cases, the means of the absolute values of the correlation differences were higher than the means of the raw correlation differences, and the overall adjusted subsample averages tended to be stronger than the actual averages. These were sometimes trivially so, other times quite substantially so. Extent of the former depended on variabilities of under- and overstatement of the adjustments, the latter on relations between magnitudes of under- or overstatement and correlation. The correlation ratios depict this, as they represent the differences in proportion to the raw differences. Because researchers tend to assume that range restriction suppresses correlations relative to their full-population levels, the not-uncommon presence of higher subsample than full-sample correlations is important. Proportions of overstated subsample correlations were greater when the means of the absolute values of the differences were relatively larger than those of the raw differences. The correlation ratios simply reinforced these observations.

Comparing results for the two subsamples offered further insight. First, it emphasized the weakness of the common assumption that range-restriction results in underestimation of correlations. Recall that indirect selection on IQ was greater in Subsample 2 than Subsample 1. Contrary to expectations based on this, at mean level, the correlation differences from the full-sample correlations were smaller in Subsample 2 than Subsample 1. Explanation for this comes from the means of the absolute values of the correlation differences. Though these were also smaller in Subsample 2 than Subsample 1, with the exception of that for Absorption, this was true to a much smaller degree than for the mean raw differences (e.g., in absolute value, the Subsample 1 raw and absolute means differed by .012 [.04, –.03]) for FSIQ, whereas those for Subsample 2 differed by .023 [.03, –.01]). This indicated that there was more variance in sign of the differences in the more highly selected Subsample 2, so that ability to guess at direction of bias *decreased* with greater indirect selection. The correlation ratios make clear that the degree to which this was true was a function of size of correlation: Bias was more likely to take the commonly assumed direction for the higher FSIQ correlations than for the lower personality correlations. Even this was not completely reliable, however, as all bets on direction and degree of bias were essentially off for the very low raw Absorption correlations, yet the mean absolute correlation ratio was moderate.

In general, adjustments resulting from the commonly used formula for direct selection were tiny, making little difference, especially in the first subsample. In the second, more selected, subsample, however, the differences between actual and estimated full-sample correlations were sometimes larger after this adjustment (compare raw and adjusted mean absolute correlation differences and ratios in Table 6). The ranges of adjusted correlations and the mean absolute correlation ratios made clear that this was mostly when raw correlations were low. Applying the full adjustment for direct selection improved this situation, but not much.

Table 7 presents analogous information for the adjustment for indirect selection and MLE of the full-sample correlation based on the subsamples. In general, the adjustment for indirect selection overestimated the full-sample correlations, sometimes rather badly, so that it was more inaccurate than the full adjustment for direct selection. A primary reason for overstatement was inaccuracy of the assumption that indirect selection results in range restriction that understates correlations. That is, in many cases it had not, and even when it did, it did not do so as much as assumed in the indirect selection adjustment process. Still, accuracy was slightly better in the more selected second subsample, except for the very low correlations with Absorption. This suggested that, when correlations were at least moderate, adjustment formula accuracy improved with greater sample selection. Maximum likelihood did considerably better, and, overall, slightly better than full adjustment for direct selection. This was true particularly in the second subsample, though again not for the lowest correlating Absorption.

Finally, the Le et al. (2016) Case V adjustment method failed slightly more than half the time, primarily due to subsample standard deviations that were larger than those of the full sample. The rather high frequency of this was probably due to sampling quirks in the relatively small ‘‘population’’ in this case, but norming samples for even widely used personality and intelligence tests are always far from completely population-representative and not uncommonly around the size of the full MISTRA sample. This means that reliance on standard deviation ratio is a rather serious limitation of this adjustment method. To be specific, of the 42 tests, the method failed for all four outcome tests 27 times in the first selected subsample. In the second, it failed for 18 for Flynn-adjusted FSIQ, 24 for each of WB and AL, and 18 for AB. Where it worked, it produced results very similarly lacking in accuracy to the others (see online supplemental table).

General Comments. We recommend first applying MLE, applying adjustments as completely as possible for direct selection preferred over the method used here for indirect selection, which is intended to be the most accurate available, even when selection is very likely indirect. Applying both maximum likelihood and adjusting for direct selection, comparing results, and interpreting accordingly is better still.

The adjustment formulas all tended to be somewhat more accurate when applied to positive than to negative correlations. This can be seen by comparing the means of the absolute values of the correlation differences for FSIQ and Well-being with those for Absorption, keeping in mind that the adjustment formulas tended to be somewhat more accurate for higher correlations. This is what made this tendency more apparent in the absolute correlation ratios. It was, however, far from completely reliable. Still, it is probably a good idea to reverse one of the measure’s scoring when the raw sample correlation is negative and estimating extent of its sampling bias is a goal.

Greater skew and kurtosis also tended to be associated with greater distortion in estimated full-sample correlations. Unfortunately, however, as for estimates due to range restriction themselves, neither skew nor kurtosis directions nor magnitudes

Table 7. Means and Ranges of Effectiveness of Range-Restriction Adjustments to Selected-Sample Correlations for Indirect Selection and Maximum Likelihood Estimates of Full-Sample Correlations.

	First selected sample				Second selected sample			
	Flynn-adjusted FSIQ	MPQ				MPQ		
		Well-being	Absorption	Alienation	Flynn-adjusted	Well-being	Absorption	Alienation
Mean full-sample correlation	.455	.133	.083	−.218	.455	.133	.083	−.218
Range of correlations	(.321, .650)	(.047, .232)	(−.044, .233)	(−.314, −.106)	(.321, .650)	(.047, .232)	(−.044, .233)	(−.314, −.106)
Adjustment for indirect selection								
Mean-adjusted correlation	.578	.210	.092	−.343	.572	.154	.233	−.303
Range of adjusted correlations.	(.204, .778)	(.032, .307)	(−.058, .343)	(−.508, −.141)	(.196, .781)	(−.002, .574)	(−.022, .505)	(−.647, −.057)
Mean-adjusted correlation difference	−.123	−.077	−.010	.125	−.117	−.022	−.150	.085
Range of differences	(−.239, .030)	(−.169, .064)	(−.110, .099)	(−.001, .220)	(−.495, .034)	(−.362, .112)	(−.504, .048)	(−.113, .489)
Mean, absolute value of differences	.123	.082	.037	.125	.120	.065	.153	.095
Mean-adjusted correlation ratio	.794	.699	2.618	.639	.807	.492	.316	.780
Range of adjusted correlation ratios	(.559, .986)	(.317, 3.006)	(−.017, 63.6)	(.358, 1.009)	(.366, 1.057)	(.342, −37.8)	(.002, 1.790)	(.244, 1.791)
Mean absolute correlation ratio	1.206	1.425	3.984	1.361	1.199	2.702	1.722	1.299
Maximum likelihood estimates								
Mean estimated correlation	.483	.167	.071	−.279	.464	.114	.185	−.239
Range of estimated correlations	(.154, .607)	(.025, .276)	(−.042, .252)	(−.414, −.104)	(.162, .626)	(−.013, .256)	(−.026, .323)	(−.355, −.035)
Mean estimated correlation differences	−.027	−.034	.012	.061	−.009	.019	−.103	.021
Range of differences	(−.127, .071)	(−.099, .071)	(−.070, .102)	(−.018, .139)	(−.159, .099)	(−.067, .124)	(−.228, −.039)	(−.137, .160)
Mean, absolute value of differences	.040	.043	.032	.062	.037	.044	.105	.049
Mean estimated correlation ratio	.955	.889	−1.633	.781	.982	2.111	.421	.984
Range of estimated correlation ratios	(.728, 1.308)	(.389, 3.79)	(−020, −102.4)	(.449, 1.144)	(.707, 1.238)	(.410, 32.4)	(.539, 2.914)	(.509, 2.509)
Mean absolute correlation ratio	1.089	1.325	5.041	1.225	1.082	3.629	1.671	1.249

Note. MPQ = Multidimensional Personality Questionnaire. Differences are full (actual) sample correlation less selected (sub)sample correlation, after adjustment as labeled. Absolute correlations ratios are |(1 − ratio) + 1|. For the ranges of ratios, the extremes are expressed as the greatest absolute ratios to 1; this does not always note the largest sign shift. FSIQ is (WAIS) full-scale IQ.

appeared to be systematically associated with degree of distortion. For example, degree of skew was associated in the first sample with greater distortion in direct selection-adjusted correlations with FSIQ, but not in correlations with the personality measures (except for a statistical ‘‘tendency’’ for association with distortion of WB). This could suggest that skew might matter more when observed (positive) correlations are greater, as that is the primary common difference between the personality measures and FSIQ in this study, but there was no association in the second sample.

Instead there was an association with degree of kurtosis in the second sample, which showed up for AL too. It makes a certain amount of sense that kurtotic associations could be picked up in correlations of smaller magnitudes in negative correlations just as easily as in positive correlations in positively selected samples such as those that occur most commonly in research studies and were modelled here, as kurtosis reflects more symmetrical deviations from normality than does skew. It also makes sense that it would be picked up in AL in this study as its correlations tended to be the strongest among the personality measures, but they were negative, which might matter for skew but would be less likely to matter for the more symmetrical kurtotic measure of deviation from normality.

These tendencies for greater degree of distortion in range-restriction adjustment formulas with greater skew and/or kurtosis were not consistent in the two samples, Even in these at most very moderately skewed and kurtotic variables (whose deviations from normality would not usually arouse any concern) they were strong enough that examining results of adjustment formulas with and without transforming variables to improve normality is a good idea. Of course, all such transformations change the way the correlation should be interpreted, but we rarely have any sense that the numerical ratio or interval scales we often use to measure psychological constructs actually capture those constructs in those ways, so this is probably not as big a problem as it might superficially appear.

Examination of the specific effects of the same sample selections on the means, standard deviations, skews, and kurtoses in the 42 cognitive tests studied here suggests no clear patterns. Greater sample selectivity did tend to produce higher means, but even this did not happen consistently. Researchers often tend, at least implicitly, to assume that sample selection produces smaller standard deviations, but in these data this was not true about as often as it was true, and patterns for differences in skew were not more regular. Moreover, patterns of combinations of effects of sample selection on standard deviation, skew, and kurtosis were even less consistent. It was possible to examine individual tests and make some inferences about the particular distributional properties of that test in the full sample, but these inferences could not get beyond the inevitable confound between specific item properties of that test that would apply in any sample and the sample-specific clusterings of participants’ actual abilities in the areas items tested.

Conclusion

We began this article lauding researchers’ increasing awareness that participants in their studies tend not to represent the population from which they come very well.

We noted that this has been addressed in two ways, through development of new methods to adjust results to offset the resulting distortions in estimates and tests of their accuracy, and by increasing application of existing formulas in reporting estimated associations in actual study samples. These efforts fully deserve the praise we offered. Based on this study, however, we have to conclude that the accuracy tests run to data have not been sufficient to resolve the problem of distortions in estimated associations from population levels that selection into research participation creates. This is of course of necessity a subjective judgment; some may see the deviations we picked up as too small to matter, noting that no adjustment formula could ever be expected to recover the intended statistic exactly. Several of the deviations in both tests in the 6-Day Sample were not small in absolute value and were quite large in relation to the observed correlations (see Table 3). The same was true of many of the specific test observations in the MISTRA sample, even when not true of the mean levels (see especially the ranges in Tables 6 and 7). We believe that the ratios of deviations to the observed correlations are particularly relevant because accuracy is always relative to the size of whatever is being measured. Lab tests are commonly considered relatively accurate when their results ''generally'' fall within 10% of actual values, both by lab course instructors and medical technicians. By this standard, 68% of the common approximations, 68% of the full adjustments for direct selection, 90% of the adjustments for indirect selection, and 73% of the EM adjustments would be considered inaccurate. Doubling tolerance to 20%, 49% of the common approximations, 47% of the full adjustments for direct selection, 72% of the adjustments for indirect selection, and 51% of the EM adjustments would be considered inaccurate. We suggest that either psychological measurement somehow merits being held to much lower standards of accuracy, or these adjustment formulas too often do not perform adequately.

The deviations we observed were apparently due to too-restrictive assumptions about the shapes of the distributions cognitive ability tests take in actual populations and the natures of the distortions in those distributions imposed by the multifaceted, largely indirect selection processes that get some targeted people to participate in research studies and others not. Sources of deviation from normality could not be pinpointed exactly in this study because sampling variability is inevitably confounded in empirical data with systematically selective participation (whether direct or indirect, random, or nonrandom), population-level deviations from normality, and incompletely uniform item coverage of the construct range, but the common ''symptom'' of all these sources of ''malaise'' was small and very typical deviations from normality in the population-level full MISTRA sample. More work is needed to develop better adjustment formulas and test them more thoroughly using the kinds of distributions that measures in common use actually produce at population levels, and models of the processes that reflect how people actually sort into both research and personnel selection samples. This work needs to recognize that such distributions basically always differ somewhat from the normal. Researchers producing estimates in existing study samples also need to be much more circumspect than they often have been in assuming that they have some idea about what form of distortion the particular patterns of selection into their samples has created. This may be especially important for meta-analysis, as its results are often considered to be more generally applicable, and the distortions in estimates that sample-selection adjustment formulas can contain may actually make them less so.

Authors' Note

The data reported in this article have been previously published and were collected as part of two completely independent larger data collections, one of which ended more than 10 years before the other began. Findings from each data collection effort have been reported in multiple separate articles, numbering in the dozens if not hundreds, over a period of 30 years. No other article has used data from both of these completely independent studies in anything close to this manner.

Acknowledgments

Funding from the Biotechnology and Biological Sciences Research Council (BBSRC) and Medical Research Council (MRC) is gratefully acknowledged.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The 6-Day Sample follow-up study is supported by the U.K. Medical Research Council's contributions to the cross-council Life Long Health and Wellbeing Initiative (MRC G1001401/1). The work was undertaken by the University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross-council Lifelong Health and Wellbeing Initiative (MR/K026992/1).

Supplemental Material

Supplemental material is available for this article online.

References

Alexander, R. A. (1990). Correction formulas for correlations restricted by selection on an unmeasured variable. *Journal of Educational Measurement*, *27*, 187-189.

Alexander, R. A., Carson, K. P., Alliger, G. M., & Cronshaw, S. F. (1989). Empirical distributions of range restricted SDx in validity studies. *Journal of Applied Psychology*, *74*, 253-258.

Berry, C. M., & Sackett, P. R. (2007). Revisiting interview-cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology*, *60*, 837-874.

Berry, C. M., & Zhao, P. (2015). Addressing criticisms of existing predictive bias research: Cognitive ability test scores still overpredict African-Americans' job performance. *Journal of Applied Psychology*, *100*, 162-179.

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of *d* for range restriction and unreliability. *Organizational Research Methods*, *4*, 46-61.

Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences form a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods*, *15*, 386-397.

Brett, C. E., & Deary, I. J. (2014). Realising health data linkage from a researcher's perspective: Following up the 6-Day Sample of the Scottish Mental Survey 1947. *Longitudinal and Life Course Studies*, *5*, 283-298.

Bryant, N. D., & Gokhale, S. (1972). Correcting correlations for range restriction due to selection on an unmeasured variable. *Educational and Psychological Measurement*, *32*, 305-310.

Chernyshenko, O. S., & Ones, D. S. (1999). How selective are psychology graduate programs? The effect of the selection ratio on GRE score validity. *Educational and Psychology Measurement*, *59*, 951-961.

Deary, I. J., & Brett, C. E. (2015). Predicting and retrodicting intelligence between childhood and old age in the 6-Day Sample of the Scottish Mental Survey 1947. *Intelligence*, *50*, 1-9.

Deary, I. J., Pattie, A., & Starr, J. M. (2013). The stability of intelligence from age 11 to age 90 years: The Lothian Birth Cohort of 1921. *Psychological Science*, *24*, 2361-2368.

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J., & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology*, *86*, 130-147.

DeFries, J. C., Vandenberg, S. G., McClearn, G. E., Kuse, A. R., Wilson, J. R., Ashton, G. C., & Johnson, R. E. (1974). Near identity of cognitive structure in two ethnic groups. *Science*, *183*, 338-339.

Desai, M. (1952). The test-retest reliability of the progressive matrices test. *British Journal of Medical Psychology*, *25*, 48-53.

Enders, C. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1-16.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3-8.

Glass, G. V. (1982). Meta-analysis: An approach to the synthesis of research results. *Journal of Research in Science Teaching*, *19*, 93-112.

Greener, J. M., & Osburn, H. G. (1980). Accuracy of corrections for restriction in range due to explicit selection in heteroscedastic and nonlinear distributions. *Educational and Psychological Measurement*, *40*, 337-346.

Gross, A. L., & Fleischman, L. (1983). Restriction of range corrections when both distribution and selection assumption is violated. *Applied Psychological Measurement*, *7*, 227-237.

Hagglund, G., & Larsson, R. (2006). Estimation of the correlation coefficient based on selected data. *Journal of Educational and Behavioral Statistics*, *31*, 377-411.

Hakstian, A. R., & Bennett, R. W. (1977). Validity studies using the Comprehensive Ability Battery (CAB) I: Academic achievement criteria. *Education and Psychological Measurement*, *37*, 425-437.

Hakstian, A. R., & Cattell, R. B. (1975). *The comprehensive ability battery*. Champaign, IL: Institute for Personality and Ability Testing.

Hanson, R. K., Hunsley, J., & Parker, K. H. (1988). The relationship between WAIS subtest reliability, g-loadings, and meta-analytically derived validity estimates. *Journal of Clinical Psychology*, *44*, 557-562.

Hoffman, C. C. (1995). Applying range restriction corrections using published norms: Three case studies. *Personnel Psychology*, *48*, 913-922.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*, 594-612.

Johnson, W., & Bouchard, T. J. (2011). The MISTRA data: Forty-two mental ability tests in three batteries. *Intelligence*, *39*, 82-88.

Johnson, W., Bouchard, T. J., McGue, M., Segal, N. L., Tellegen, A., Keyes, M., & Gottesman, I. I. (2007). Genetic and environmental influences on the Verbal-Perceptual-Image Rotation (VPR) model of the structure of mental ability in the Minnesota Study of Twins Reared Apart. *Intelligence*, *35*, 542-562.

Johnson, W., Brett, C. E., Calvin, C., & Deary, I. J. (2016). Childhood characteristics and participation in Scottish Mental Survey 1947 Follow-Up Study: Implications for participation in aging studies. *Intelligence*, *54*, 70-79.

Johnson, W., Corley, J., Starr, J. M., & Deary, I. J. (2011). Psychological and physical health at age 70 in the Lothian Birth Cohort 1936: Links with early life IQ, SES, and current cognitive function and home environment. *Health Psychology*, *30*, 1-11.

Kuse, A. R. (1977). *Familial resemblances for cognitive abilities estimated from two test batteries in Hawaii* (Unpublished doctoral dissertation). University of Colorado at Boulder, Boulder, CO.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh*, *62*, 28-30.

Le, H., & Schmidt, F. L. (2006). Correcting for indirect range restriction in meta-analysis: Testing a new meta-analytic procedure. *Psychological Methods*, *11*, 416-438.

Little, R. J., & Rubin, D. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.

Littman, A. J., White, E., Satia, J. A., Bowen, D. J., & Kristal, R. (2006). Reliability and validity of two single items of psychosocial stress. *Epidemiology*, *17*, 398-403.

Lonnqvist, J. E., Paunonen, S., Verkasalo, M., Leikas, S., Tuulio-Henriksson, A., & Lonnqvist, J. (2007). Personality characteristics of research volunteers. *European Journal of Personality*, *21*, 1017-1030. doi:10.1002/per655

MacPherson, J. S. (1958). *Eleven-year-olds grow up*. London, England: University of London Press.

Maxwell, J. (1969). *Sixteen years on: A follow-up of the 1947 Scottish Survey*. London, England: University of London Press.

Mendoza, J. L. (1993). Fisher transformations and correlations corrected for selection and missing data. *Psychometrika*, *58*, 601-615.

Mendoza, J. L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics*, *12*, 282-293.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Moors, J. J. (1986). The meaning of kurtosis: Darlington re-examined. *American Statistician*, *40*, 283-284.

National Records of Scotland. (2013). *Vital events Reference Tables, 2013*. Retrieved from National Records of Scotland website: www.gro-scotland.gov.uk

Nishiwaki, Y., Clark, H., Morton, S. M., & Leon, D. A. (2005). Early life factors, childhood cognition and postal questionnaire response rate in middle age: The Aberdeen Children of the 1950s study. *BMC Medical Research Methodology*, *5*. doi:10.1186/1471-2288-5-16

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formula. *Journal of Applied Psychology*, *79*, 298-301.

Roth, P. L., Bobko, P., Switzer, F. S., III, & Dean, M. A. (2001). Prior selection causes biased estimates of standardized ethnic group differences: Simulation and analysis. *Personnel Psychology*, *2001*, 591-617.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology*, *55*, 807-825.

Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling*, *19*, 477-494.

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.

Scottish Council for Research in Education. (1933). *The intelligence of Scottish children*. London, England: University of London Press.

Scottish Council for Research in Education. (1949). *The trend of Scottish intelligence*. London, England: University of London Press.

Scottish Council for Research in Education. (1953). *Social implications of the 1947 Scottish Mental Survey*. London, England: University of London Press.

Segal, N. L. (2000). *Entwined lives: Twins and what they tell us about human behavior*. New York, NY: Plume.

Tarescavage, A. M., Fischler, G. L., Cappo, B. M., Hill, D. O., Corey, D. M., & Ben-Porath, Y. S. (2015). Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF) predictors of police officer problem behavior and collateral self-report test scores. *Psychological Assessment*, *27*, 125-137.

Taylor, A. (2004). The consequences of selective participation on behavioral genetic findings: Evidence from simulated and real data. *Twin Research*, *7*, 485-504.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In J. G. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality and assessment: Vol. 2. Personality measurement and testing* (pp. 261-292). Thousand Oaks, CA: Sage.

Thorndike, R. L. (1949). *Personnel selection*. New York, NY: Wiley.

Volken, T. (2013). Second-stage no-response in the Swiss Health Survey: Determinants and bias in outcomes. *BMC Public Health*, *13*, 167. doi:10.1186/1471-2458-13-167

Watkins, M. P. (1979). *Similarities between newlywed spouses (assortative marriage) with respect to specific cognitive abilities, socioeconomic status, and education*. Berkeley: University of California.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale*. New York, NY: Psychological Corporation.

Weiner, I. B. (2003). *Handbook of psychology*. Hoboken, NJ: Wiley.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.