An Investigation of the Generalizability of a College Subsample for Personality and Mental Health Research

Philip TROMOVITCH*

(Received October 2, 2023)

Although the reference standard for a sample in social science research is usually a nationally representative sample, in practice, college samples are widely used. The use of non-national samples raises concerns as to the validity and generalizability of the findings as a result of possible sample bias. *The Multinational Life Experience and Personality Project* (MLEPP) is collecting data from general population samples of adults aged 18 to 59 in multiple countries, and the current version of the MLEPP questionnaire asks respondents if they are college students. Consequently, by running an analysis on a multinational dataset, and then running the identical analysis on the college student subsample, it is possible to compare the results of using national samples versus using college samples. Similarly, by the use of matching, a subsample of aged-matched non-college data can also be analyzed to see if college student samples produce practically significantly different results than aged-matched non-college samples. The current paper presents an exploration of the generalizability of various sample types. It is concluded that although some small differences emerge between sample types, in terms of broad interpretation in social science research the use of college samples is non-problematic if participant age is not an important variable.

Keywords : generalizability, practical significance, research interpretation, MLEPP, MMCS1

1. Introduction

The Multinational Life Experience and Personality Project (MLEPP) is a large, multiphase, multinational set of studies. The MLEPP is collecting cross-sectional data in waves on a funds-available basis from adults aged 18 to 59; these data are combined to form larger samples for analysis. The second phase of the MLEPP started in September 2018 and data collection is expected to complete in three to five years. At the present time, data collection has completed in the United Kingdom, France, and Germany and has begun in six other territories. The data collection goal is to collect data from $N > 1000$ men and $N > 1000$ women in each country/territory in order to have samples large enough to analyze the possible effects of low prevalence experiences (e.g., those with a prevalence rate of 1%).

The MLEPP collects data on three mental health indicators: self-esteem, level of depressive symptoms, and level of anxiety symptoms. Personality traits assessed include: altruism, warmth, and being an understanding person. Intellectuality is also assessed. The aforementioned seven traits are assessed using multi-item scales comprised of *International Personality Item Pool*[1] items which were translated from the English versions into French and German for use in France and Germany, respectively. Each measure used in the present analyses is composed of 9 or 10 items, with each item using a 5-point Likert-like scale.

* Harris Science Research Institute, Doshisha University, Kyotanabe City, Kyoto 610-0394
Telephone: +81-774-65-6671, E-mail:ptromovi@mail.doshisha.ac.jp

The MLEPP additionally collects data on the degree to which respondents are comfortable with their sexuality. Comfort with sexuality is assessed using the activities-personal subscale of the *Multidimensional Measure of Comfort with Sexuality (MMCS1)*[2]. This measure is composed of 8 items, with each item using a 6-point Likert-like scale.

In addition, the MLEPP collects data on the respondents' family background using numerous author generated items (e.g., prior to age 16: socioeconomic status; verbal or physical fighting between parents; parental mental health; experiencing corporal punishment in the form of being hit, kicked, or punched; experiencing corporal punishment as having been abusive; being made to feel loved and cared for; receiving adequate provision of food, shelter, and medical care).

In social science research, one common concern is the possible biasing effects of varying sample types. In general, while samples designed to be nationally representative are considered to be the reference standard for unbiased samples, one of the most widely used sample types is college student samples. Given that a very large proportion of the population of developed countries goes to college (even if all attendees do not complete a college degree), a college sample is likely to have at least good representativeness for investigations that are not highly sensitive to participant age (college samples will almost always have a notably younger mean age than national samples). There can also be concerns with the use of college samples since it is generally not known to what degree socioeconomic status (SES) may differ from aged-matched non-college individuals (with an assumption that college students have a higher SES due to the financial cost of college education), how mental health may differ, how political views may differ, and how IQ (or intellectuality) may differ (with an assumption that college students have higher intellectuality and IQ due to the historically scholarly nature of college education).

The second phase of the MLEPP is collecting national data from adults aged 18 to 59. Respondents are recruited by market research firms (i.e., panel providers) which try to provide nationally representative samples. The questionnaire asks respondents if they are currently college students. Thus, the second phase of the MLEPP allows exploration of issues related to generalizability of findings from college samples by using the college student subsamples of the national datasets that are being collected. Furthermore, because of the large size of these datasets, it is possible to extract and analyze age-matched non-college samples as well, to compare the results of using college samples to the results from using non-college community members of similar age.

Depending on a researcher's concerns, the issue of interest might be to know if college student samples produce meaningfully different findings than age-matched non-college samples, or alternatively and more commonly, to know if college student samples produce meaningfully different findings compared to the use of national samples.

The goal of the present analyses is to explore these issues. The goal is not to see if college student samples produce *statistically* significantly different results because even trivial (but systematic) differences will be found to be statistically significant in large samples[3], but rather, to see if there is a *practically* significant difference based on sample type used for analysis.

All eight of the scales used for the present analyses can be scored on a scale of approximately 0 to 40 (i.e., a 5-point Likert-like scale can be scored from 0 to 4 and the items summed; hence a 10-item measure would have a range of 0 to 40). For the current article, the results for all eight measures were scaled to a 0 to 40 range. Consequently, a 1-point difference between groups can be taken to indicate that the two groups

responded equivalently except for one of ten items, on which the groups differed by only a single Likert-like scale point. Similarly, a 10-point difference between groups would indicate that, on average, all items on the measure differed by a single Likert-like scale point. For the present investigation, a practically significant difference was arbitrarily defined as a 2-point difference between groups -- readers may wish to define their own criterion for a practically significant difference before reading further.

2. The Samples

In order to ensure a sufficient sample size for the present analyses, the datasets collected from the United Kingdom, France, and Germany were combined. A separate dataset was created by copying the records of the college students from the multinational dataset. The SPSS case-control matching procedure was used to create a third dataset of non-college student data that is sex and age matched to the college students. The college student and age-matched non-college datasets were created by copying the relevant records -- not removing them -- from the multinational dataset. Due to missing data, the exact *N* for each analysis that follows varied from one analysis to another, but in all cases the *N* was greater than the values presented in Table 1.

3. The Investigations

3.1 *Levels of Traits, College vs National*

The first investigation examined the average difference between the multinational dataset and the college dataset on the eight variables of interest. The average difference across the female analyses was 0.83; the male data showed a similar average difference of 0.95; thus on average, there is no practically significant difference between college students and 18-59 year old adults on these eight traits. The largest difference for females was on intellectuality, showing a difference of 1.56 points (not a practically significant difference). The largest difference for males was 2.05, which occurred on the comfort with sexuality trait. Thus, there is arguably a practically significant difference between college students and general population adults on comfort with sexuality for males. It should be noted, however, that 16 difference values were calculated for these analyses (8 traits by 2 sexes) and this was the only difference to reach the 2.0 level.

3.2 *Levels of Traits, College vs Non-College*

The second investigation examined the average differences between the college dataset and the aged-matched non-college dataset. The average difference across the female analyses was 0.47; the male data showed a similar average difference of 0.48; thus on average, there is no practically significant difference between college students and age-matched non-college individuals on these eight traits. The largest difference for females was on depression, showing a difference of 1.13 points (not a practically significant difference). The largest difference for males was on intellectuality, showing a difference of 1.26 points (not a practically significant difference). Thus, even the largest difference among the 16 analyses was clearly not of practical significance.

3.3 *Causal Modeling*

The analyses just presented show that there is little or no practically significant difference between using college students and using national samples for assessing levels of personality and mental health traits. However, much social science research goes beyond measuring levels and tries to predict such traits from antecedents. Prior research has shown that the mental health and personality traits assessed by the MLEPP are substantially predicted by family background variables.

Table 1. Minimum sample sizes used.

Sample Type	Females	Males
Multinational	$N > 4000$	$N > 3900$
College	$N > 420$	$N > 460$
Non-College	$N > 375$	$N > 425$

Notes: The college and non-college samples are subsamples of the multinational dataset. The non-college samples are smaller than the college samples because exact matching on age was used and the multinational dataset did not contain appropriate matches for all college students.

Table 2. Depression levels predicted from family background variables -- female samples.

	Multinational Sample			College Sample			Non-College Sample		
Forward Regression	FBV	r^2	p-value	FBV	r^2	p-value	FBV	r^2	p-value
step 1	PMH	7.9%	<.001	VF	6.8%	<.001	PHM	10.7%	<.001
step 2	VF	9.6%	<.001	PMH	9.1%	<.001	VF	13.4%	<.001
step 3	SES	9.9%	<.001	loving	10.2%	.019	–	–	–
step 4	PF	10.1%	.010	–	–	–	–	–	–
step 5	loving	10.2%	.005	–	–	–	–	–	–
step 6	fsmc	10.4%	.003	–	–	–	–	–	–
step 7	cpA	10.5%	.024	–	–	–	–	–	–

Notes: r^2 = r-squared value after adjusting for the number of variables in the regression equation; p-value = statistical significance of the model at that step; FBV = family background variable which entered the model: VF = verbal fighting between parents; PF = physical fighting between parents; PMH = parental mental health; loving = made to feel important, loved, and cared for; SES = socioeconomic status; fsmc = adequate provision of food, shelter, and medical care; cpA = corporal punishment self-reported as abusive. The FBV which assessed being hit, punched, or kicked by their parents as part of corporal punishment did not enter any of the equations.

Table 3. Depression levels predicted from family background variables -- male samples.

	Multinational Sample			College Sample			Non-College Sample		
Forward Regression	FBV	r^2	p-value	FBV	r^2	p-value	FBV	r^2	p-value
step 1	PMH	10.2%	<.001	PMH	12.3%	<.001	loving	10.8%	<.001
step 2	loving	12.3%	<.001	SES	13.5%	.009	PMH	15.3%	<.001
step 3	VF	12.5%	.002	loving	14.3%	.022	SES	16.6%	.007
step 4	SES	12.7%	.004	–	–	–	–	–	–
step 5	cpA	12.8%	.020	–	–	–	–	–	–

Notes: Same as Table 2.

In research on Japanese adults it was found that levels of depressive systems was the trait best predicted by family background variables for both females and for males with these antecedents predicting approximately 10% of the variance in adult depression scores (female $r^2 = 9.9\%$; male $r^2 = 11.4\%$)[4].

In order to investigate a typical causal modeling approach, depression (level of depressive systems) was selected for the present analyses. Depression scores were predicted via multiple regression using the forward stepwise procedure to create models predicting depression from a collection of eight family background variables using the common "$p < .05$ to enter" criterion.

3.3.1 *Depression Modeling, Female Samples*

The three models generated from the three female samples appear in Table 2. The number of variables that entered each equation was monotonically related to the size of the samples: the larger the sample, the more variables that entered the equation. As can be seen comparing the three models, the first two variables that entered the equations were always the same (parental mental health & verbal fighting between the parents), though the order of entry was reversed in the college student sample.

Although the non-college sample model had only two variables enter, the somewhat larger college sample had an additional variable enter the equation with its third and final step (being made to feel loved and cared for). This variable (loving) also entered the multinational sample model in step 5, again demonstrating similarity among the results from different types of samples.

Interestingly, the smaller two-variable model of the non-college individuals explained more variance in depression scores (13.4%) than the other models (10.2% & 10.5%). It may also be of note that although the large multinational dataset which created greater statistical power allowed many more predictor variables to enter the regression equation, the difference in variance explained between the multinational 4-variable model and the final 7-variable model was negligible with the three additional predictor variables only explaining an additional 0.4% of the variance (10.1% vs. 10.5%).

3.3.2 *Depression Modeling, Male Samples*

As can be seen in Table 3, the college and the non-college male samples produced very similar models, each selecting three predictor variables. Although the order of entry was different between the two models, out of the eight possible predictor variables the exact same three variables entered the equations. Two of these three variables (parental mental health & loving) were the first two variables to enter the multinational model, and the third (SES) entered the multinational model in the fourth step, showing further similarity across sample types.

As occurred in the female analyses, the notably larger multinational sample made use of the most predictor variables (five, versus three in the college and non-college samples). Also, as with the female analyses, the non-college sample produced the largest r^2 value.

4. Observations & Conclusions

Regarding assessing general levels of the three personality traits examined in this research (altruism, warmth, & being an understanding person), the three mental health traits (self-esteem, depression, & anxiety), intellectuality, and comfort with sexuality, 32 comparisons were made. Only one produced a difference equivalent to two scale points or larger on a single item of a 10 item measure. These findings suggest that for general investigations, the use of college students will likely produce results that are sufficiently close to those that would be obtained by using national samples. Put another way, there is little or no practically significant difference in the results between

college samples and those of national samples for the traits assessed in the present research.

These analyses also found that the average differences in level of traits between college and age-matched non-college individuals is about half the size of the difference found between college and national samples. Since the larger differences occurred with the multinational sample comparisons, and the multinational samples have a higher average age, it seems likely that the small differences found in the degree of difference are largely a result of age, suggesting that age should be controlled for, when possible.

Regarding modeling psychological traits from family background variables, the analyses show that although the models produced from different sample types do show differences, in terms of broad interpretation the models are quite similar and can be considered to be essentially equivalent.

There is always a high risk of overinterpretation when scientists examine a single analysis, and it is easy to forget that when doing social science research there are almost always uncontrolled confounding variables. Furthermore, variables that are assessed and show replicable associations may in fact be proxy variables for a different trait that was not considered. For example, depression and anxiety are correlated; if research assessed only one of these variables, the researcher would have no way to know if the results were due to the trait they thought they were assessing or the trait for which that measure is also a proxy. This problem of intercorrelation can occur at other levels, for example, measures used to assess family background (e.g., verbal fighting between parents) may in fact be a proxy variable for something else (e.g., alcohol abuse; since parents may be more likely to fight when intoxicated, or parents might fight over the topic of alcohol use if one parent believes the other parent is drinking too much alcohol).

In conclusion, when considering the larger issues that are faced by social scientists when interpreting findings, it seems that the issue of possible bias due to using college samples, rather than, for example, national samples, is minor and not one which should be of substantial concern, especially in the early stages of investigations as well as in investigations with limited statistical power.

This research was supported in part by grants-in-aid from the Harris Science Research Institute of Doshisha University and a grant-in-aid from MEXT (grant number 21K02971).

References

1) L. R. Goldberg, "International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences". Available at http://ipip.ori.org/ updated September 23, 2019.
2) P. Tromovitch, "The Multidimensional Measure of Comfort with Sexuality (MMCS1)", 34-39. In Fisher et al. (Eds.) *Handbook of Sexuality-related Measures* (3rd ed.). Routledge: New York, New York, (2011).
3) P. Tromovitch, "The lay public's misinterpretation of the meaning of 'significant': a call for simple yet significant changes in scientific reporting". *Journal of Research Practice*, **11**(1), article P1 (2015).
4) P. Tromovitch, "The degree to which age and educational level predict pro-social personality and mental health among Japanese adults". *The Harris Science Review of Doshisha University*, **59**(4), 39-44 (2018).