COMPREHENSIVE PSYCHOLOGY

2015, Volume 4, Article 1
ISSN 2165-2228

DOI: 10.2466/03.CP.4.1

Received August 7, 2014
Accepted January 13, 2015
Published February 5, 2015

CITATION
Marchant, G. J. (2015) How plausible is using averaged NAEP values to examine student achievement? *Comprehensive Psychology*, 4, 1.

AmSci
Ammons Scientific
www.AmmonsScientific.com

How plausible is using averaged NAEP values to examine student achievement?[1]

Gregory J. Marchant

Department of Educational Psychology, Ball State University

Abstract

Software limitations and difficulty of use have led to some using averaged values instead of plausible values. This study compared the results of plausible values using AM Statistical Software with averaged values using SPSS for descriptive statistics multiple regressions using the 2009 4th and 8th grade math and reading National Assessment of Educational Progress (NAEP) for the total sample in addition to a subsample of Native American students. Gender and poverty and absenteeism and enjoyment of the subject material were used in separate equations to predict achievement. The subgroup means were almost identical, but differences were evident in the standard deviations. The regressions yielded similar results for both approaches. The results suggested that averaged values might present a viable, although not ideal, approach to plausible values.

In a perfect world of educational assessment, students might be administered annually a lengthy, unbiased, reliable, valid examination of their abilities. The sample would be the population, and every standard and skill would be adequately measured. The data would run easily using any common statistical package, allowing flexibility of analyses and choice of platform. Instead, we compromise in an effort to achieve working estimates within the real limitations. For student achievement, that means trying to obtain a representative sample of the population of each state and weighing each student's scores accordingly. To minimize the time burden, each student gets only a portion of the total test. Somehow, these selected students with their selected items must yield meaningful scores. For a number of large data sets, such as the National Assessment of Educational Progress (NAEP) test, this goal is achieved using "plausible values."

Plausible values are not test scores *per se*. They are computed approximations or estimates of scores (Monseur & Adams, 2009; Von Davier, Gonzales, & Mislevy, 2009). As is the case with NAEP, usually five plausible values are generated (Wu, 2004, 2005). The values then become part of a more complex analysis than used with simple scores. Although scripts can be written for popular statistical programs like SPSS or SAS, specialty programs like AM Statistical Software are often preferred for running plausible values analyses (American Institute for Research, 2014). Unfortunately, AM software also has its limitations. Although it does utilize fairly straightforward "drag and drop" and right-click procedures, knowing what goes where can be a bit of a challenge. For plausible values the analyses are limited to regressions, multiple regressions, and the generation of group means, standard errors, and standard deviations without an intuitive way to test the means for those groups. AM software also only runs on Windows, which may be problematic for Apple MacIntosh users. AM is currently described as "still in Beta release (2014)" and the following responses from Jon Cohen are presented on their Common Questions and Answers website (2002–2006):

1. The correlations procedure does not currently produce standard errors or significance tests.
2. Currently, AM tables procedures do not produce marginal estimates in the same run as cross-tabs.
3. AM does not currently do chi-squared tests.
4. The released version of AM does not estimate IRT models.
5. AM does not output odds ratios.

[1]Address correspondence to Gregory J. Marchant, Department of Educational Psychology, Ball State University, Muncie, IN or e-mail (gmarchant@bsu.edu).

6. AM can do a limited set of specific models (confirmatory factor analysis/structural equation modeling), but does not currently have a general capability to estimate these.

7. AM does not generate standardized coefficients.

8. AM does not currently calculate Cronbach's α.

9. Unfortunately, currently the only way to get AM to calculate correlations with pair-wise deletion is to calculate each correlation separately.

10. Currently AM does not have a test to determine whether the pattern of data across cells is likely to be random.

11. There is no F-test.

12. AM will not yet run on a MacIntosh.

13. AM will run without administrator rights, but someone with administrator rights must install the software.

Because of the simplicity and versatility of statistical programs like SPSS and SAS, researchers have looked for shortcuts to plausible values analyses. The most popular "fix" is to create *a score* by averaging the five plausible values. However, the NAEP Technical Documentation Website warns, "that appropriate point estimates of individual scale scores cannot be calculated by averaging the five plausible values attached to a student's file. Using averages of the five plausible values attached to a student's file is inadequate to calculate group summary statistics, such as proportions above a certain level or to determine whether group means differ from one another." It has been argued that the misuse of plausible values results in "a substantial departure from the estimates obtained from the correct and intended analysis" (Carstens & Hastedt, 2010). Carstens and Hastedt, in their study using the Trends in International Mathematics and Science Study data, found country means changed as much as almost 15% and standard errors by up to 50%; however, little difference was found in *t* test comparisons.

This article reports the comparison of descriptive statistics and multiple regression results using NAEP data. Plausible values analyses were run using AM software and averaged plausible values "scores" were run using SPSS. This paper is not so much a statistical or even conceptual comparison of approaches, but rather a practical comparison between plausible value analyses and the use of averaged plausible values in the same analyses. The question is not, "which of these techniques is the correct or most accurate approach"; running plausible values as estimate analyses is designed to avoid biases and yield the most accurate results with the structure and limitations of the data. However, for some the limitations of plausible values software in terms of ease of various analyses present real issues. The question is: "How different are the results when using the shortcut averaged plausible values 'score'?"

Method

The 2009 individual-level restricted-use NAEP data for Grades 4 and 8 math and reading were analyzed using SPSS and AM software. Student weights (ORIGWT) were applied for the SPSS analyses. For the AM software, student weights (ORIGWT), strata variable (REPGRP1), cluster variable (JKUNIT), and 62 replicate weights were set. Descriptive statistics were calculated by gender and eligibility for the free lunch program. Two separate multiple regression analyses were conducted predicting composite achievement. For one multiple regression, the dichotomous variables of gender and poverty were used. For another multiple regression, the continuous variables were days absent the previous month and how much they like math or how fun they think reading is. Each of these analyses was conducted for Grade 4 and Grade 8 and for reading and math. In addition to analyses for the total sample, the analyses were also run for just the Native American students (the smallest ethnic group). Hypothetically, the analyses could be testing the questions: (1) Do gender and poverty predict Grade 4 and 8 math and reading achievement for all students and just for Native American students? (2) Does absenteeism and enjoyment of the subject predict Grade 4 and 8 math and reading achievement for all students and just for Native American students?

Results

The answers to the research questions were the same regardless of the approach taken (see Tables 1 and 2). For the total sample, gender and poverty predicted math and reading achievement for both grade levels, with poverty exhibiting a stronger relationship. For the Native American sample, gender and poverty also predicted achievement; however, gender did not contribute a significant amount of unique variance for math achievement. Absenteeism and enjoyment were also significant predictors of math and reading achievement for both grade levels and both samples.

Descriptives

For the total sample, all of the scores for the averaged values were within one point of the plausible value mean scores. Eighty-nine percent of the averages were within one-tenth of a point, and 75% were within one-hundredth of a point of each other. Even for the subsample of Native Americans, all of the mean scores were within one point of the plausible value means. Eighty-six percent of the averages were within one-tenth of a point, and 68% were within one-hundredth of a point of each other.

The standard deviation for the plausible values averaged 1.37 higher than the average for the averaged scores for the total sample of 29.74. For the Native American population, the mean standard deviation for the plausible value values averaged 1.54 higher than the mean for the averaged scores of 31.29.

TABLE 1
Descriptive statistics and multiple regression results for NAEP math

	Grade 4		Grade 8	
	Plausible	Averaged	Plausible	Averaged
Total Sample	$N=3{,}188{,}890$	$N=142{,}370$	$N=3{,}739{,}840$	$N=161{,}680$
M (*SD*)	238.17 (28.67)	238.17 (27.28)	282.91 (36.39)	282.91 (35.05)
Male Scores	239.06 (29.49)	239.06 (28.29)	283.95 (37.22)	283.95 (35.91)
Poor	227.04 (27.68)	227.04 (26.27)	267.13 (34.98)	267.13 (33.38)
Non-Poor	250.51 (26.42)	250.51 (25.21)	294.67 (34.42)	294.67 (33.14)
Female Scores	237.24 (27.76)	237.24 (26.46)	281.86 (35.49)	281.86 (34.13)
Poor	226.68 (26.23)	226.67 (24.75)	265.67 (32.90)	265.67 (31.22)
Non-Poor	247.75 (25.06)	247.75 (23.79)	292.34 (32.83)	292.34 (31.50)
Multiple Regression				
Sex/Poor	$R^2=0.15$	$R^2=0.17$	$R^2=0.14$	$R^2=0.15$
Sex	$z=-7$	$\beta=-0.03$	$z=-7$	$\beta=-0.03$
Poor	$z=-66$	$\beta=-0.41$	$z=-56$	$\beta=-0.38$
Absent/Like	$R^2=0.05$	$R^2=0.05$	$R^2=0.07$	$R^2=0.08$
Absent	$z=-32$	$\beta=-0.18$	$z=-35$	$\beta=-0.18$
Like	$z=26$	$\beta=0.12$	$z=38$	$\beta=0.19$
Native American	$n=41{,}454$	$n=1{,}850$	$n=41{,}470$	$n=1{,}790$
M (*SD*)	225.28 (29.22)	225.37 (27.69)	265.58 (37.84)	265.58 (36.18)
Male Scores	225.81 (29.95)	225.84 (28.44)	267.43 (37.87)	267.57 (36.42)
Poor	221.23 (28.92)	221.23 (27.31)	258.27 (35.97)	258.27 (34.13)
Non-Poor	237.28 (29.28)	237.28 (28.00)	280.83 (36.95)	280.83 (35.49)
Female Scores	224.80 (28.54)	224.95 (27.01)	263.58 (37.07)	264.66 (35.67)
Poor	219.43 (27.76)	219.43 (26.13)	259.18 (36.02)	259.18 (34.14)
Non-Poor	236.66 (26.45)	236.66 (25.08)	274.01 (37.73)	274.01 (36.32)
Multiple Regression				
Sex/Poor	$R^2=0.07$	$R^2=0.08$	$R^2=0.06$	$R^2=0.07$
Sex	$z=-1$, $p=ns$	$\beta=-0.03$, $p=ns$	$z=-1$, $p=ns$	$\beta=-0.03$, $p=ns$
Poor	$z=-9$	$\beta=-0.28$	$z=-8$	$\beta=-0.26$
Absent/Like	$R^2=0.06$	$R^2=0.07$	$R^2=0.07$	$R^2=0.08$
Absent	$z=-9$	$\beta=-0.20$	$z=-8$	$\beta=-0.23$
Like	$z=6$	$\beta=0.16$	$z=6$	$\beta=0.15$

Note All $p<.001$ unless noted.

Multiple Regressions

All of the R^2 values were within 0.02 of each other, and all of the equations significantly predicted the achievement variable (see Tables 1 and 2). All of the coefficients were significant for the total sample, and none of the coefficients were significant for gender for the Native American students in reading. The relative strength of the two coefficients were similar across approaches; however, the relative difference in coefficients was somewhat greater for the averaged values. Such for the total sample predicting Grade 4 math scores, poverty was about nine times as strong as gender using plausible values, but poverty was about 14 times as strong as gender using averaged values.

Discussion

If different research reports were written for the two previously posed questions using the two different approaches, the results would not look much different and the interpretations could be almost identical. Yes, gender and poverty predict achievement, and poverty is a stronger predictor than gender. Boys have higher scores than girls in math, and girls have higher scores than boys in reading. Gender is not a significant factor in the equation for Native Americans in math. Absenteeism and positive feelings about the subject are related to achievement.

In general, the results and conclusions were similar for the analyses on the large-scale database using plausible values and averaged values. This suggests that using averaged values could be a viable option when software limitations are problematic. Although it may be viable, the use of averaged values should never be considered preferable. Plausible values were developed as a means of dealing with the limitations and imperfections of datasets like NAEP. One potentially ignores those limitations at one's own peril. Other research has raised concerns over the use of averaged values. As in-

TABLE 2
Descriptive statistics and multiple regression results for NAEP reading

	Grade 4		Grade 8	
	Plausible	Averaged	Plausible	Averaged
Total Sample	N=3,648,080	N=178,800	N=3,693,370	N=160,870
M (SD)	220.95 (35.48)	220.94 (33.32)	264.01 (34.33)	264.01 (32.27)
Male Scores	217.61 (36.15)	216.47 (33.97)	259.43 (34.69)	258.29 (32.60)
Poor	202.55 (35.15)	202.55 (32.81)	244.36 (34.05)	244.36 (31.92)
Non-Poor	228.70 (32.66)	228.70 (30.05)	268.12 (31.52)	268.12 (29.32)
Female Scores	224.37 (34.46)	223.46 (32.34)	268.64 (33.33)	267.57 (31.23)
Poor	209.60 (33.24)	209.60 (30.89)	253.15 (32.65)	253.15 (30.53)
Non-Poor	235.86 (30.70)	235.86 (28.31)	278.35 (29.56)	278.35 (27.13)
Multiple Regression				
Sex/Poor	R^2=0.15	R^2=0.17	R^2=0.14	R^2=0.16
Sex	z=25	β=0.11	z=30	β=0.15
Poor	z=−71	β=−0.39	z=−58	β=−0.38
Absent/Fun	R^2=0.05	R^2=0.06	R^2=0.11	R^2=0.12
Absent	z=−27	β=−0.13	z=−27	β=−0.15
Fun	z=38	β=0.20	z=55	β=0.31
Native American	n=40,520	n=1,990	n=39,910	n=1,740
M (SD)	203.95 (40.97)	203.95 (38.47)	251.23 (36.68)	251.23 (34.30)
Male Scores	200.53 (43.34)	200.14 (40.95)	245.95 (37.14)	246.23 (34.83)
Poor	192.57 (42.50)	192.57 (39.96)	239.87 (36.95)	239.87 (34.61)
Non-Poor	216.49 (40.83)	216.49 (38.27)	257.63 (34.57)	257.63 (32.27)
Female Scores	207.57 (37.95)	207.52 (35.44)	256.21 (35.52)	256.12 (33.24)
Poor	200.12 (38.02)	200.12 (35.40)	249.10 (35.25)	249.10 (32.65)
Non-Poor	222.18 (33.27)	222.18 (30.67)	267.84 (33.40)	267.84 (30.88)
Multiple Regression				
Sex/Poor	R^2=0.08	R^2=0.09	R^2=0.08	R^2=0.09
Sex	z=3	β=−0.09	z=4	β=−0.14
Poor	z=−7	β=−0.28	z=−9	β=−0.26
Absent/Fun	R^2=0.05	R^2=0.06	R^2=0.09	R^2=0.10
Absent	z=−6	β=−0.18	z=−4	β=−0.14
Fun	z=6	β=0.15	z=8	β=0.28

Note All p<.001.

dicated by Carsten and Hasredt (2010), differences in variance were a more serious issue than differences in means. Therefore, any analysis relying on variances is more likely to be influenced by using the averaged values.

The developers of AM software are trying to address the limitations of the program. The latest "Beta Version 0.06.00" has added graphics for the first time, as well as a couple more analyses. Unfortunately the question for many wanting to analyze individual level plausible values data is, do I do the analysis using averaged values or do I not do the analysis? Unless plausible values become an easy choice as part of popular statistical programs, research should continue to explore the utility and dangers in using alternative approaches. Despite concerns over the use of averaged plausible values, there is no evidence to suggest the results using them yield significantly different results.

References

American Institute for Research. (2014) *AM Software*. Washington, DC: Authors. Retrieved August 1, 2014, from http://am.air.org/.

Carsten, R., & Hasredt, D. (2010, June) The effect of not using plausible values when they should be: an illustration using TIMSS 2007 grade 8 mathematical data. Unpublished document, Institute for Objective Measurement, Durham, NC. Retrieved from http://www.rasch.org/rmt/rmt182c.htm.

Cohen, J. (2002-2006) *Common questions and answers: responses*. Retrieved August 1, 2014, from http://am.air.org/amfaqs.asp.

Monseur, C., & Adams, R. (2009) Plausible values: how to deal with their limitations. *Journal of Applied Measurement*, 10(3), 320-334.

Von Davier, M., Gonzales, E., & Mislevy, R. J. (2009) What are plausible values and why are they useful? *IERI Monograph Series*, Vol. 2. Retrieved August 1, 2014, from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf.

Wu, M. (2004) Plausible values. *Rasch Measurement Transactions*, 18, 976-978.

Wu, M. (2005) The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, 31(2-3), 114-128.