

Attitudes, Incentives, and Test Performance

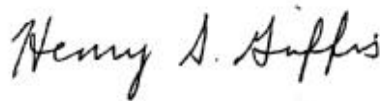
William H. Sims • Catherine M. Hiatt



4825 Mark Center Drive • Alexandria, Virginia 22311-1850

Approved for distribution:

December 2003

A handwritten signature in black ink that reads "Henry S. Griffis". The signature is written in a cursive style with a large initial 'H' and 'G'.

Henry S. Griffis, Director
Workforce, Education and Training Team
Resource Analysis Division

CNA's annotated briefings are either condensed presentations of the results of formal CNA studies that have been further documented elsewhere or stand-alone presentations of research reviewed and endorsed by CNA. These briefings represent the best opinion of CNA at the time of issue. They do not necessarily represent the opinion of the Department of the Navy.

Approved for Public Release; Distribution Unlimited. Specific authority: N00014-00-D-0700.
For copies of this document call: CNA Document Control and Distribution Section (703)824-2123.

Copyright © 2003 The CNA Corporation



Attitudes, Incentives, and Test Performance

31 December 2003

William H. Sims and Catherine M. Hiatt

Summary



- Payment for testing in surveys, such as PAY97, is essential
 - About 50% of respondents say they would only test for pay. They tend to have higher test scores.
- Respondents tried equally hard throughout 9 months of testing
 - About 83% said they tried to do their best. They tend to have higher test scores.
 - Trying their best is not correlated to payment.
- Fewer members of some groups claim to have tried to do their best:
 - Minorities, STP males, ETP 8th graders

Our findings draw on results from the Profile of American Youth (PAY97) and are summarized on this slide:

- Payment for testing in surveys is essential.
- Respondents tried equally hard throughout the 9 months of testing.
- Fewer members of some groups claim to have tried to do their best.

Objective



- To examine the effect of attitudes and incentives on test performance
 - Does pay increase participation?
 - Does pay increase test scores?
 - Do some demographic groups try harder on tests?
 - Do respondents try equally hard throughout a lengthy data collection period?

The objective of this analysis is to examine the effect of attitudes and incentives on test performance. These issues were important considerations in planning the data collection for the National Longitudinal Survey of Youth (NLSY97) and the subset known as PAY97. They are also likely to be important in planning similar data collections in the future.

We will examine the following issues:

- Does pay increase participation?
- Does pay increase test scores?
- Do some demographic groups try harder on tests?
- Do respondents try equally hard throughout a lengthy period of data collection?

Approach



- Use the PAY97 data set for test scores and demographics
 - ETP97 (wt6eout)
 - STP97 (wt6s)
- Use the PAY97 online questionnaire for information on attitudes and incentives

As part of PAY97, participants were asked to take a computerized version of the Armed Services Vocational Aptitude Battery (ASVAB). ASVAB is the enlistment test used by all services to estimate the potential of enlistees for training. This test takes about 1.5 hours to complete. All persons agreeing to take the test were given \$75 for their time.

We will use the PAY97 data set for test scores and demographics. The PAY97 data set is described in the *Profile of American Youth 1997 (PAY97) Technical Sampling Report*, by Whitney Moore, Steven Pedlow, and Kirk Wolter (NORC, August 1999).

These data include the respondents age 18-23 (who are designated as ETP97) and the respondents entering grades 10, 11, and 12 in the fall of 1997 (who are designated as STP97). All data are subsets of the NLSY97. The data were weighted to be nationally representative.

As part of the test administration in PAY97, an online questionnaire was administered to all participants. We will use this questionnaire for information on attitudes and incentives.

PAY97 online questionnaire



- Q29
 - “If you were not offered any money, would you still have taken the test?”
- Q31
 - “I tried to do my best on the test.”
 - Strongly disagree
 - Disagree
 - Neither agree nor disagree
 - Agree
 - Strongly agree

} Disagree

} Agree

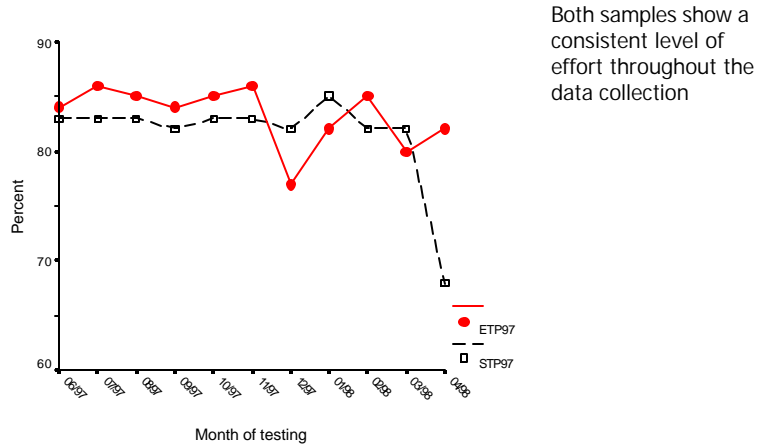
We will use two questions from the online questionnaire:

- Q29: “If you were not offered any money, would you still have taken the test?”
- Q31: Do you agree or disagree with the statement, “I tried to do my best on the test”?

We aggregated the five response options for Q31 into a reduced set of three options for analysis: disagree, neither agree nor disagree, and agree.

Clearly, opinions expressed in the questionnaire are subject to uncertainty. Nonetheless, they represent an important window into the attitudes of test takers that is difficult to get otherwise and may provide useful insights for future data collection efforts.

Percentage saying that “I tried to do my best on the test” by month tested

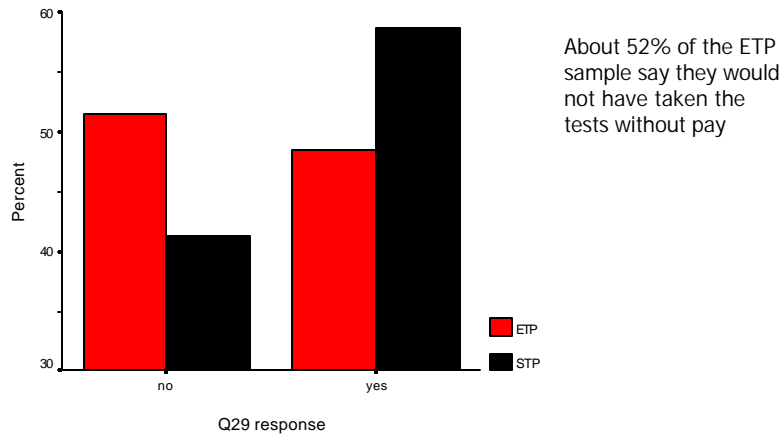


This slide shows the percentage of respondents saying that “I tried to do my best on the test” by month tested.

We see that both the ETP and STP samples show a consistent level of effort throughout the 9-month data collection. There were only a few cases tested during April 1998, so the apparent low level of effort during this month is likely a statistical aberration.

It would have been reasonable to expect that persons brought in for testing toward the end of the data collection were more difficult to persuade and might make less of an effort on the test. This expectation does not seem to be borne out by the data.

Q29: If you were not offered any money, would you still have taken the tests?



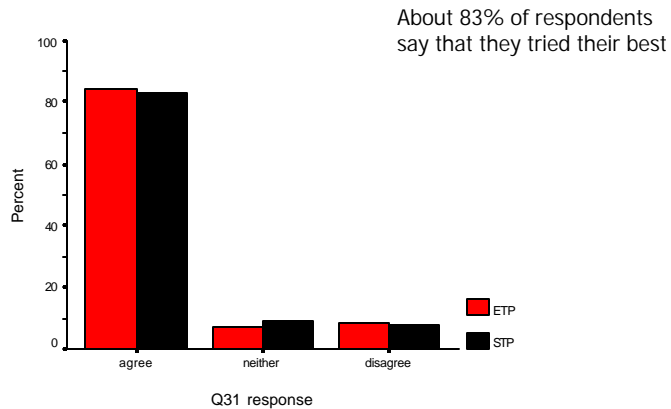
This slide shows responses to the question, “If you were not offered any money, would you still have taken the tests?”

About 52 percent of the ETP sample and 42 percent of the STP sample indicate that they would not have participated in the testing without pay. Apparently, the younger STP sample was somewhat more altruistic about the intrinsic importance of the project.

The percentage saying that they would not have tested without pay is rather large. Even allowing for some exaggeration by respondents, the data indicate that participation is influenced by pay. This should not really be surprising. After all, the respondents were being asked to take a test that would require about 1.5 hours plus administrative time and time to travel to and return from the test site.

It appears that participation in testing is positively influenced by pay and that the lack of pay would have a large negative effect on participation.

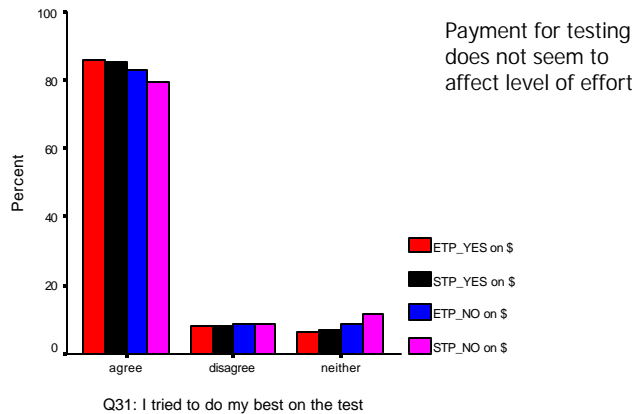
Q31: Do you agree with the statement that “I tried to do my best on the test”?



This slide shows responses to the question: Do you agree with the statement that “I tried to do my best on the test”?

The data indicate that about 83 percent of both ETP and STP respondents say that they tried to do their best on the test.

Does pay affect the level of effort?



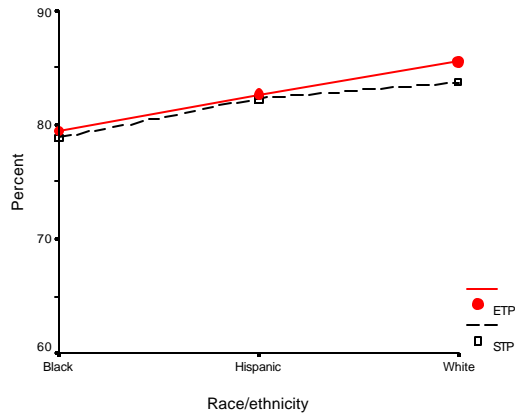
This slide addresses the issue of whether pay affects level of effort.

The four groups of respondents were:

1. ETP respondents who say they would have tested without pay
2. ETP respondents who say they would not have tested without pay
3. STP respondents who say they would have tested without pay
4. STP respondents who say they would not have tested without pay.

The data show that all groups are about equally likely to say that they tried to do their best on the test. Apparently, pay brings them to the test site but does not affect their level of effort.

Percentage saying “I tried to do my best on the test” by race/ethnicity

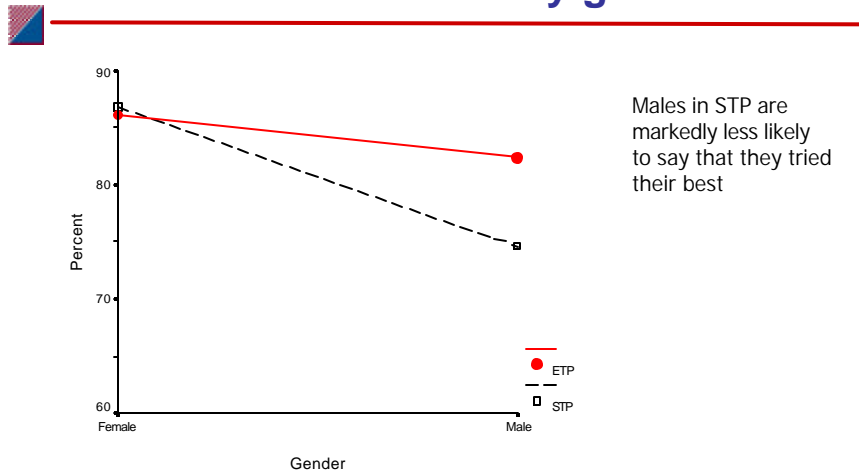


Minorities are slightly less likely to say that they tried their best

Next we examine whether some demographic groups are more or less likely to say that they tried hard on the test.

This slide shows the percentage saying that they tried their best by race/ethnicity. We see that minorities are slightly less likely to say that they tried their best.

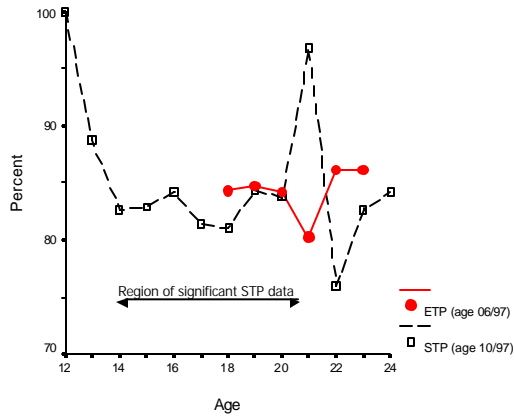
Percentage saying “I tried to do my best on the test” by gender



This slide shows the percentage saying that they tried their best by gender.

Males are markedly less likely to say that they tried to do their best. The gender disparity is greatest for the younger group (STP), where 87 percent of females but only 75 percent of males said that they tried to do their best.

Percentage saying “I tried to do my best on the test” by age



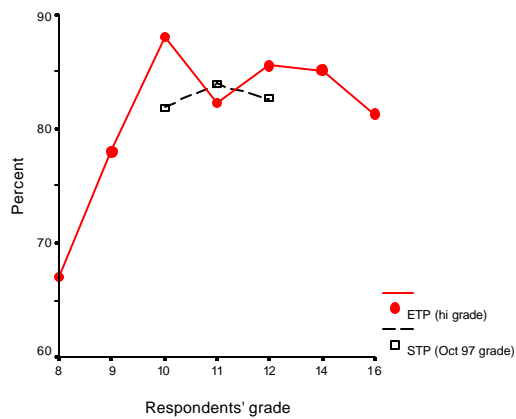
Trying is not a function of age

This slide shows the percentage saying that they tried to do their best by age.

Trying to do one's best does not seem to be a function of age.

The ETP data set (solid circles) seems to be constant with age. The STP data set (open circles) shows more fluctuations, but they are confined to ages for which data are limited. If one focuses only on STP data for ages 14-20, which would include almost all persons entering the 10th, 11th, or 12th grades, the relationship is seen to be constant.

Percentage saying “I tried to do my best on the test” by grade

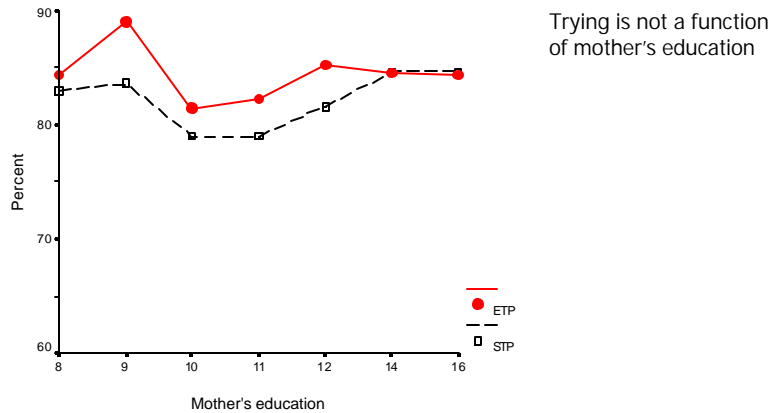


Respondents at 8th grade and below are less likely to say that they tried their best

This slide shows the percentage saying that they tried to do their best by grade.

ETP respondents at the 8th grade level and below are less likely to say that they tried to do their best.

Percentage saying “I tried to do my best on the test” by mother’s education



This slide shows the percentage saying that they tried to do their best by mother’s educational level.

Trying does not seem to be a function of mother’s educational level.

Regression analysis



- AFQT = A
 - + B(Black)
 - + C(Hispanic)
 - + D(Age)
 - + E(Female)
 - + F(Respondent's education)
 - + G(Mother's education)
 - + H(Tried to do my best)
 - + I (Only tested for money)

In the next section, we use regression analysis to control for various demographics and estimate the effect of attitude (tried to do my best) and incentives (only tested for money) on test score.

Weights



- Case weights
 - ETP: wt6eout
 - STP: wt6s
- Case weights scaled by estimated design effect to approximate a simple random sample (SRS)
 - Allows the interpretation of standard regression statistics

The case weights were scaled by an estimated design effect to approximate a simple random sample. This procedure is necessary to allow us to interpret the standard regression statistics. We describe the procedure in detail in the appendix.

Regression for AFQT: ETP

Variable	Coefficient	T-statistic	Significance	Cum. r^2	Delta r^2
Constant	-56.0	-6.7	.000	.140	.140
Black	-18.4	-10.8	.000		
Hispanic	-9.1	-4.7	.000		
Age	-0.6	-1.6	.108		
Female	-2.7	-2.3	.024		
Resp. edu.	6.6	17.4	.000	.377	.237
Mom's edu.	2.8	9.3	.000	.419	.042
"Tried my best"	4.4	2.7	.006	.422	.003
Tested for \$	5.3	4.4	.000	.430	.008

This slide summarizes the regression results for the ETP sample. The variables are shown in the first column. The next three columns show the regression coefficient, T-statistic, and significance level at the *final step* in the regression when all variables have been entered. The next two columns show the cumulative r^2 and change in r^2 as each variable or group of variables is entered in the regression.

The results show that, after controlling for the known demographic influences on AFQT, those who "tried their best" did about 4.4 AFQT points better than those who did not say that they tried their best. The results also show that those who only tested for pay did about 5.3 AFQT points better than those who would have tested without pay. Clearly, respondents' stated attitudes and incentives make a unique difference in the resulting test scores.

Regression for AFQT: STP

Variable	Coefficient	T-statistic	Significance	Cum. r ²	Delta r ²
Constant	-74.6	-6.4	.000	.021	.021
Black	-21.9	-3.5	.001		
Hispanic	-9.2	-1.3	.181		
Age	-2.8	-3.3	.001		
Female	-1.6	-1.1	.267		
Resp. edu.	9.4	7.9	.000	.083	.062
Mom's edu.	4.2	12.9	.000	.201	.118
"Tried my best"	9.4	5.1	.000	.217	.016
Tested for \$	2.2	1.5	.130	.218	.001

This slide summarizes the regression results for the STP sample.

The results are very similar to the ETP sample in the previous slide. STP respondents who say that they tried their best did about 9.4 AFQT points better. STP respondents who said that they only tested for pay did about 2.2 AFQT points better than others; however, this variable was not statistically significant at the standard .05 level.

Summary



- **Payment for testing in surveys is essential**
 - About 50% of respondents say they would only test for pay. They tend to have higher test scores.
- **Respondents tried equally hard throughout the 9 months of testing**
 - About 83% say they tried to do their best. They tend to have higher test scores.
 - Trying their best is not correlated to payment.
- **Fewer members of some groups claim to have tried to do their best:**
 - Minorities, STP males, ETP 8th graders

Based on the foregoing results, we conclude that:

- Payment for testing in such surveys as PAY97 is essential. About 50 percent of respondents say that they would only test for pay. These respondents tend to have higher test scores.
- Respondents say that they tried equally hard throughout the 9 months of testing. About 83 percent say that they tried to do their best on the test, and they tend to have higher test scores. Trying to do their best was not correlated to only being willing to test for pay.
- Fewer members of some demographic groups say that they tried to do their best on the test. Minorities, STP males, and ETP 8th graders are less likely to say that they tried to do their best.

Questions?



Appendix: Design effect



In this appendix, we discuss the design effect.

What is the design effect?



- It is a factor that expresses the inefficiency of a sample relative to a simple random sample:
 - Clustering reduces sampling efficiency
 - Oversampling reduces sampling efficiency
 - Stratification increases sampling efficiency
- Why do we need to know it?
 - To estimate statistical errors
 - To interpret regression statistics

The design effect is a factor that expresses the inefficiency of a sample relative to a simple random sample. A sample with a design effect of 1.0 is equivalent to a simple random sample. A sample with a design effect of 2.0 requires twice as many cases as a simple random sample to be statistically equivalent to a simple random sample.

Clustering and oversampling both reduce sampling efficiency. Stratification, however, increases sampling efficiency. All three procedures were used in PAY80 and PAY97 and are routinely used in other large sampling efforts.

It is essential that we have an estimate for the design effect. For example, the PAY80 data set is based on about 12,000 cases and is weighted by case weights to approximate the total youth population of about 30 million. Neither the raw number of cases nor the weighted number of cases is appropriate for use in statistical tests because neither represents a simple random sample (which is assumed by most common statistical packages). For this reason, we must use the design effect to estimate new scaled case weights that will approximate a simple random sample.

How will we estimate the design effect for PAY97?



- NORC has not yet computed a design effect for PAY97
- We will estimate the design effect for PAY97 by generalizing the design effect computed by NORC for PAY80

NORC has not yet computed a design effect for the PAY97 data set.

We will estimate the design effect for PAY97 by generalizing the design effect computed by NORC for PAY80. The procedure is described in a CNA publication,¹ but we reproduce it here for the convenience of the reader. We believe that the generalization is reasonable because both PAY80 and PAY97 had similar clustered, stratified sampling designs.

1. William H. Sims and Catherine M. Hiatt, *Follow-on Analysis of PAY97 Test Scores*, July 2001 (CNA Annotated Briefing D0003839.A2).

Design effect for mean AFQT in PAY80^a

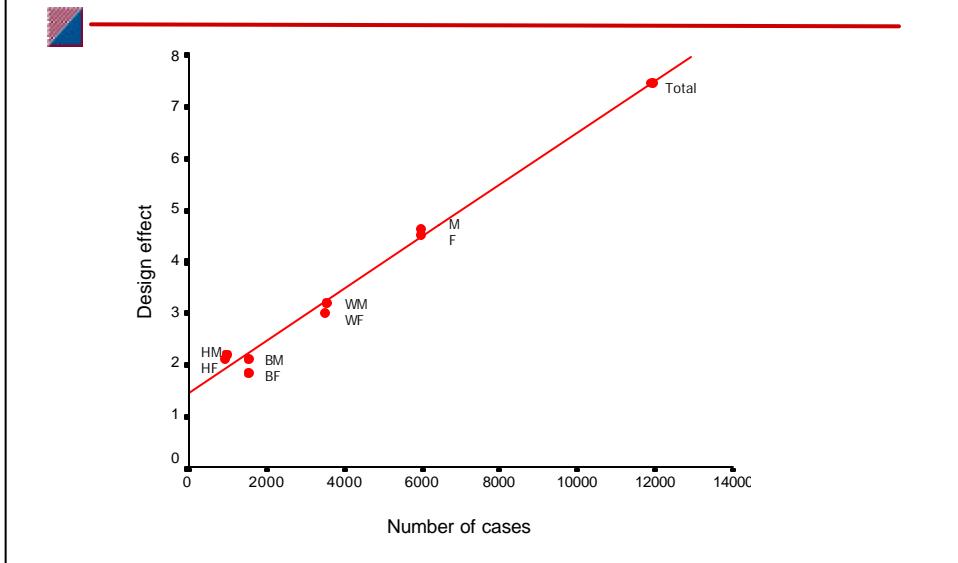
Gender	Race/ethnicity	Number of cases	Design effect
Male	White	3,544	3.2164
	Black	1,517	1.8253
	Hispanic	908	2.1018
	Subtotal	5,969	4.6307
Female	White	3,499	2.9946
	Black	1,511	2.1147
	Hispanic	935	2.2091
	Subtotal	5,945	4.5057
Total		11,914	7.4373

a. *Profile of American Youth, User's Guide and Codebook*, NORC, March 1982

This slide shows the design effects calculated by NORC² for various subpopulations in PAY80. These are the data that we will generalize for use in PAY97.

² *Profile of American Youth (PAY80), User's Guide and Codebook*, NORC, March 1982.

Design effect and sample size: PAY80



This slide shows a chart of the design effect computed by NORC for PAY80 for various subpopulations versus the number of cases in each subpopulation. The data were taken from the previous slide. A simple regression line is seen to fit the data very well.

Scaling case weights for PAY80



- Design effect
= $1.441 + (.0005056) * (\text{sample size})^1$
- Effective sample size
= $\text{sample size} / \text{design effect}$
- Scaled case weight
= $(\text{case weight} / \text{sum of case weights}) * (\text{effective sample size})$

1. Relationship developed for the PAY80 data set. See CNA CAB D0003839.A2, July 2001.

This simple regression equation fits the NORC PAY80 design effects very well. The equation is:

$$\text{Design effect} = 1.441 + .0005056 * (\text{sample size}).$$

We then use this equation to compute the design effect for our various subsamples and apply the result to estimate the size of an effective simple random sample as shown:

$$\text{Effective sample size} = \text{sample size} / \text{design effect}.$$

We then scale the case weights of the sample or subsample as:

$$\text{Scaled case weight} = (\text{case weight} / \text{sum of case weights}) * (\text{effective sample size}).$$

Estimation of design effect in PAY97



Sample	Case weighted sample ¹	Cases	Design effect ²	Equivalent simple random sample (SRS) ³
ETP	17,793,945	5,029	3.9837	1,262
STP	10,132,001	4,077	3.4872	1,169

1. Reduced sample, includes only cases with all regression variables present.
2. Estimated using equation developed for PAY80 that had a similar clustered stratified sample.
Design effect = $1.441 + .0005056$ (cases)
3. Estimated size of equivalent simple random sample = Cases/design effect

In this slide, we estimate the design effect for PAY97 by using the generalized equation developed from PAY80.

The equation is:

$$\text{Design effect} = 1.441 + .0005056 (\text{cases}).$$

Using this equation, we estimate that the design effects for the ETP97 and STP97 samples are 3.9837 and 3.4872, respectively.

