APPENDIX M

Power Analysis:
NTEWS Methodological Experiments
and Seeded Sample

1. Introduction and Overview

This appendix describes the power analysis for the methodological experiments that will be incorporated into the nationally representative National Training, Education, and Workforce Survey (NTEWS) production sample, and the separate power analysis for the seeded sample. While this appendix provides a brief recap of the experimental design, a more detailed description of the methodological experiments and seeded sample is provided in Appendix C.

This planned administration of the NTEWS will be its first cycle. Although the design of the NTEWS is heavily informed by two other federal surveys—the ongoing National Survey of College Graduates (NSCG) and the discontinued Adult Training and Education Survey (ATES) module of the National Household Education Survey (NHES)—it differs in keyways from both collections. For this reason, the administration designs used for the NSCG and ATES cannot be assumed to be optimal for the NTEWS.

Accordingly, the first cycle of NTEWS administered to a nationally representative sample will incorporate two methodological experiments to provide baseline data needed to identify optimal contact strategies and incentive structures for future cycles of the NTEWS.

For the methodological experiments incorporated into the national sample, NCSES expects to conduct a number of post-collection analyses to evaluate the impact of contact strategies and incentive levels on various outcomes. Findings from these analyses are contingent on the size of the NTEWS sample; a sample that is too small will yield insufficient statistical power to detect differences across treatment groups. To evaluate the sufficiency of the initial NTEWS sample for detecting differences among treatment groups, a power analysis was conducted. The overarching goal of the power analyses for the methodological experiments was to confirm that, given the planned national sample size of 43,200, the planned comparisons could be conducted with acceptable minimum detectable differences between treatment groups.

Additionally, and separate from the nationally representative sample, the NTEWS will be administered to a convenience sample of 1,000 known postsecondary certificate holders. This "seeded" sample is included to assess measurement error in the postsecondary certificate attainment item, which has been substantially revised from the prior version included on the ATES. A separate analysis was conducted to determine the requisite sample size for evaluating measurement error in this certificate item.

1.1. Overview of planned methodological experiments and seeded sample

The planned national sample size for the initial cycle of the NTEWS (excluding the seeded sample) is 43,200. This sample size was chosen to achieve precision targets for the estimated prevalence of key credentials (certificates and licenses, postsecondary certificates, and work experience programs) within specified subpopulations. The power analysis for the methodological experiments was then conducted conditional on this expected sample size. That is, the purpose of the power analysis was to determine what differences in response rates and estimates could be detected between experimental treatments given the selected sample size, rather than to inform the choice of the sample size (which was determined independently of the experiments). Refer to Appendix C for a more detailed discussion of the sample design for the NTEWS.

Two randomized methodological experiments will be incorporated into the national sample:

1. Contact-strategies experiment
    - NSCG-style alternating modes (*NSCG*): 9,720 sample members
    - NHES-style sequential modes (*NHES*): 9,720 sample members
    - Choice (no contingent incentive) (*CHOICE*): 4,860 sample members
    - Choice-plus ($20 contingent incentive) (*CHOICEPLUS*): 4,860 sample members
    - Paper-focused (*PAP*): 9,720 sample members
    - CATI-focused (*CAT*): 4,320 sample members
2. Noncontingent-incentive experiment (incentives in week 1 mailing, except for late-stage incentive)
    - No incentive: 10,800 sample members, further divided into:
        - No incentive with week 23 mailing (*INC0*): 5,400 sample members
        - $30 late-stage debit card with week 23 mailing (*INC0L*): 5,400 sample members
    - $10 debit card (*INC10*): 10,800 sample members
    - $20 debit card (*INC20*): 10,800 sample members
    - $30 debit card (*INC30*): 10,800 sample members

Note that the above sample sizes may vary slightly due to rounding. The treatment groups for both experiments will be assigned using a fully crossed factorial design, leading to a total of 30 treatment groups. The national sample is designed to permit statistical evaluations of main effects for each experiment, but not of interactions between treatments (except on an exploratory basis).

Separate from the 43,200-person production sample, the seeded sample will consist of 1,000 known postsecondary certificate holders selected from a convenience frame assembled from several certificate providers (e.g., community colleges). The seeded sample will not be nationally representative, nor will it implement the methodological experiments (though seeded sample members will receive a $10 debit card with the week 1 mailing to encourage response). The primary statistical goal of the seeded sample is to estimate the percentage of known postsecondary certificate holders who report that they do *not* hold a certificate (the certificate underreporting rate).

1.2. Structure of power analysis

The power calculations assumed a power requirement of 0.8 and an alpha requirement of 0.05 (equivalently, a 95 percent confidence level for a two-sided statistical test). Therefore, the minimum detectable difference for a given comparison is the smallest *true* difference for which, given the expected sample size, there is at least an 80 percent probability that a two-sided statistical test would show a statistically significant difference at the 0.05 level. NCSES expects that many comparisons (especially comparisons within subgroups) will be conducted on an exploratory basis and that, in certain circumstances, decisions could be based on observed differences that do not meet standard statistical significance levels. However, for the purpose of calculating the minimum detectable difference, the standard power of 0.8 and alpha of 0.05 were assumed. Results are also shown with an alpha of 0.1, which implies smaller minimum detectable differences.

Power analyses were conducted for several types of comparisons, reflecting several types of planned post-collection analyses. First, minimum detectable differences were calculated for comparisons of both *overall response rates* and of *response rates within subgroups*. The overall response rate is a critical outcome of any data collection, as it determines the cost of obtaining a sufficient number of responses and is also a potential driver of nonresponse bias. Accordingly, NCSES plans to compare overall response rates between treatment groups to identify the single contact strategy and single incentive structure that obtain responses from the largest percentage of the NTEWS sample. At the same time, it is possible that the optimal contact strategy and/or incentive structure will vary between identifiable subgroups within the NTEWS sample. For example, incentives may be necessary for sample members with relatively low response propensities, but not for those with relatively high response propensities. Comparisons of response rates within subgroups will help to determine whether this is the case and, if so, enable NCSES to develop adaptive design strategies that could further improve response rates and/or mitigate nonresponse bias in future NTEWS cycles.

Minimum detectable differences were also calculated for comparisons of *substantive NTEWS estimates* obtained from respondents in different treatment groups. Examples of substantive estimates would include the percentage of respondents reporting a specific type of credential and the distribution of a demographic variable among respondents. NCSES plans to compare substantive estimates to assess the extent to which different contact and incentive strategies affect the composition of NTEWS respondents. This will allow NCSES determine whether a given contact strategy or incentive protocol would affect officially reported statistics from the NTEWS, and to evaluate the effect of these design features on nonresponse bias. For the contact-strategies experiment, comparisons of substantive estimates will also allow the assessment of mode effects to determine whether measurement error in key NTEWS items (e.g., whether the respondent reports a certification) varies among response modes (web, paper, or computer-assisted telephone interview [CATI]).

For the seeded sample, the structure of the power analysis differed. The power analysis for the seeded sample was conducted to determine a sample size that would allow the estimation of the certificate underreporting rate at a reasonable level of precision—specifically, a margin of error (MOE) of 4 percentage points or below.

2. Analysis for methodological experiments in national sample

2.1. Parameters to be compared

As noted above, power analyses were conducted for comparisons of response rates between any two experimental treatment groups, both overall and within key subgroups. For the purpose of these calculations, the response rate was defined as the percentage of the sample that completed the NTEWS questionnaire.[1]

Power analyses were also conducted for two types of comparisons of estimates. For both the contact-strategies and noncontingent-incentive experiments, detectable differences were calculated for *comparisons of overall estimates*—those obtained from *all* respondents from a given treatment group. For the NHES-style, paper-focused, and CATI-focused treatments within the contact-strategies experiment, detectable differences were also calculated for *mode effects comparisons*. To isolate the impact of the response mode on responses to key NTEWS items, these comparisons use only those respondents from the first 11 weeks of data collection. As shown in table C.2 of the experimental design appendix (Appendix C), the NHES-style, paper-focused, and CATI-focused treatments offer only web, paper, or CATI response modes (respectively) for the first 11 weeks of data collection.

2.2. Assumptions

Because of the complex sample design, the NTEWS sample will be subject to weighting variability that will likely inflate the variances of most weighted estimates relative to what would be obtained from a simple random sample (SRS) of the same size. This variance inflation factor attributable to departures from simple random sampling is referred to as the *design effect*. Minimum detectable differences for the NTEWS experiments were calculated in two ways: (1) ignoring the design effect (i.e., using SRS formulas) and (2) incorporating the design effect. For the latter, a design effect of 2 was assumed based on the weighting variation observed in a test sample drawn by the Census Bureau using the planned NTEWS sample design. NTEWS sample members will be randomly assigned to treatment groups using a systematic design; therefore, despite the complex sample design, unweighted comparisons between treatment groups (which will not be subject to the design effect) will be internally valid. That is, it will be possible to conclude (with 95 percent confidence) that a statistically significant unweighted effect is a true effect *within the 2022 NTEWS sample*. However, weights and the design effect will need to be used to obtain externally valid estimates of the treatment effect over the NTEWS target population (U.S. adults ages 16 through 75 and not enrolled in primary or secondary school). Therefore, detectable differences that ignore the weights and the design effect are still detectable differences at the *sample* level but may be biased estimates of differences at the population level, with incorrect uncertainty estimates associated with them. Differences that incorporate weights and the design effect are detectable differences at the *population* level. Sample-level differences may be sufficient for making decisions about future NTEWS cycles if the sample design remains approximately unchanged; whereas population-level differences would be more appropriate if there are meaningful changes to the sample design.

An assumed *response rate* was needed to determine sample sizes for comparisons of estimates (which can be conducted only among respondents). For this analysis, a response rate of 62.5

[1] For the purpose of the power analysis, it was assumed that sampled persons who indicate on the NTEWS that they are still enrolled in grades K-12—and thus are outside the target population for the NTEWS—would be counted as respondents because the knowledge that a sampled person is outside the target population is, operationally, useful information. For this reason, the *operational* response rates used for these experimental comparisons may differ from the officially reported NTEWS response rates, which would treat such persons as ineligible.

percent was assumed for most treatment groups. This was the same response rate used in the NTEWS burden calculations. The exceptions were for the treatment groups receiving no incentive at week 1, for which a 57.5 percent response rate (5 percentage points lower) was assumed; and the treatment group receiving a $30 incentive at week 1, for which a 67.5 percent response rate (5 percentage points higher) was assumed.

For comparisons within subgroups, the minimum detectable difference depends on the *size of the subgroup* as a proportion of the total sample. For the purpose of these calculations, the subgroup of interest was assumed to constitute 25 percent of the total sample. In practice, a number of subgroups are likely to be of interest in the analysis of the experiments, some of which will constitute less or more than 25 percent of the sample. Detectable differences for subgroups constituting 25 percent of the sample were calculated to provide a general idea of how well the selected sample sizes would support evaluations of the effects of the treatments within subgroups.

In general, when conducting power calculations for comparisons of proportions, it is necessary to choose a *baseline* proportion. For a given sample size, the minimum detectable difference is highest for a 50 percent estimate and decreases the closer an estimate is to either 0 or 100 percent. For comparisons of NTEWS estimates (e.g., the percentage reporting a credential), a baseline proportion of 25 percent was assumed. This is the approximate population prevalence of certifications and licenses estimated by the NCES 2016 ATES, a predecessor collection to the NTEWS. The population prevalence for each of the other key characteristics measured by the NTEWS—attainment of postsecondary certificates and completion of a work experience program—is expected to be the same or lower than that of certifications and licenses, so detectable effects for comparisons of these other key substantive estimates may be smaller than shown here.

Finally, some comparisons are expected to focus only on respondents as of a certain point in data collection. Evaluations of mode effects will be restricted to respondents from the *first 11 weeks of data collection* to focus on the period during which only a single response mode was offered (in the NHES-style, paper-focused, and CATI-focused treatments). For these mode effects comparisons, it was assumed, based on the experience of the NSCG, that 60 percent of respondents will complete the NTEWS in the first 11 weeks of collection. Evaluations of the effect of the week 1 incentive will be restricted to respondents from the *first 22 weeks of data collection* to avoid confounding by the late-stage incentive that will be sent to half of the $0-incentive group in week 23. Similarly, evaluations of the effect of this late-stage incentive will focus on sample members that have *not* responded by week 23, to determine the incremental effect of the late-stage incentive. For these comparisons, it was assumed, again based on the experience of the NSCG, that 90 percent of respondents will complete the NTEWS in the first 22 weeks of the collection.

2.3. Results

Table L.1 on page 7 shows detectable differences, in percentage point terms, for main effect comparisons between any two contact-strategy treatments, from power analyses using the assumptions described above and a significance level of 0.05. Table L.2 on page 8 shows the same for main effect comparisons between any two noncontingent-incentive treatments, again with a significance level of 0.05.

Tables L.3 and L.4 (pages 9 and 10, respectively) show detectable differences for the contact-strategies and noncontingent-incentive experiments (respectively) assuming a significance level of 0.1.

3. Analysis for seeded sample

Because the 1,000-person seeded sample is not included in the methodological experiments, no experimental comparisons will be conducted within the seeded sample. Rather, the primary statistical goal of the seeded sample is to estimate the percentage of known postsecondary certificate holders who report that they do *not* hold a certificate (the certificate underreporting rate). For this reason, the size of the seeded sample was chosen to allow estimation of the certificate underreporting rate with an MOE of 4 percentage points or below.

The precision with which this rate can be estimated will depend on the response rate to the seeded sample. This response rate may differ from the 62.5 percent response rate assumed for the national sample, because the seeded sample will be drawn from a different frame (lists of known certificate holders provided by a convenience sample of postsecondary institutions) and will not be nationally representative.[2] For the purpose of this calculation, a response rate of 60 percent was assumed for the seeded sample. This implies that the 1,000-person seeded sample would yield about 600 respondents. Because weights will not be calculated for the seeded sample, there will be no design effect.

In the 2016 ATES seeded sample, an underreporting rate of close to 50 percent was observed. Expecting 600 respondents, and having no design effect, NCSES and NCES estimate that the NTEWS seeded sample would estimate a 50 percent underreporting rate with an MOE of about 4 percentage points. If in fact the underreporting rate with the revised certificate item is either greater than or less than 50 percent, the MOE will be lower.

[2] A convenience sample was used because an exhaustive list of certificate holders in the U.S. does not currently exist. Assembling such a list for the purpose of the NTEWS would have been cost-prohibitive.

Table L.1. Detectable differences for comparisons between any two contact-strategy treatments, 0.05 significance level

Parameter	Control group	Treatment group	Detectable difference (percentage points) Unweighted	Weighted
Response rate	NSCG, NHES, or PAP	NSCG, NHES, or PAP	2.0	2.8
	NSCG, NHES, or PAP	CHOICE or CHOICEPLUS	2.4	3.4
	CHOICE	CHOICEPLUS	2.8	3.9
	NSCG, NHES, or PAP	CAT	2.5	3.5
	CHOICE or CHOICEPLUS	CAT	2.9	4.0
Response rate within 25% subgroup	NSCG, NHES, or PAP	NSCG, NHES, or PAP	3.9	5.6
	NSCG, NHES, or PAP	CHOICE or CHOICEPLUS	4.8	6.8
	CHOICE	CHOICEPLUS	5.6	7.9
	NSCG, NHES, or PAP	CAT	5.0	7.1
	CHOICE or CHOICEPLUS	CAT	5.7	8.2
Estimate (all respondents)	NSCG, NHES, or PAP	NSCG, NHES, or PAP	2.2	3.2
	NSCG, NHES, or PAP	CHOICE or CHOICEPLUS	2.7	3.9
	CHOICE	CHOICEPLUS	3.2	4.5
	NSCG, NHES, or PAP	CAT	2.8	4.1
	CHOICE or CHOICEPLUS	CAT	3.3	4.7
Estimate (mode effects)[1]	NHES	PAP	3.0	4.3
	NHES or PAP	CAT	3.9	5.5

[1] Mode effects comparisons would use data only from the first 11 weeks of data collection from the NHES, paper-focused, and CATI-focused groups, to isolate the impact of mode on responses to key items.

NOTE: Calculations assume a total sample size of 43,200. The unweighted detectable difference is calculated ignoring the design effect and is therefore the detectable difference at the sample level. The weighted detectable difference is calculated assuming a design effect of 2 and is therefore the detectable difference at the population level. Detectable differences reflect a power requirement of 0.8 and a required significance level of 0.05.

Table L.2. Detectable differences between any two noncontingent-incentive treatments, 0.05 significance level

			Detectable difference (percentage points)	
Parameter	**Control group**	**Treatment group**	**Unweighted**	**Weighted**
Response rate before week 23[1]	INC0+INC0L, INC10, INC20, or INC30	INC0+INC0L, INC10, INC20, or INC30	1.9	2.7
Incremental response rate after week 23[2]	INC0	INC0L	2.6	3.8
Response rate before week 23 within 25% subgroup[1]	INC0+INC0L, INC10, INC20, or INC30	INC0+INC0L, INC10, INC20, or INC30	3.8	5.4
Incremental response rate after week 23 within 25% subgroup[2]	INC0	INC0L	5.5	8.0
Estimate (respondents prior to week 23)[3]	INC0+INC0L, INC10, INC20, or INC30	INC0+INC0L, INC10, INC20, or INC30	2.2 - 2.3	3.1 - 3.2
Estimate (all respondents)[4]	INC0	INC0L	3.1	4.5

[1] Response rate comparisons by prepaid incentive amount use response rates prior to week 23 to avoid confounding by the late-stage incentive.

[2] Response rate comparisons of the late-stage vs. no late-stage groups focus on the incremental response rate after week 23, because this is the mailing at which the late-stage incentive is sent.

[3] Comparisons of key estimates by prepaid incentive amount are restricted to respondents prior to week 23 to avoid confounding by the late-stage incentive. Detectable differences vary for this comparison because the response rate was assumed to be lower than average in INC0+INC0L and higher than average in INC30; therefore, the detectable difference depends in the specific groups between which estimates are compared.

[4] Comparisons of key estimates by the late-stage incentive amount use all respondents to evaluate the effect of the late-stage incentive on the final respondent composition.

NOTE: "INC0+INC0L" refers to INC0 and INC0L combined. Calculations assume a total sample size of 43,200. The unweighted detectable difference is calculated ignoring the design effect and is therefore the detectable difference at the sample level. The weighted detectable difference is calculated assuming a design effect of 2 and is therefore the detectable difference at the population level. Detectable differences reflect a power requirement of 0.8 and a required significance level of 0.05.

Table L.3. Detectable differences for comparisons between any two contact-strategy treatments, 0.1 significance level

			Detectable difference (percentage points)	
Parameter	**Control group**	**Treatment group**	**Unweighted**	**Weighted**
Response rate	NSCG, NHES, or PAP	NSCG, NHES, or PAP	1.7	2.5
	NSCG, NHES, or PAP	CHOICE or CHOICEPLUS	2.1	3.0
	CHOICE	CHOICEPLUS	2.5	3.5
	NSCG, NHES, or PAP	CAT	2.2	3.1
	CHOICE or CHOICEPLUS	CAT	2.5	3.6
Response rate within 25% subgroup	NSCG, NHES, or PAP	NSCG, NHES, or PAP	3.5	4.9
	NSCG, NHES, or PAP	CHOICE or CHOICEPLUS	4.3	6.1
	CHOICE	CHOICEPLUS	4.9	7.0
	NSCG, NHES, or PAP	CAT	4.4	6.3
	CHOICE or CHOICEPLUS	CAT	5.1	7.2
Estimate (all respondents)	NSCG, NHES, or PAP	NSCG, NHES, or PAP	2.0	2.8
	NSCG, NHES, or PAP	CHOICE or CHOICEPLUS	2.4	3.4
	CHOICE	CHOICEPLUS	2.8	4.0
	NSCG, NHES, or PAP	CAT	2.5	3.6
	CHOICE or CHOICEPLUS	CAT	2.9	4.1
Estimate (mode effects)[1]	NHES	PAP	2.7	3.8
	NHES or PAP	CAT	3.4	4.9

[1] Mode effects comparisons would use data only from the first 11 weeks of data collection from the NHES, paper-focused, and CATI-focused groups, to isolate the impact of mode on responses to key items.

NOTE: Calculations assume a total sample size of 43,200. The unweighted detectable difference is calculated ignoring the design effect and is therefore the detectable difference at the sample level. The weighted detectable difference is calculated assuming a design effect of 2 and is therefore the detectable difference at the population level. Detectable differences reflect a power requirement of 0.8 and a required significance level of 0.1.

Table L.4. Detectable differences between any two noncontingent-incentive treatments, 0.1 significance level

			Detectable difference (percentage points)	
Parameter	**Control group**	**Treatment group**	**Unweighted**	**Weighted**
Response rate before week 23[1]	INC0+INC0L, INC10, INC20, or INC30	INC0+INC0L, INC10, INC20, or INC30	1.7	2.4
Incremental response rate after week 23[2]	INC0	INC0L	2.3	3.3
Response rate before week 23 within 25% subgroup[1]	INC0+INC0L, INC10, INC20, or INC30	INC0+INC0L, INC10, INC20, or INC30	3.4	4.8
Incremental response rate after week 23 within 25% subgroup[2]	INC0	INC0L	4.8	7.0
Estimate (respondents prior to week 23)[3]	INC0+INC0L, INC10, INC20, or INC30	INC0+INC0L, INC10, INC20, or INC30	1.9 - 2.0	2.8 - 2.9
Estimate (all respondents)[4]	INC0	INC0L	2.8	4.0

[1] Response rate comparisons by prepaid incentive amount use response rates prior to week 23 to avoid confounding by the late-stage incentive.

[2] Response rate comparisons of the late-stage vs. no late-stage groups focus on the incremental response rate after week 23, because this is the mailing at which the late-stage incentive is sent.

[3] Comparisons of key estimates by prepaid incentive amount are restricted to respondents prior to week 23 to avoid confounding by the late-stage incentive. Detectable differences vary for this comparison because the response rate was assumed to be lower than average in INC0+INC0L and higher than average in INC30; therefore, the detectable difference depends in the specific groups between which estimates are compared.

[4] Comparisons of key estimates by the late-stage incentive amount use all respondents to evaluate the effect of the late-stage incentive on the final respondent composition.

NOTE: "INC0+INC0L" refers to INC0 and INC0L combined. Calculations assume a total sample size of 43,200. The unweighted detectable difference is calculated ignoring the design effect and is therefore the detectable difference at the sample level. The weighted detectable difference is calculated assuming a design effect of 2 and is therefore the detectable difference at the population level. Detectable differences reflect a power requirement of 0.8 and a required significance level of 0.1.