An Analysis of Mode Effects in the 2010 Course Experience Questionnaire

DAVID CARROLL

Graduate Careers Australia

Accepted as Best Paper 2011 AAIR Forum

Abstract

Historically, responses to the Course Experience Questionnaire (CEQ) were required to be collected by self-administered paper or online questionnaire to be eligible for official analysis. CEQ responses collected by telephone were excluded from the final analysis file to minimise the potential for bias due to mode effects: systematic variation in responses obtained using different data collection methods. For the 2010 CEQ, however, telephone data collection was permitted to maximise response rates, with responses collected in this manner included in the final analysis file for the first time. In all, nearly a tenth of all valid responses to the 2010 CEQ were collected by telephone, with institutional use of telephone data collection ranging from 18 to 56% of all responses received for that institution. Using regression and matching methods, this article seeks to identify mode effects in the 2010 CEQ data that cannot be attributed to compositional differences between the telephone and self-administered respondent samples. Implications for survey practice are also discussed.

Keywords: mode effects; survey mode; mixed mode; data comparability; data collection

Each year, graduates from all Australian higher education institutions who complete a coursework (non-research) degree are invited to complete the Course Experience Questionnaire (CEQ), which consists of attitudinal statements rated on a five-point Likert response format from *strongly disagree* to *strongly agree*. CEQ data are widely used in Australian higher education for the purposes of course and program evaluation and development, institutional performance measurement and, more recently, allocation of performance-based funding to institutions. Due to the importance placed on these data, the Graduate Careers Australia (GCA) Code of Practice governing the public disclosure of data from the Australian Graduate Survey (AGS), of which the CEQ is a component, mandates a minimum response rate of 50% to allow its public release. This longstanding requirement was originally implemented to enhance the face validity of the survey and maximise the number of cases available for detailed analysis (GCA, 2010a).

As shown in Figure 1, national CEQ response rates over the decade to 2009 hovered in the high 40% range. Indeed, for the 2009 AGS, nearly half of all institutions failed to achieve a 50% response rate for the CEQ component (GCA & ACER, 2010). To combat these low response rates, the sector-wide Survey Reference Group (SRG) that advises on the conduct of the AGS agreed, with caveats, to a request from Universities Australia (UA) that

responses to the 2010 CEQ be collected by telephone interview (as long as the interviewing was undertaken by an independent third party), and that responses collected in this manner be included for official analysis (GCA, 2009). Historically, only those responses collected by self-administered paper or online survey were included for official analysis; CEQ responses collected by telephone interview were excluded from the national data file and did not count toward an institution's CEQ response rate.

Nine higher education institutions that participated in the 2010 CEQ used telephone interviewing in conjunction with paper and/or online surveying. The share of CEQ responses gathered by telephone ranged from 56.0% to 18.4%, with a median of 29.2%. In total, 11,720 responses were collected by telephone, representing 9.2% of all responses in 2010. These nine institutions typically used paper and/or online surveys as their primary means of data collection, employing more costly telephone interviewing as a means of following up graduates who did not respond to the initial invitation, nor subsequent email or postal reminders. As anticipated, telephone data collection led to a relatively high national CEQ response rate of 52.6% in 2010 (GCA, 2011a).
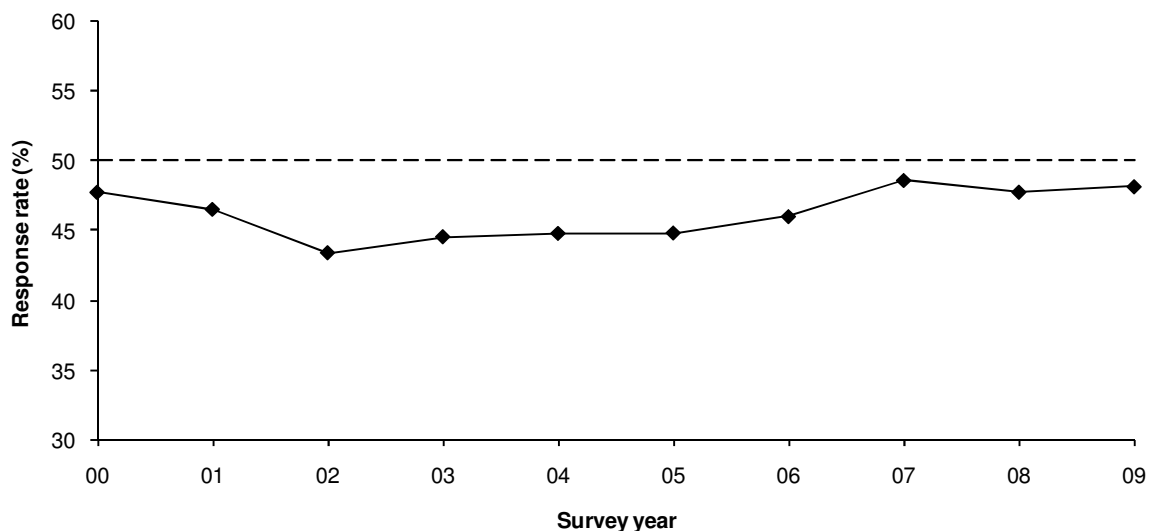

Figure 1

National CEQ response rates, 2000–09 (Adapted from GCA & ACER, 2010, p. 8). The dashed line indicates a response rate of 50%.

In spite of the success of telephone data collection in increasing the CEQ response rate, it is an open question as to whether the responses collected by telephone interview are comparable to those collected by self-completed paper or online survey. A well-documented problem with mixed-mode surveys is caused by mode effects, which refers to systematic variation in responses obtained using different data collection methods (van Nunspeet, Cuppen, & van der Laan, 2011). The purpose of this current article is to investigate whether the 2010 CEQ was subject to significant mode effects; in other words, whether CEQ responses gathered by telephone interview differ significantly to those gathered by self-administered survey after controlling for potential confounding factors.

The rest of this article is organised as follows. Section 1 presents a review of relevant literature and details our specific contribution. Section 2 provides a brief overview of the data and variables used in this study, while Section 3 outlines our empirical methodology. Section

4 presents the results of our mode effects analyses. Finally, conclusions, implications for survey practice, and limitations of this study are presented in Section 5.

1. Background

The proliferation of mixed-mode surveys in recent decades has seen the emergence of a body of literature concerning how the nature of response differs between data collection modes. A common theme in this literature is that different data collection modes often produce different answers to the same questions, with many studies demonstrating that survey responses differ between data collected via an interviewer-administered mode (e.g., telephone interview) and a self-administered mode (e.g., paper survey, online survey). Christian, Dillman and Smyth (2008), for example, found that telephone survey respondents tend to give significantly more positive responses than online survey respondents across various scale questions, including fully labelled and endpoint-only labelled scales. Dillman et al. (2009) found that, while combining different data collection modes was an effective means of improving response rates, individuals who responded via an aural data collection mode (telephone and interactive voice response) were significantly more likely to give positive responses than those who responded via paper or online survey. Kelly, Harper and Landau (2008) observed the opposite effect, with responses collected by an interviewer tending to be more negative than those collected by online survey. Differences in response between interviewer-administered and self-administered surveys have also been demonstrated by Dillman, Sangster, Tarnai and Rockwood (1996), Fowler, Roman and Di (1998), Krysan, Schuman, Scott and Beatty (1994), and Tarnai and Dillman (1992), among others. Although outside the scope of this current study, mode effects have also been observed in surveys with two interviewer-administered modes (e.g., Aquilino & Lo Sciuto, 1990), or with two self-administered modes (e.g., Yang, Falcone, & Milan, 2009).

With specific regard to mode effects in the CEQ, the most notable study is an unpublished report addressed to GCA by Edwards (2008), which examined the responses to the 2007 CEQ to identify whether differences existed between those collected by telephone and by self-administered survey. He concluded that responses collected by telephone were marginally more positive, but attributed this to the difference in composition of the telephone and self-administered respondent samples. He also concluded that the individual CEQ items underlying the scales performed similarly, regardless of the collection method employed. This analysis was limited somewhat by the small number of telephone responses to the 2007 CEQ. A total of 1,806 telephone responses were received in 2007, representing just 1.5% of all CEQ responses in that year. (Recall that telephone responses to the 2007 CEQ were ineligible for official analysis.)

By way of theoretical background, three explanations for why different data collection modes can produce different responses to otherwise identical questions include social desirability, acquiescence and primacy/recency effects (Dillman & Christian, 2005). *Social desirability* refers to the tendency for individuals to offer responses that they feel will be viewed favourably by others. Respondents to interviewer-administered surveys in particular may choose to respond more positively than if they were completing a self-administered survey because they do not want to displease the interviewer (McFarlane & Garland, 1994). *Acquiescence* refers to the tendency for respondents to agree with attitude statements presented to them (Schuman & Scott, 1989). Since respondents to interviewer-administered surveys typically have less time to weigh the issues carefully before responding, they tend to be more prone to acquiescence than respondents to self-administered surveys (Ayidiya & McClendon, 1990). *Recency* is the tendency for respondents to interviewer-administered

surveys to choose from the last offered response categories, while *primacy* is the tendency for respondents to self-administered surveys to choose from the first offered categories (Dillman & Christian, 2005; Krosnick & Alwin, 1987). Since disentangling these effects is practically impossible using only the observational data available to us, the focus of this article is to identify whether there is a significant difference in responses to the 2010 CEQ between those who completed the survey by telephone and those who completed a self-administered survey (either paper or online). Establishing the cause of any observed mode effect is outside the scope of this article, but is an area for further research. We restrict our analysis to bachelor degree graduates to minimise the potential that our results are confounded by extraneous factors. Considering that these graduates comprise nearly two thirds of all responses to the 2010 CEQ, this restriction will have little bearing on the implications of our study.

2. *Data*

This study is based on data from the 2010 CEQ, administered as a component of the 2010 AGS by GCA. All students who qualified for the award of a degree or diploma from an Australian higher education institution in 2009 were invited to complete the survey. Students who completed their studies in the first half of the year were surveyed as at 31 October, while those who completed their studies in the second half were surveyed as at 30 April the following year. The CEQ comprises eleven scales underpinned by 49 Likert-type items, which are evaluated using a five-point response format with categories *strongly disagree*, *disagree*, *neither agree nor disagree*, *agree* and *strongly agree*. All participating institutions are required to administer three 'core' scales (*Good Teaching*, *Generic Skills* and *Overall Satisfaction*) and may then choose to add one or more of the eight optional scales to their questionnaire. Graduates may provide responses for up to two fields of education on the CEQ, with each response conventionally treated as a separate case for the purposes of data analysis. Scale scores are computed as the mean of the constituent item scores after recoding the five categories of the response format to -100, -50, 0, 50 and 100 respectively (GCA, 2011b). The resulting scale scores follow an approximately normal distribution.

Starting with the national CEQ data file, we firstly excluded all responses from institutions other than the nine that undertook data collection by means of telephone interview and self-administered survey. Next, we excluded non-bachelor degree respondents and respondents who did not provide a valid response to all of the variables used in our study. Since exploratory analysis showed that self-administered respondents provided a response concerning their second field of education 1.5 times more often than did telephone respondents, we excluded all responses not related to a graduate's first field of education. These exclusions resulted in a total analysis sample of 20,845 graduates, including 6,226 telephone respondents and 14,619 self-administered respondents. The dependent variable in our study is the six-item *Good Teaching* scale (GTS). Our analysis is limited to one scale in the interest of concision. We specifically selected the GTS because of the vital importance of teaching in the higher education sector, and also because the GTS received the most responses out of any scale for the 2010 CEQ. While the use of Likert scale data in parametric statistical procedures such as multiple linear regression is a somewhat contentious issue, we follow the view of Carifio and Perla (2007), among others, that Likert scales can produce interval-level data. Values of the GTS range from -100 to 100. Table 1 presents summary statistics showing differences between the telephone and self-administered respondent groups, with *t*-statistics greater than (less than) 1.96 (-1.96) indicating a significant difference at the 5% level.

Table 1

Summary Statistics, By Respondent Group

Variable	Name	Telephone		Self-administered		H₀: Equal means
		Mean	*SD*	**Mean**	*SD*	*t*
Good Teaching scale	gts	34.087	32.553	27.663	36.829	**12.53**
Age in years	ageyrs	24.498	5.565	25.185	6.626	**-7.69**
Male	male	0.452	0.498	0.351	0.477	**13.68**
Bachelor degree (honours)	bhons	0.059	0.236	0.097	0.295	**-9.74**
Studied full-time	ftstudy	0.882	0.323	0.891	0.312	-1.86
Studied on campus	oncmode	0.882	0.323	0.873	0.333	1.70
Australian citizen/resident	austres	0.863	0.344	0.875	0.331	**-2.21**
Language other than English	nesb	0.242	0.428	0.217	0.412	**3.84**
Work type: full-time	worka	0.485	0.500	0.489	0.500	-0.54
Work type: part-time	workb	0.305	0.460	0.300	0.458	0.69
Seeking work	seek	0.291	0.454	0.349	0.477	**-8.31**
Further study: full-time	furstuda	0.211	0.408	0.210	0.407	0.14
Further study: part-time	furstudb	0.055	0.228	0.061	0.239	-1.62
Located in Australia	inaust	0.957	0.203	0.929	0.257	**8.30**
Deferred some or all course fees	deferfee	0.574	0.494	0.704	0.457	**-17.70**
Advanced standing towards qualification	advstand	0.310	0.463	0.291	0.454	**2.80**
Double degree	dbldeg	0.109	0.312	0.124	0.330	**-3.13**
Disability identified	disab	0.029	0.167	0.024	0.152	**2.12**
Number of years spent enrolled	enryrs	3.949	1.733	3.958	1.775	-0.33
Field: Natural and physical sciences	majora	0.086	0.281	0.104	0.305	**-3.99**
Field: Information technology	majorb	0.036	0.187	0.037	0.190	-0.43
Field: Engineering and related	majorc	0.052	0.222	0.057	0.231	-1.32
Field: Architecture and building	majord	0.029	0.167	0.026	0.161	0.91
Field: Agriculture, environmental and related	majore	0.011	0.105	0.013	0.114	-1.24
Field: Health	majorf	0.158	0.365	0.180	0.385	**-3.93**
Field: Education	majorg	0.075	0.263	0.065	0.246	**2.54**
Field: Society and culture	majorh	0.173	0.378	0.196	0.397	**-3.99**
Field: Creative arts	majori	0.097	0.297	0.083	0.275	**3.40**
N		14,619		6,226		

Notes. Computations based on data from the 2010 CEQ. *SD* = standard deviation; *t* = *t*-statistic. All variables listed are 0/1 dummies, except for Good Teaching scale, age in years, and number of years spent enrolled. Significant *t*-statistics at the 5% level are in boldface.

3.   *Empirical Methodology*

Inferring a causal link between data collection mode and GTS scores is hampered by the strong likelihood of selection bias. Using experimental terms, respondents were not randomly assigned to 'treatment' (telephone) and 'control' (self-administered) groups; they essentially self-selected into these groups by virtue of whether they responded to the survey in a timely fashion. As first noted by Edwards (2008) and illustrated in our study in Table 1, telephone and self-administered respondents to the CEQ differ across a number of characteristics. Failure to control for this selection bias may result in confounded estimates of the relationship between data collection mode and GTS scores. To address this we used propensity scores to match groups in regard to their likelihood of providing a response by telephone. Propensity adjustment is well-documented as reducing the bias inherent in retrospective studies (Braitman & Rosenbaum, 2002).

First, we calculated the propensity (predicted probability) for each respondent to provide a CEQ response by telephone through multiple logistic regression, using age, sex, degree level, attendance type and mode, residency, language spoken at home, work status, work-seeking status, further-study status, geographic location, fee type, advanced standing, double degree status, disability status, years spent enrolled, broad field of education and institution as covariates. Next, all responses were weighted by these propensity scores so that the two groups had the same overall propensity to be assigned to either collection mode (Kertesz et al., 2009). Propensity weights were computed as $1/(P)$ for telephone respondents and $1/(1-P)$ for self-administered respondents, where P is the propensity score (Hirano & Imbens, 2001). Propensity adjustment resulted in the two respondent groups being well-balanced as to their observed characteristics, as shown in Table 2.

We estimate the effect of telephone data collection on GTS scores using multiple linear regression, weighted as previously described, controlling for age, sex, degree level, attendance type and mode, residency, language spoken at home, broad field of education, work status, work-seeking status, further-study status, and institution. We also produce unweighted estimates as a basis for comparison.

Table 2

Propensity-Adjusted Summary Statistics, By Respondent Group

Variable	Name	Telephone Mean	Telephone SD	Self-administered Mean	Self-administered SD	H_0: Equal means t
Good Teaching scale	gts	34.087	33.016	27.597	36.775	**12.55**
Age in years	ageyrs	25.066	6.345	24.984	6.391	0.85
Male	male	0.380	0.486	0.380	0.485	0.06
Bachelor degree (honours)	bhons	0.090	0.287	0.086	0.280	1.11
Studied full-time	ftstudy	0.887	0.316	0.889	0.315	-0.24
Studied on campus	oncmode	0.874	0.332	0.875	0.330	-0.33
Australian citizen/resident	austres	0.872	0.335	0.871	0.335	0.07
Language other than English	nesb	0.228	0.419	0.225	0.418	0.42
Work type: full-time	worka	0.491	0.500	0.489	0.500	0.29
Work type: part-time	workb	0.298	0.458	0.300	0.458	-0.30
Seeking work	seek	0.330	0.470	0.331	0.471	-0.13
Further study: full-time	furstuda	0.209	0.407	0.210	0.407	-0.15
Further study: part-time	furstudb	0.061	0.240	0.059	0.237	0.47
Located in Australia	inaust	0.935	0.246	0.937	0.243	-0.51
Deferred some or all course fees	deferfee	0.658	0.474	0.662	0.473	-0.57
Advanced standing towards qualification	advstand	0.293	0.455	0.296	0.456	-0.38
Double degree	dbldeg	0.120	0.325	0.120	0.325	-0.04
Disability identified	disab	0.025	0.155	0.025	0.156	-0.06
Number of years spent enrolled	enryrs	3.978	1.710	3.956	1.783	0.83
Field: Natural and physical sciences	majora	0.101	0.301	0.099	0.298	0.53
Field: Information technology	majorb	0.036	0.186	0.037	0.188	-0.28
Field: Engineering and related	majorc	0.053	0.224	0.055	0.228	-0.51
Field: Architecture and building	majord	0.026	0.159	0.027	0.162	-0.46
Field: Agriculture, environmental and related	majore	0.013	0.112	0.013	0.112	0.00
Field: Health	majorf	0.171	0.377	0.174	0.379	-0.49
Field: Education	majorg	0.067	0.251	0.068	0.251	-0.05
Field: Society and culture	majorh	0.192	0.394	0.189	0.392	0.53
Field: Creative arts	majori	0.084	0.278	0.086	0.280	-0.47
N		14,619		6,226		

Notes. Computations based on data from the 2010 CEQ. *SD* = standard deviation; *t* = *t*-statistic. All variables listed are 0/1 dummies, except for Good Teaching scale, age in years, and number of years spent enrolled. Significant *t*-statistics at the 5% level are in boldface.

4. Results

Consistent with much of the existing literature, we find a significant mode effect at the national level for the GTS. As shown in Table 3, graduates who responded by telephone rated their experiences on the GTS around 6.6 points higher, on average, than graduates who completed a self-administered survey when other characteristics are taken into account. (Recall from Section 2 that values of the GTS range from -100 to 100.) A similar mode effect was seen in the propensity-adjusted model, with an average GTS score 6.4 points higher for telephone respondents, all else being roughly equal. This is equivalent to around one fifth of a standard deviation on the GTS, which is hardly a trivial effect. The similar results from our raw and propensity-adjusted models suggest that the mode effect is largely independent of the difference in composition of the telephone and self-administered samples. Moreover, the size of the coefficient on the *telephone interview* variable is quite large when compared with other

covariates in our propensity-adjusted model. When considering covariates other than those related to field of education, which together explain the most variation in scores out of all collected variables (GCA & ACER, 2008), providing a response by telephone is second in effect only to graduating with an honours degree, which itself is associated with an average GTS score 11.5 points higher than graduating with a pass degree.

Table 3

Mode Effect Regression Estimates Before and After Propensity Adjustment

Variable	Name	Raw (unweighted) estimates			Propensity-adjusted estimates		
		B	*SE*	*t*	*B*	*SE*	*t*
Telephone interview	phone	6.5673	0.517	**12.71**	6.3667	0.547	**11.64**
Age in years	ageyrs	0.2705	0.045	**5.96**	0.2255	0.065	**3.45**
Male	male	0.3512	0.535	0.66	0.3570	0.598	0.60
Bachelor degree (honours)	bhons	11.4314	0.955	**11.97**	11.5443	1.220	**9.46**
Studied full-time	ftstudy	1.3773	0.877	1.57	1.1347	1.028	1.10
Studied on campus	oncmode	4.0705	0.905	**4.50**	3.8527	1.025	**3.76**
Australian citizen/resident	austres	-4.8278	0.866	**-5.57**	-5.0509	0.941	**-5.37**
Language other than English	nesb	-0.6642	0.703	-0.94	-1.1414	0.785	-1.45
Work type: full-time	worka	-1.8472	0.749	**-2.47**	-1.9861	0.884	**-2.25**
Work type: part-time	workb	1.4137	0.710	**1.99**	1.7656	0.818	**2.16**
Seeking work	seek	-2.0551	0.569	**-3.61**	-2.3172	0.679	**-3.41**
Further study: full-time	furstuda	2.4009	0.709	**3.39**	2.4133	0.807	**2.99**
Further study: part-time	furstudb	-1.5623	1.093	-1.43	-2.0970	1.336	-1.57
Field: Natural and physical sciences	majora	8.5682	0.929	**9.22**	7.8921	1.078	**7.32**
Field: Information technology	majorb	-0.3886	1.414	-0.27	0.8917	1.493	0.60
Field: Engineering and related	majorc	-8.1820	1.198	**-6.83**	-6.8160	1.269	**-5.37**
Field: Architecture and building	majord	3.1074	1.576	**1.97**	4.3844	1.637	**2.68**
Field: Agriculture, environmental and related	majore	14.0330	1.940	**7.23**	12.0037	2.249	**5.34**
Field: Health	majorf	3.7202	0.783	**4.75**	4.0997	0.904	**4.54**
Field: Education	majorg	4.7613	1.078	**4.42**	3.8784	1.219	**3.18**
Field: Society and culture	majorh	7.3441	0.793	**9.26**	6.6437	0.911	**7.29**
Field: Creative arts	majori	10.4318	1.007	**10.36**	10.3679	1.059	**9.79**
N		20,845			20,845		
Prob > *F*		0.000			0.000		
R^2		0.05			0.06		

Notes. Computations based on data from the 2010 CEQ. Dependent variable is GTS score. All variables listed are 0/1 dummies, except for age in years. *B* = unstandardised regression coefficient; *SE* = robust standard error; *t* = *t*-statistic; Prob > *F* = probability associated with *F*-statistic. Additional controls included for institution. Omitted reference categories for variables with more than one category are not working (work type), not studying (further study), and management and commerce (field). Both models are significant at p < 0.001. Significant *t*-statistics at the 5% level are in boldface.

While a national perspective is a useful point of departure, an examination of mode effects at the institutional level is a key focus of this study. There are two reasons for this. First, CEQ data are used primarily to monitor the performance of individual institutions, with broad national figures arguably of secondary importance. Second, while participating institutions were precluded from conducting their own telephone interviewing, the data collection process itself was not conducted centrally by a single agency, nor were institutions

required to follow a standard interview script (although doing so was recommended). As such, there was not a single telephone data collection process for the 2010 CEQ; there were potentially *nine*, although the extent to which these processes varied is unknown. We investigate these potential institutional differences by replicating the analysis described in Section 3 separately for each institution. For brevity, only the *telephone interview* coefficients from the nine propensity-adjusted models are presented in Table 4. Institutions have been de-identified.

Table 4

Propensity-Adjusted Mode Effect Regression Estimates, By Institution

Institution	Variable	Propensity-adjusted estimates					
		B	*SE*	*t*	*N*	**Prob > *F***	*R*^{2}
Institution 1	Telephone interview	5.3296	1.355	**3.93**	2,931	0.000	0.07
Institution 2	Telephone interview	11.3108	2.104	**5.37**	1,154	0.000	0.11
Institution 3	Telephone interview	9.4192	4.767	**1.98**	328	0.016	0.12
Institution 4	Telephone interview	6.7989	1.529	**4.45**	2,318	0.000	0.05
Institution 5	Telephone interview	11.7815	1.529	**7.71**	2,683	0.000	0.10
Institution 6	Telephone interview	4.6117	2.298	**2.01**	2,416	0.000	0.06
Institution 7	Telephone interview	7.4001	1.447	**5.11**	2,738	0.000	0.06
Institution 8	Telephone interview	3.3112	1.490	**2.22**	2,574	0.000	0.06
Institution 9	Telephone interview	4.7521	2.197	**2.16**	3,613	0.000	0.07

Notes. Computations based on data from the 2010 CEQ. Dependent variable is GTS score. All variables listed are 0/1 dummies. *B* = unstandardised regression coefficient; *SE* = robust standard error; *t* = *t*-statistic; Prob > *F* = probability associated with *F*-statistic. Additional controls included for age, sex, degree level, attendance type, attendance mode, residency, language spoken at home, broad field of education, work status, work-seeking status, and further-study status. All models are significant at $p < .05$. Significant *t*-statistics at the 5% level are in boldface.

The first point of interest in Table 4 is that significant mode effects were observed for all nine of the institutions that supplemented their CEQ data collection with telephone interviewing, even after controlling for an extensive array of background variables. Even more notable, however, is the extent to which these mode effects vary between institutions. Telephone respondents from Institution 5, for instance, provided mean GTS responses around 11.8 points higher than self-administered respondents from the same institution. Sizeable effects were also observed for institutions 2, 3, 4 and 7. Conversely, telephone respondents from Institution 8 rated their experiences on the GTS only 3.3 points higher, on average, than respondents from that institution who completed a self-administered survey.

5.   *Conclusions, Implications and Limitations*

The introduction of telephone data collection in the 2010 CEQ was unquestionably effective in improving response rates to the survey, which, at a national level, had been stagnating below the 50% for the entire decade through 2009. The relatively high national CEQ response rate of 52.6% in 2010 was undoubtedly assisted by the 11,720 CEQ responses collected by telephone interview. The results of this analysis suggest, however, that the adoption of telephone interviewing in conjunction with self-administered surveys as an official collection mode for the 2010 CEQ may have introduced bias as a result of these two collection modes producing non-equivalent results. Graduates who provided a response by telephone tended to rate their course experience more positively than graduates who

completed a self-administered paper or online survey, which held even after adjusting for selection bias and controlling for a wide range of other characteristics. This effect was far from uniform across participating institutions, which suggests variability in the non-standardised telephone data collection process employed for the 2010 CEQ. While these findings are notable for the CEQ, they are hardly unprecedented in the literature. A broad range of studies have observed similar effects in surveys that combine interviewer-administered and self-administered data collection modes.

It would be an overreaction to conclude from these findings that all future responses to the CEQ collected by telephone should be automatically excluded from official analysis. These findings do, on the other hand, make a very strong case for standardisation—and ideally centralisation—of the telephone data collection process. The mode effect observed for Institution 8 in our study was quite weak, equivalent to only around one tenth of a standard deviation on the GTS, which provides some evidence that mode effects can potentially be minimised. The manual governing the administration of the 2011 AGS, which was underway at the time of writing, specifies that all institutions collecting CEQ responses by telephone must adhere to a standard interview script (GCA, 2010b). It will be interesting to see whether this initiative minimises mode effects in the 2011 CEQ, or at least reduces the variation in these effects between participating institutions. Moreover, complete outsourcing of telephone data collection for the CEQ to a single agency, preferably one with experience in mixed-mode surveys that combine both interviewer-administered and self-administered modes, would surely be a 'gold standard' for standardised telephone data collection, and would reduce the cost to institutions through economies of scale. Edwards (2008) made similar recommendations.

It is important to acknowledge the limitation of our empirical approach. Since propensity scores are estimated solely on the basis of observed covariates, there remains the possibility of bias resulting from the omission of unobserved, and indeed *unobservable*, covariates that potentially could affect whether respondents complete the CEQ by telephone or self-administered survey. We have attempted to address this in our study by conditioning on a rich set of observed covariates, including ones related to a graduate's personal characteristics, previous course enrolment, labour market status and further study status at the time of the survey. Moreover, as noted by Stuart (2010), unobserved covariates are a cause for concern only when they are unrelated to observed covariates, since controlling for observed covariates also controls for unobserved covariates that are correlated with them. As such, judicious use of observed covariates can go some way to minimising the bias associated with unobservables (Bryson, Dorsett, & Purdon, 2002). As an example, motivation, which is not measured on the AGS, could affect whether graduates respond to the CEQ in a timely fashion. Observed covariates that are thought to be correlated with motivation, such as completing a double degree, may capture some of this effect. A further limitation is the lack of detailed information on the specific telephone collection methods employed by the nine institutions in our study. This information should ideally be collected, as it would provide important contextual material that could help to explain the considerable variation in mode effects observed for different institutions, and could be used to inform good practice.

Finally, with regard to the data from the 2010 CEQ, it is our view that all users should be mindful of these findings when analysing and interpreting figures produced from these data. While the impact of this mode effect is relatively minor at the national level, comparisons made at an institutional level are likely to be affected to a much greater degree.

References

Aquilino, W.S., & Lo Sciuto, L.A. (1990), Effects of interview mode on self-reported drug use. *Public Opinion Quarterly*, *54*, 362–395.

Ayidiya, S.A., & McClendon, M.J. (1990), Response effects on mail surveys, *Public Opinion Quarterly*, *54*, 229–247.

Braitman, L.E., & Rosenbaum, P.R. (2002). Rare outcomes, common treatments: analytic strategies using propensity scores. *Annals of Internal Medicine*, *137*, 693–695.

Bryson, A., Dorsett, R., & Purdon, S. (2002). *The use of propensity score matching in the evaluation of active labour market policies* (Working Paper No. 4). London: Department for Work and Pensions.

Carifio, J., & Perla, R.J. (2007), Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3*, 106–116.

Christian, L.M., Dillman, D.A., & Smyth, J.D. (2008). The effects of mode and format on answers to scalar questions in telephone and web surveys. In J. Lepkowski, C. Tucker, M. Brick, E. De Leeuw, L. Japec, P. Lavrakas, M. Link & R. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 250–275). New York: Wiley-Interscience.

Dillman, D.A., & Christian, L.M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods*, *17*, 30–52.

Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B.L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, *38*, 1–18.

Dillman, D.A., Sangster, R., Tarnai, J., & Rockwood, T. (1996). Understanding differences in people's answers to telephone and mail surveys. In M.T. Braverman & J.K. Slater (Eds.), *Current issues in survey research: New directions for program evaluation series* (pp. 45–62). San Francisco: Jossey-Bass.

Edwards, D. (2008). *CEQ phone response analysis*. Unpublished report, Australian Council for Educational Research, Melbourne.

Fowler, F.J., Jr., Roman, A.M., & Di, Z.X. (1998). Mode effects in a survey of Medicare prostate surgery patients. *Public Opinion Quarterly*, *62*, 29–46.

Graduate Careers Australia. (2009, July 8). *Minutes of the 38th meeting of the Survey Reference Group*.

Graduate Careers Australia. (2010a). *Code of Practice for the public disclosure of data from the Australian Graduate Survey*. Melbourne: Graduate Careers Australia. Retrieved September 16, 2011, from http:// start. graduatecareers.com.au/ucm/groups/ content/documents/document/gca001239.pdf

Graduate Careers Australia. (2010b). *Australian Graduate Survey 2011 manual: October 2010 round*. Melbourne: Graduate Careers Australia. Retrieved September 16, 2011, from http://start.graduatecareers.com.au/ucm/ groups/content/documents/document/gca001320.pdf

Graduate Careers Australia. (2011a) .*Graduate Course Experience 2010*. Melbourne: Graduate Careers Australia.

Graduate Careers Australia. (2011b) .*2010 CEQ Methodology*. Melbourne: Graduate Careers Australia.

Graduate Careers Australia & Australian Council for Educational Research. (2008). *Graduate Course Experience 2007*. Melbourne: Graduate Careers Australia.

Graduate Careers Australia & Australian Council for Educational Research. (2010). *Graduate Course Experience 2009*. Melbourne: Graduate Careers Australia.

Hirano, K., & Imbens, G. (2001). Estimate of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology*, *2*, 259–278.

Kelly, D., Harper, D.J., & Landau, B. (2008). Questionnaire mode effects in interactive information retrieval experiments. *Information Processing and Management*, *44*, 122–141.

Kertesz, S.G., Posner, M.A., O'Connell, J.J., Swain, S., Mullins, A.N., Shwartz, M., & Ash, A.S. (2009). Post-hospital medical respite care and hospital readmission of homeless persons. *Journal of Prevention & Intervention in the Community*, *37*, 129–142.

Krosnick, J., & Alwin, D.F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*, 201–219.

Krysan, M., Schuman, H., Scott, L.J., & Beatty, P. (1994) Response rates and response content in mail versus face-to-face surveys. *Public Opinion Quarterly*, *58*, 381–399.

McFarlane, B., & Garland, R. (1994). A comparison of mail and face-to-face survey methods: New Zealand health professionals. *Marketing Bulletin*, *5*, 71–81.

Schuman, H., & Scott, J. (1989). Response effects over time: Two experiments. *Sociological Methods and Research*, *17*, 398–408.

Stuart, E.A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21.

Tarnai, J., & Dillman, D.A. (1992). Questionnaire context as a source of response differences in mail versus telephone surveys. In N. Schwarz & S. Sudman (Eds.). *Context effects in social and psychological research* (pp. 115–129). New York: Springer-Verlag.

van Nunspeet, W., Cuppen, M., & van der Laan, P. (2011). *Redesigning Dutch social surveys: Efficiency and mixed mode*. Netherlands: Statistics Netherlands. Retrieved September 16, 2011, from http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/EN/Content/Events/ DGINS/Document__StatisticsNetherlands,property=file.pdf

Yang, M.Y., Falcone, A.E., & Milan, L.M. (2009, August). *Survey mode effects in two military surveys*. Paper presented at the Joint Statistical Meeting of the American Statistical Association, Washington, DC. Retrieved September 16, 2001, from http://www.amstat.org/ sections/srms/proceedings/y2009/Files/304857.pdf