Extract from article:

Duncan D. Nulty (2008) The adequacy of response rates to online and paper surveys: what can be done?, Assessment & Evaluation in Higher Education, 33:3, 301-314, DOI: 10.1080/02602930701293231

What is an adequate response rate?

It might be strictly more correct at this point to be asking what an adequate sample size is. However, in the context of teaching and course evaluation surveys, sampling is not likely to be in the minds of academics. It is much more likely that they will ask a question about response rates. Furthermore, if a determination is made regarding sample size, the size of the population being sampled needs to be known first and so the corresponding response rate can be readily calculated from these two figures.

Whether or not a response rate is adequate depends (in part) on the use that is being made of the data. If the data gathered from a teaching evaluation survey were to be used only to bring about improvements by that teacher, and there is even one response that provides information which can be used in this way, the survey's purpose has, at least in part, been served and the response rate is technically irrelevant. If such a single useful response were just one from (say) a hundred or more possible respondents, that is of no consequence—unless that response is entirely at odds with what the majority of other students would have said. A more likely outcome would be that a single response would be regarded as completely inadequate in the context of a summative appraisal of the performance of the teacher. Generally, course and teaching evaluation data are used for both of these purposes, and increasingly the latter (Ballantyne 2003).

Accepting that course and teaching evaluations are rarely conducted for solely formative purposes, there is certain to be widespread concern about the adequacy of the responses to these surveys. In part, this will translate into a concern about response rates. It should be noted however, that this concern occurs without sufficient awareness of the importance of sample size and population size.

Richardson (2005) cited Babbie (1973, 165) and Kidder (1981, 150–151) when stating that 50% is regarded as an acceptable response rate in social research postal surveys. Baruch (1999) researched the response rates reported by 141 published studies and 175 surveys in five top management journals published in 1975, 1985 and 1995. He found that the overall average response rate was 55.6%. Richardson (2005), however, indicated that the Australian Vice-Chancellors' Committee & Graduate Careers Council of Australia (2001) regarded 'an overall institutional response rate for the Course Experience Questionnaire (CEQ) of at least 70% [to be] both desirable and achievable' (4). But, in concluding comments, he stated: 'Response rates of 60% or more are both desirable and achievable for students who have satisfactorily completed their course units of programmes.' (p.409.), despite having noted earlier that this rate 'clearly leaves ample opportunity for sampling bias to affect the results' (406).

Assertions regarding the adequacy or otherwise of a particular percentage response rate appear to be made without reference to any theoretical justification—or to the total number of potential respondents. Behind the assertions appears to be a balance between rational and political considerations of acceptability. It would be better if there was a theoretically justified, systematic way to calculate the response rate required.

Calculating required response rates

When academics survey their students to gather opinions on their teaching, or the quality of courses, they may either ask every student enrolled in those courses to respond, or may select only a smaller sub-set of students. If every student is surveyed, the purpose is to establish the views of the entire group of students. In this instance the population is every student enrolled on the course.

When academics elect to survey a sub-set of the enrolled students, there is one of two purposes. They might only be interested in the opinions of that particular sub-set of students because they possess some characteristic that is of particular interest. For instance, the population could consist of only the mature-age students who are enrolled in the course. In these circumstances it follows that the academics have neither the interest nor the intent to deduce anything about other students, nor subsequently to take actions that in any way relate to those students or their views.

Alternatively, an academic might be interested in the views of all students enrolled on his/her course but simply finds it more practical to survey only one sub-set. In this case, the population remains all students enrolled on the course. The sub-set which is surveyed is a sample of that population. It is common that an academic may survey those students who attend a particular class on a particular day of the week and not other students who attend on other days. In these circumstances, the academic will seek to extrapolate findings from the sample to the population. Whether it is valid to do so is the issue.

In all three scenarios outlined above, it is unlikely that every student who is asked to respond to a survey will actually do so. As a result, there are a number of matters to consider before it is possible to determine whether it is valid to extrapolate findings derived from the students who did respond to either the sample from which they came or the population to which they belong.

In the first two scenarios, every student in the population is surveyed but not all respond. The respondents represent a non-random sample of the population. An appropriate question is whether the respondents differ systematically from the non-respondents, and if so, whether these differences would cause them to respond differently to the questions asked. If the answer to both questions is 'yes', the sample is biased and simple extrapolation of findings from the sample to the population is not valid.

It is reasonable to expect that any survey that samples a population (or that achieves only a sample by way of respondents) will incur some sampling error and possibly also some sample bias. The former is the extent to which any statistical measure applied to the sample (such as the mean) gives a result that deviates from the mean of the population as a result of random variation in the membership of the sample. The latter is where a statistical measure applied to the sample deviates from the population measure because of systematic bias in the membership of the sample. In principle, both can be reduced by increasing the sample size and/or response rate—however, neither of these steps *guarantees* a reduction in either error or bias (Dillman 2000).

There are different ways in which sample bias can be introduced. In the context of course and teaching evaluation surveys, sample bias might be introduced if the academic chooses to administer a survey in a daytime lecture in preference to an evening lecture. The evening lecture might consist of a higher proportion of people who are in full employment, study part time, and are older. The views of these people may deviate systematically from the views expressed by those who attend the daytime lecture.

Sample bias can also be introduced as a product of the survey method that is chosen. Watt et al. (2002, 329) have reported that web users are demographically different from other users. Salmon et al. (2004) reported that variance in data from web surveys was less than for paper surveys. It is reasonable to suppose that an online survey will attract responses from students who are demographically different from students who would respond to a paper survey.

Third, sample bias can be introduced because of systematic differences between respondents and non-respondents. As noted by Richardson (2005, 406), research shows that 'demographic characteristics of people responding to surveys are different from those who do not respond in terms of age and social class' (Goyder 1987, Chapter 5). While that may not matter to most academics conducting evaluations of their teaching and courses, Goyder more importantly reported that 'respondents differ from non-respondents in their attitudes and behaviour' (Goyder 1987, Chapter 7) and other research has shown that 'students who respond to surveys differ from those who do not respond in terms of their study behaviour and academic attainment …' (Astin 1970; Neilsen et al. 1978; Watkins & Hattie 1985, 406).

Richardson (2005) concluded: 'It is therefore reasonable to assume that students who respond to feedback questionnaires will be *systematically* different from those who do not respond in their attitudes and experience of higher education' (406, emphasis added) and furthermore, 'it is not possible to *predict* attitudes or behaviours on the basis of known demographic characteristics' (Goyder 1987, Chapter7, emphasis added). This means it impossible to use demographic data concerning students to construct a sampling frame that might seek to overcome sampling bias.

Thus, not only are the expressed views of respondents likely to be different from those of non-respondents but responses gathered using web surveys are likely to be different from those gathered using paper-based surveys.

In the face of evidence of this kind, are we still prepared to accept response rates of 50%–60%–70% as adequate? It seems reasonable to argue that despite our best efforts it will often be difficult and/or expensive to obtain response rates above 70%. Politically, it is discomforting to accept low response rates because the proportion of non-respondents may be too high for us to be sure that those who responded are representative of the others who did not. The issue becomes 'what are we prepared to accept?'. As such, there is some degree of arbitrariness about the decision.

But there is some theory to guide us in the domain of statisticians and mathematicians beginning with a seminal paper by Neyman (1934), which discusses 'the method of stratified sampling' compared with 'the method of purposive selection', followed in 1955 by a paper entitled 'A unified theory of sampling from finite populations' (Godambe 1955) and more recently a paper by Smith (1983), 'On the validity of inferences from non-random sample'. A more accessible account of the salient points has been provided in Chapter 5 of Dillman (2000, 194–213).

First, there is a systematic way to calculate the sample size required for a specified level of confidence in the result, in relation to a population of a specified size, with a specified degree of sampling error, given a specified level of probability for a particular answer to be provided by a respondent (Dillman 2000, 206–207).

Specifically, and in relation to the context of teaching evaluation, under the following conditions it is possible to use a formula provided by Dillman (2000) to calculate how many respondents are required (and therefore also the required response rate).

The conditions are:

- The total number of students in the population that is being surveyed is known.
- All students in the population are surveyed. (Note: It is not actually necessary to survey all the students, but this assumption is necessary for the argument being made about *response*

rate. In practice, if the reader wants to calculate sample size instead, the requirement to survey all the students can be removed.)

- There is a known probability of any one student providing a certain answer to a question on a survey.
- The required/desired level of accuracy of result is known or set.
- There is a known or chosen level of confidence required/desired for the same result to be obtained from other samples of the same size from the same total group of students in the course.

In order to seek to present data representing the 'best possible scenario' (i.e. one that maximizes the probability of needing the lowest response rates) the formula supplied by Dillman (2000) was initially applied with liberal conditions set. These were: to set a 10% sampling error (higher than the normal 3%), to assume a simple yes/no question is to be answered equally by respondents in 50:50 ratio (the most conservative situation), and to accept an 80% confidence level (much lower than the normal 95% used by statisticians).

However, in practice it is known that students' responses to questions on teaching and course evaluation surveys use the top ratings more frequently than the lower ones. Considering data gathered in one Australian university over an eight-year period with over 25,000 surveys using a 1 to 5 scale, actual percentages are 72% of students responding with a rating of 4 or 5, the remainder using a rating of 1, 2 or 3. Thus, the assumption of a 50:50 split on a 'yes/no' question can be altered to a (nominal) 70:30 split. Applying this more liberal condition yields lower required response rates, which are tabulated in Table 3 in the columns headed 'Liberal conditions'.

Columns under the heading 'Stringent conditions' present the required responses and response rates when more stringent (and more common) conditions are set: specifically 3% sampling error, and 95% confidence level.

Starting with the data from the liberal conditions, the table shows that for class sizes below 20 the response rate required needs to be above 58%. This is greater than the maximum achieved by all but one of the universities cited earlier when using paper-based surveys (that maximum was only a little higher at 65%). In other words, the table suggests that even the relatively good response rates obtained to paper surveys of teaching and courses are only adequate when the class size is 20 or higher—and, even then, only when liberal conditions in relation to the acceptable sampling error and required confidence level are acceptable.

Similarly, considering the response rates achieved with online surveys, the table shows that the highest response rate reported earlier (47%) is only adequate when class sizes are above (approximately) 30—and again, even then, only when liberal conditions in relation to the acceptable sampling error and required confidence level are acceptable.

In other institutions, such as Griffith University for example, class size (at best) needs to exceed 100 before its existing response rate of 20% can be considered adequate. In other words, for this institution, unless the response rate can be boosted, online surveys should not be used on classes with less than 100 students.

When the more traditional and conservative conditions are set, the best reported response rate obtained for on-paper surveys (65%) is only adequate when the class size exceeds approximately 500 students. The best reported response rates for online surveys (47%) are only adequate for class sizes above 750 students. The 20% response rate achieved for online surveys by Griffith University would not be adequate even with class sizes of 2000 students.

Table 3 is, however, only a *guide* as it is based on the application of a formula derived from a theory that has random sampling as a basic requirement. With teaching and course evaluations *this requirement is not met*. If the total enrolment of a course is sampled, it is generally a convenience sample—selecting all students who show up to the Monday daytime lecture for example.

Table 3. Required response rates by class size.

	‘Liberal conditions’		‘Stringent conditions’	
	10% sampling error; 80% confidence level; 70:30 split responses 4 or 5 compared with 1, 2, 3		3% sampling error; 95% confidence level; 70:30 split responses 4 or 5 compared with 1, 2, 3	
Total no. of students on the course	Required no. of respondents	Response rate required (%)	Required no. of respondents	Response rate required (%)
10	7	75%	10	100%
20	12	58	19	97
30	14	48	29	96
40	16	40	38	95
50	17	35	47	93
60	18	31	55	92
70	19	28	64	91
80	20	25	72	90
90	21	23	80	88
100	21	21	87	87
150	23	15	123	82
200	23	12	155	77
250	24	10	183	73
300	24	8	209	70
500	25	5	289	58
750	25	3	358	48
1000	26	3	406	41
2000	26	1	509	25

If all students enrolled are surveyed, or if a random selection of these are surveyed, random sampling is still not achieved in practice because those who respond are not a random selection. Indeed, those who respond are systematically different from those who do not, and that those who respond will be different depending on the method of evaluation selected (Astin 1970; Neilsen et al. 1978; Watkins & Hattie 1985; Goyder 1987; Watt et al. 2002).

Discussion

What are the consequences of ignoring these facts? If the sample size is too small, results obtained will not be representative of the whole group of students. That is, the results will suffer from both sample error and sample bias. This means that the results obtained (from a sample) are not likely to be an indication of what the group as a whole (the population) would have said. Given that the respondents may be systematically different from non-respondents it is possible that the feedback provided could influence an academic to respond in ways that are counter to what they would do if they had feedback from all students. Similarly, if the data are used summatively to judge a teacher’s performance, it may lead a person to make an erroneous judgement. Although academics (like the rest of us) have to make judgements all the time in the absence of useful information, it would be helpful if the parameters affecting the feedback were more transparently obvious. It would also be helpful if the information available was not itself misleading—as may be the case.

For example, let us consider a hypothetical scenario. If an online survey is used, the respondents are more likely to be students who are familiar with and able to use this medium. As such, these students may also comment more favourably regarding online teaching matters than the other students would. Hypothetically, these students may also constitute a minority. The result will be a survey with a low overall response rate, made up of students who are mostly familiar with, able to use and favourably disposed toward online teaching and learning provisions of the course. If this happens, and these are the only data considered, the academic concerned could form a false view that she/he should do more to boost the use of online teaching approaches.

It should be noted that the problem here is not simply that the responses to the survey have come from a minority of students, but that the survey results suffer from systematic bias. This means that these data may also misrepresent and misinform summative judgements regarding the performance of the teacher. Unfortunately, it is not possible to determine the direction of that bias. Although (in this hypothetical case) students responding to online surveys may be more positively disposed towards online teaching approaches, this does not mean that they will also be more positively disposed towards the teacher's teaching.

The hypothetical scenario above serves to illustrate another problem too: imagine an online survey of all students yields a 30% response and an on-paper survey of the same students yields a 60% response. The temptation would be to regard the results of the latter as more valid and more worthy of consideration. However, as already described above, it may be that the online survey attracted responses from those who predominantly make use of online teaching and learning resources, while the respondents to the paper survey may contain few of these people. Effectively the two surveys have sampled two different sub-groups of students with systematically different views which may (or may not) be reflected in the nature of their answers to survey questions (depending on the questions). *Neither* survey may be a valid reflection of the whole group but each one may be a valid reflection of each sub-group.

In practice, it is likely that only one of these two surveys would be conducted—the academic will not have both sets of data for comparison. The academic's responses to improve his/her teaching and/or his/her course might therefore be erroneous. Similarly, the data for either survey may be misleading if used for summative purposes. This is not a problem resulting from low response rate per se but, rather, a problem associated with the potential for systematic sample bias in respect of the respondents to any one survey type—or, indeed, any survey.

This last point takes us into territory that is beyond the scope of this paper. Suffice it to say that the design of a survey, not only the mode of administration, may also affect who responds to it and what they say. Thus, when interpreting survey results, it is important to think about what was asked, how it was asked and how these variables may have resulted in bias in respect of who responded, what they said *and* how these responses may have differed if the survey itself, the mode of administration and the resultant pool of respondents had been different. The implication is that data derived from surveys are likely to be somewhat more easily and validly used if the surveys themselves are appropriately designed and used for particular targeted purposes. Given that doing this is difficult, even in the best of worlds, this observation underscores the need to evaluate courses and teachers using multiple methods, and to carefully consider the differences between the pictures that emerge from each in order to triangulate a more accurate position.

It follows from all this discussion that, although Table 3 gives us a guide for response rates which could (in a theoretically ideal world) be considered adequate, the reality is that even if the response rates suggested are achieved, great care is needed to be sure that results for a survey are representative of the whole group of students enrolled. Although this is known, current practice frequently ignores this need for caution. Generic course and teaching surveys are often used to evaluate situations they were not designed for, and response rates which are below those advocated by Table 3 are generally accepted. Despite this a high weight is simultaneously placed on student evaluation results.