Highlights

◗ *Building on USA validation and over 200,000 responses in Australasia, ongoing and diverse validation is conducted of the Student Engagement Questionnaire (SEQ).*

◗ *Validation of the SEQ ensures it has robust technical properties, is underpinned by rigorous technical foundations, is appealing, is efficient to administer, and produces results which can be analysed and reported.*

◗ *The content validity of the instrument is assured by the instrument's derivation from research literature, and ongoing validation and review against new findings.*

◗ *Face validity has been checked via interviews and focus groups conducted in many institutions over the last four years.*

◗ *The instrument displays appropriate levels of construct validity, demonstrated via confirmatory structure and item response modelling, and relations between the scales*

◗ *The SEQ's response scales provide appropriate specificity of measurement*

◗ *There is a need to conduct a cross-institutional study of the criterion validity of engagement conditions and practices*

◗ *The SEQ scales display appropriate levels of internal consistency, and which vary across institutions and fields of education*

An overview of psychometric properties of the AUSSE Student Engagement Questionnaire (SEQ)

Overview

The quality of education is a product of what students do, and how teachers, support professionals and institutions support good educational practice. This means that measuring students' participation in good educational practices and measuring how institutions support such participation goes to the heart of educational quality.

An important link in this line of reasoning is that the instruments used for measurement provide valid, reliable and efficient measurement. This is essential, for otherwise insights into how students engage in education will be biased or diffuse and wrong decisions may be made that have serious implications for policy and practice.

To that end, this briefing provides an overview of the psychometric properties of the Student Engagement Questionnaire (SEQ). The SEQ is administered as part of the Australasian Survey of Student Engagement (AUSSE) (ACER, 2011), and in a range of other applied and scholarly research studies. Since 2006, the SEQ has been deployed to over 600,000 students at all but one Australasian universities and in a growing number of other higher education providers. Over 200,000 people have completed the inventory.

Psychometric properties of the Student Engagement Questionnaire (SEQ)

The SEQ is used under license from the Center for Postsecondary Education at Indiana University in the USA who run the National Survey of Student Engagement (NSSE) (NSSE, 2011). The SEQ has close links with the instrument used in NSSE, the College Student Report (CSR), but has been extensively revised, developed and validated for Australasian higher education.

The CSR and SEQ are based on decades of scholarly research, and since 1999 the CSR has been administered at over 1,300 institutions and subjected to numerous tests and improvements. This guide does not report all of this research, but provides a high-level overview of key outcomes and properties. Related resources have been prepared by Kuh (2009) and Coates (2006, 2010), which are recommended as companion documents to this briefing.

Instrument contents

The SEQ is designed for administration to undergraduate students in under 15 minutes in online or paper form. The instrument is designed for English-speaking students. A copy of the 2010 paper version of the instrument is provided on the AUSSE website: http://ausse.acer.edu.au. This same basic content is used with all students, although there are multiple versions of the online form to ameliorate the influence of unreached item-non-response. Modified versions of the SEQ have been developed for academic staff, postgraduate students and for use within non-university training contexts. This briefing analyses the properties of the main higher education student version of the SEQ.

The SEQ contains around 150 items that operationalise the concept of student engagement. This concept is concerned with whether students participate in effective educational practices, and whether institutions support such engagement. Most of the items from the NSSE instrument have been retained, and a number of new items have been added to measure phenomena of particular relevance in Australasia. A number of 'experimental' items are tested in the online version each year to facilitate the ongoing growth of the instrument. The SEQ measures key demographic and context information, and aligns with all relevant Australian standards classifications.

More than a dozen different response scales are used, which enhances the character of the form and the richness of the resulting data. The SEQ items use an eclectic range of individual activity response scales (for instance: 'Never', 'Sometimes', 'Often', 'Very Often'; or 'None', '1 to 2', '3 to 4', '5 to 6', 'More than 6') and response scales which invite reflection on institutional emphasis (for instance: 'Very little', 'Some', 'Quite a bit', 'Very much'). The careful choice of response scales is important as it enhances the 'actionability' of the results – the capacity to convert evidence into change. The instrument does not measure student happiness but the deeper notion of educational engagement, and as a result 'agreement' response scales are not used.

In addition to providing information of their own, many SEQ items tap a range of more general important educational phenomena. A selection of these items are grouped together to measure seven scales. These scales are summarised in Table 1.

The SEQ also provides measurement of the six outcome measures summarised in Table 2. The measure of average grade is captured through a single item on the SEQ. The other five measures are composite measures which reflect responses to a group of items.

Table 1 AUSSE scale descriptions

Scale	Label	Description
Academic Challenge	AC	Extent to which expectations and assessments challenge students to learn
Active Learning	AL	Students' efforts to actively construct their knowledge
Student and Staff Interactions	SSI	Level and nature of students' contact with teaching staff
Enriching Educational Experiences	EEE	Participation in broadening educational activities
Supportive Learning Environment	SLE	Feelings of legitimation within the university community
Work Integrated Learning	WIL	Integration of employment-focused work experiences into study
Career Readiness	CRE	Preparation for participation in the professional workforce

Table 2 AUSSE outcomes measures

Measure	Label	Description
Higher Order Thinking	HOT	Participation in higher-order forms of thinking
General Learning Outcomes	LRN	Development of general competencies
General Development Outcomes	DEV	Development of general forms of individual and social development
Average Overall Grade	GRD	Average overall grade so far in course
Departure Intention	MOB	Non-graduating students' intentions on not returning to study in the following year
Overall Satisfaction	OVL	Students' overall satisfaction with their educational experience
Career Readiness	CRE	Preparation for participation in the professional workforce

It is important to note that the CSR and hence the SEQ are instruments specifically designed to measure a reasonably large number of aspects of student engagement. An important process for people to work through is determining the relevance of various items to their context. A related activity might involve using the items to form a number of additional scales. A range of different scales have been developed by researchers in the USA (see: http://www.nsse.iub.edu). People with access to AUSSE data can engage in a similar process in Australasia. This briefing reports results from analyses undertaken on the scales and measures in Table 1 and Table 2.

Validation approach

Validation of the SEQ has ensured that the instrument has robust technical properties, that it is underpinned by rigorous technical foundations, that it is appealing, that it is efficient to administer, and that it produces results which can be analysed and reported.

A range of qualitative and quantitative procedures were used to validate the SEQ in 2006 and 2007 before deploying it in Australasia. The multifaceted validation work replicated and advanced the approach implemented in developing other instruments used widely in Australian higher education. The work is ongoing in nature, and supports the ongoing growth and refinement of the instrument.

The validation work included focus groups, cognitive interviews, expert review, pilot testing and review, psychometric modelling (Rasch item response modelling and congeneric measurement modelling), analysis of differential item functioning (DIF), and reliability analyses. The qualitative analyses were used to refine item wording, for instance, while the quantitative analyses helped ensure that the scales measured the target constructs with acceptable levels of bias and precision. Of course, this work built on the extensive validation undertaken over a decade of use in the USA (Kuh, 2009).

The summary that follows reports on relevant research and validation work undertaken during the development of the SEQ, and more recent analyses of the 2008 cross-institutional data. It does not report a comprehensive technical examination of the instrument – which could run to many hundreds of pages – but surveys what are likely to be among the most important properties.

Establishing content validity

Steps were taken during instrument development to account for the content validity of the Student Engagement Questionnaire. In general, content validity refers to the extent to which an instrument provides measurement of all facets of a construct. What this means is that the SEQ should measure all relevant facets of the construct of student engagement. First, then, it is necessary to define and document this construct. Having done this it is possible to map the instrument against the construct to ensure content coverage. A range of strategies can then be used to evaluate and refine the content coverage.

Student engagement is an emergent and complex psychosocial phenomenon which has its substance defined in terms of higher education research and practice. Conceptually, 'student engagement' has been defined as a meta-construct for bringing together the large and diverse amount of research undertaken on student learning and development. This includes many thousands of analytical and empirical studies. Useful syntheses of these as they pertain to student engagement have been given by Pascarella and Terenzini (2001, 2005), Kuh, Pace and Vesper (1997), Kuh, Schuh and Whitt (1991), Kuh (2004, 2008), Ewell and Jones (1996), Pace (1979, 1988, 1995), Tinto (1993), Astin (1985, 1990, 1993) and Coates (2006).

The College Student Report on which the SEQ is based was developed to operationalise the construct of student engagement for the purposes of measurement. There is an intimate link between the conceptual foundations and the instrument. A critical feature of the SEQ is its foundation in empirically based theories of student learning. Items in the SEQ are based on findings from decades of research on the activities and conditions linked with high-quality learning. This

foundation helps assure the educational importance of the phenomena measured by the instrument. Items are not included in the instrument because they are simply seen to reflect good ideas or reflect the consensus of stakeholders. Indeed, a criterion for including any item on the form is that it is based on empirical research (normally of a longitudinal nature) which affirms the educational significance of the phenomenon.

A range of validation strategies were used to verify the link between the instrument and the research foundations. Items were mapped against key themes in several meta-analyses to ensure sufficient content coverage. The instrument was also assessed by dozens of practitioners and research experts on university student learning and development. These processes resulted in certain additions and deletions. Items on 'spirituality' were dropped from the SEQ, for instance, and items on online learning and careers advice were added. Several items were added to measure early departure intentions and the concept of work-integrated learning. This empirical work provides a means of ensuring the relevance of the instrument and its underpinning constructs to the Australasian context.

While a considerable amount of this work was undertaken in 2007 during the initial validation of the SEQ (see: Coates, 2008a), it also forms part of the ongoing continuous improvement process. As new research and insights come to light and further studies are conducted on the SEQ (Radloff & Coates, 2010), changes are made to the instrument and its underpinning research foundations. Since 2007, feedback from thousands of people has been used to improve the instrument. Good examples are work on leadership and career readiness (Edwards & Coates, 2008; Scott, Coates & Anderson, 2008), which led to new items being trailed in the 2009 collection. As with other aspects of the instrument, the specification of principles and processes provides a foundation for ongoing improvement.

Checking face validity

The purpose of face validity is to determine the extent to which a test on its face value appears to measure the constructs that it purports to measure. Face validity assessments are subjective in nature, and are best obtained from representatives of stakeholders associated with the instrument. While among the least 'technical' forms of validity, face validity may indeed be one of the most

important. If key stakeholders do not see the instrument as relevant and have confidence in what it measures, then it is unlikely to be seen as relevant and useful. This is particularly important in a post-compulsory environment involving autonomous institutions – engaging leaders and learners in the process is a formative and necessary part of the enhancement equation.

The face validity of the instrument can cover a range of areas, including item wording, appearance, length, language, response scale characteristics, relevance, and the usability of results. A considerable amount of validation work has been undertaken in the USA to establish the face validity of the CSR. Work with the SEQ built on this through an extensive and iterative process of consultation and review.

During the initial development of the SEQ, draft items were sent to people at all universities in Australia for reflection and review. Feedback was received from a large number of people working as DVCs, PVCs, policy and planning staff, student support professionals, academic developers, and administrative personnel. This feedback played a vital role in improving the relevance and appeal of the items.

The items were compiled into an operationalisable format and a paper and online form were designed. These forms provided the basis for a series of further reviews. They were sent to stakeholders at universities and a range of other organisations. Feedback from experts on higher education policy and practice were factored into the emerging instrument.

Building on the extensive validation work undertaken by Coates (2006), several student focus groups were also conducted at this stage, with these being led by a five different people. The focus groups were conducted with a heterogeneous groups of students at a range of institutions in both Australia and New Zealand. In addition to the focus groups a series of cognitive interviews were undertaken with first and later year students. These helped probe student interpretations of items and aspects of the questionnaire, such as language and presentation that could be improved.

Confirming construct validity

While face and content validity rely to a large extent on the judgement of selected individuals, construct validity

is established through psychometric modelling. An integrated series of psychometric procedures were used during instrument development to analyse the primary data collected in the research and explore the construct validity of the SEQ. These were initially run as part of a pre-fieldwork pilot test conducted in 2007, and have been revised following each cross-institutional administration.

Construct validity can be approached in various ways, but one of the most important considerations is that the items in each scale provide what is referred to as 'unidimensional' measurement of its target construct. Unidimensionality is most commonly investigated using exploratory or confirmatory factor analysis, or item response modelling. Results from this kind of analysis can be extensive, and a small selection is included to indicate the analytical methods that have been used.

Figure 1 shows correlations between the SLE construct and its six constituent items. These estimates range from 0.44 to 0.77, indicating that the observed items load well on the underpinning scale. This also affirms that the single construct explains approximately between 20 and 60 per cent of the observed variability in each item. Confirmatory factor modelling produced similar results for the other scales.

Figure 2 shows a variable map for the Active Learning scale, which is produced via Rasch item response modelling (IRT). The top part of this map shows linear scoring units on the left hand side followed by the distribution of student responses (each 'x' reflects around 150 students). The items are listed on the right side of the variable in the form of letter and number combinations. With these, for instance, "G.3" reflects the point on the variable at which a student has a 50 per cent chance of responding to the third or fourth response options for that item. The legend that maps letters to items is shown in the bottom-half of the figure. This lower portion of the figure also shows a plot of item fit statistics. These reflect the level of relationship between each item and the underpinning variable. Normally, these should range between 0.80 and 1.20.

Together, this display shows that the items in the Active Learning scale spread out well across the underpinning construct. They are not all clumped in the middle of the variable, for instance, indicating that they provide measurement of students who have low scores on this facet of student engagement as well as those who have high scores. That the item and student distributions are reasonably well aligned means that the items are well targeted to students' engagement.

The fit statistics at the base of Figure 2 show that the active learning items have a high level of fidelity to the variable. Similar results were found from item response modelling of all other scales, with no fit statistics lying above the upper threshold of 1.20. Figures above this threshold reflect a more random relationship between the item and underpinning variable. As shown in this Figure, all items A to G have ordered item threshold estimates (for example, A.3 > A.2 > A.1) on the scale. The fact that categorical item locations are in order reflects that response options of all items in this scale work as expected, that a higher response option of an item indicates a higher location of this item option on the scale.

Divergent validity is another facet of construct validity, which is concerned with whether the instrument measures scales that are not too highly related and hence could be considered discrete. This can be tested by looking at how each individual item relates to each scale – an elaborate process – or simply by looking at the percentage of shared variability between each scale.

Table 3 reports the percentage of shared variance between the AUSSE scales and outcome measures. These statistics indicate how much variation in the data is shared by two scales. By way of example, 17 per cent of the variability in AC scores is shared with AL scores. There is only 11 per cent of joint variation between DEV and HOT. The largest amount of joint variation is between LRN and DEV, which are understandably correlated constructs.

The performance of SEQ response scales was investigated. There has been debate about the a priori validity of different response scales (Preston and
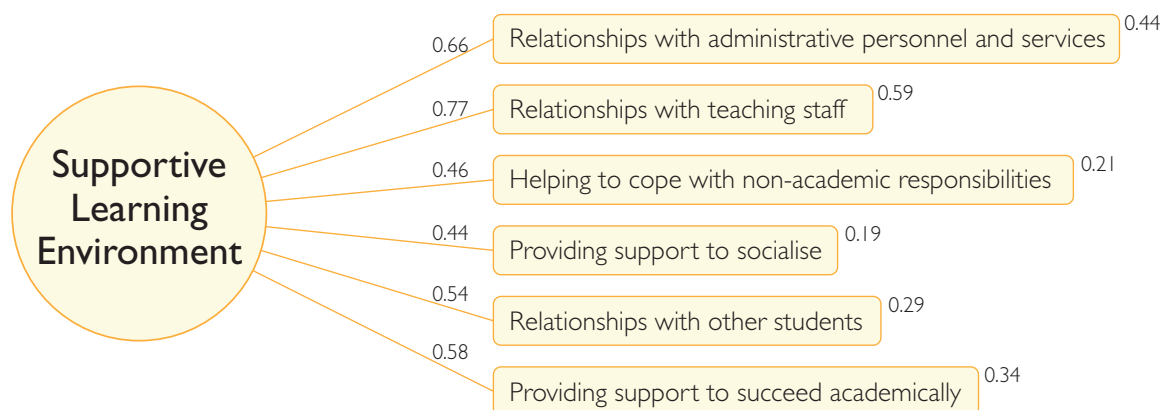

Figure 1: Supportive Learning Environment (SLE) cogeneric measurement model

```
SCORE                STUDENTS    |  ITEMS
                                 |
                                 |
100                              |  G.3
                                 |  D.3
                          X      |
                                 |  B.3
                          X      |
                         XX      |  A.3 G.2
                                 |  D.2 E.3
 84               XXX            |  F.3
                XXXXXX           |  C.3
                                 |
              XXXXXXXX           |  G.1
            XXXXXXXXXXX          |  D.1
                      X          |  B.2
 67        XXXXXXXXXXXXXX        |
          XXXXXXXXXXXXXXXX       |
                     XX          |
         XXXXXXXXXXXXXXXXX        |  A.2 C.2 E.2
                     XX          |  F.2
        XXXXXXXXXXXXXXXXXXXX     |
                      X          |
 50        XXXXXXXXXXXXXXXXX     |
                      X          |
           XXXXXXXXXXXXXX        |
                      X          |  B.1
                                 |
            XXXXXXXXXXX          |
                                 |  A.1
 33                              |  E.1
                XXXXXX           |
                                 |
                                 |
                                 |
                                 |  F.1
                    XXX          |
 17                              |
                                 |  C.1
                                 |
                                 |
                                 |
                     X           |
                                 |
  0                              |
```

```
                  FIT STATISTICS

          0.77      0.83      0.91      1.00      1.10      1.20
------------+---------+---------+---------+---------+---------+-----
A Asked questions           *          |
B Made a presentation           *      |
C Collaborated in class                |    *
D Collaborated beyond class            |      *
E Tutored others            *          |
F Service learning                     |                 *
G Discussed ideas                      |        *
------------+---------+---------+---------+---------+---------+------
```
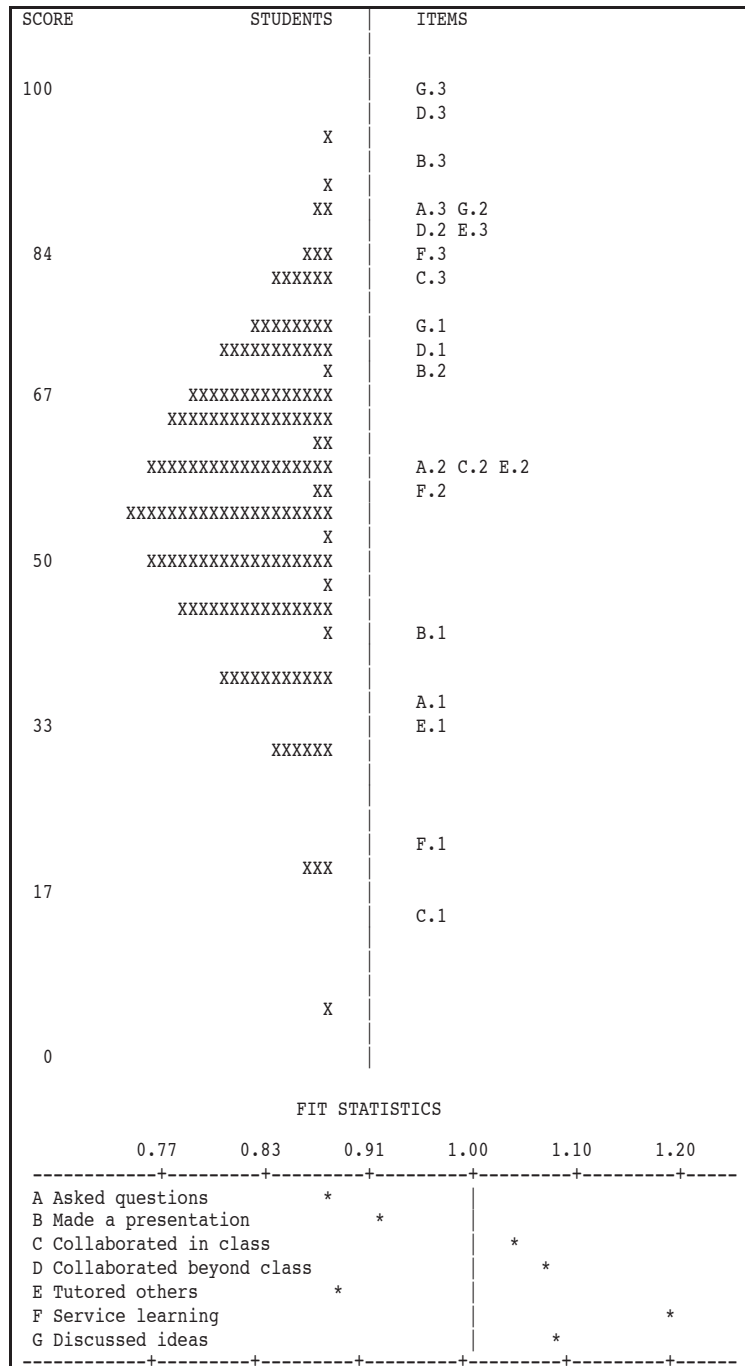
Figure 2: Active Learning scale variable map

Colman, 2000; Barnett, 1999; Dubois and Burns, 1975; Bock and Jones, 1968; Lecointe, 1995; Linacre, 1999). What is most important, however, is that the response scales function appropriately in empirical contexts.

Item response modelling was used to explore the functionality of the several response scales used on the instrument. According to the measurement model used for the analysis (Wright & Masters, 1982), the probability of a student selecting a particular response is a function of both student and item parameters. Most importantly, this measurement model allows one to test if invariance item ordering and invariance person ordering can be achieved. Invariance item ordering means that any estimated item location on the scale is independent from sample distribution, and

Table 3 AUSSE scale shared variance (per cent)

	Engagement scales				
	AC	AL	SSI	EEE	SLE
AL	17				
SSI	14	26			
EEE	12	21	14		
SLE	8	8	13	8	
WIL	11	13	10	15	6
	Outcome measures				
	HOT	LRN	DEV	GRD	DI
LRN	18				
DEV	11	40			
GRD	2	1	0		
DI	1	4	1	1	
OS	6	22	14	3	13

invariance person ordering means that any estimated person location on the same scale is independent from subset of items. One popular way to checking the attainment of invariance ordering is thought the testing of goodness of fit to model, which is carried out by checking global fit statistics and fit statistics at item level. In the current study, these parameters are each student's sense of their own engagement, and the aspect and extent of engagement being probed by the item. When reasonably good fit statistics are observed, it implies that for each item, as for all items, students with higher levels of engagement are more likely to select higher level response categories than students reporting lower levels of engagement.

By way of example, Figure 3 plots the likelihood of a student selecting a particular response category as a function of measured level of engagement. The graph shows one line per response category. The horizontal axis shows the range of scores that can be obtained on the items. The graph shows that each response category is at some point the most likely to be chosen. The graphs show that the 'sometimes' category was the least significant, although there is also evidence that this category is certainly playing a role in the measurement process.

All measurement instruments contain bias. Hence the existence of bias can be taken for granted. What is important is understanding the magnitude and prevalence of bias. If bias is large and widespread, this is more concerning than if is moderate and related to a specific group. Bias is a dimension of validity, for if a measuring instrument is biased in a certain context then it would not be said to provide valid measurement of the construct in that context. It is important to get a feel for such variations in performance, and this is typically done through the analysis of differential item functionality. Differential item functionality examines
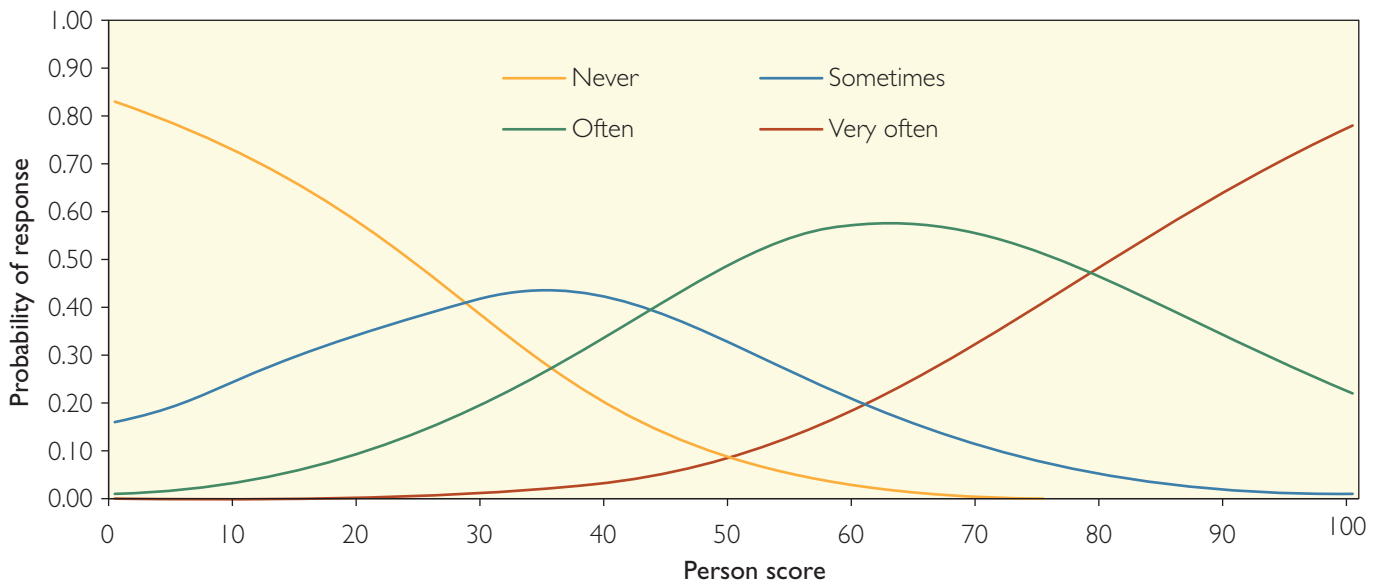

Figure 3: Sample SEQ response category probability curves

bias in an instrument by testing whether respondents from different populations have the same level of engagement score the same on an item. If they do not, then the item can be said to demonstrate differential performance across these groups.

An analysis of SEQ item bias was conducted using Conquest (ACER, 2011) item response modelling software. This compares item estimates for specified groups and considers whether the differences are statistically significant. The analysis was replicated for group of potential relevance, including sex, campus, country, field of education and institution.

Figure 4 shows a typical result for items in the Academic Challenge scale. The reporting units are called logits, a metric that is a by-product of the Rasch item response modelling. The estimates are similar for both males and females, particularly after taking into account uncertainties due to the randomness in the data (displayed in terms of 95% confidence intervals). Males appear to find the item about 'working harder than you thought you could to meet standards' harder than females. However, after taking into account error margins this difference is shown to be insignificant. Similar variations were evident for other item and group combinations. Importantly, there did not appear to be any systematic patterns of differential functioning across groups.

An important bias consideration is whether the instrument performs differently with campus-based and distance students. This is an important issue given the role of distance education to Australasian higher education. The issue is complex, however, for while it is often commonly argued that distance education is educationally equivalent to face-to-face provision, it is often simultaneously argued that distance education is a different paradigm that should be evaluated using different criteria. In relation to the SEQ, what is important is that the phenomena are valid in a distance context and that they are interpreted by respondents in an appropriate way.

Consideration for distance students were embedded in the development of the SEQ. This involved consultation with several experts on distance education, review of relevant research (Coates, 2008b), and conceptual analysis of items and scales.

This led to various additions to existing items, and to the addition of whole new items. Certainly, there are a few items on the SEQ that are not likely to be applicable to distance learners – as with different items for other groups of learners – and specific instructions were added to the form that respondents should leave an item blank if it did not apply to them. Extensive analysis of the responses provided by distance learners
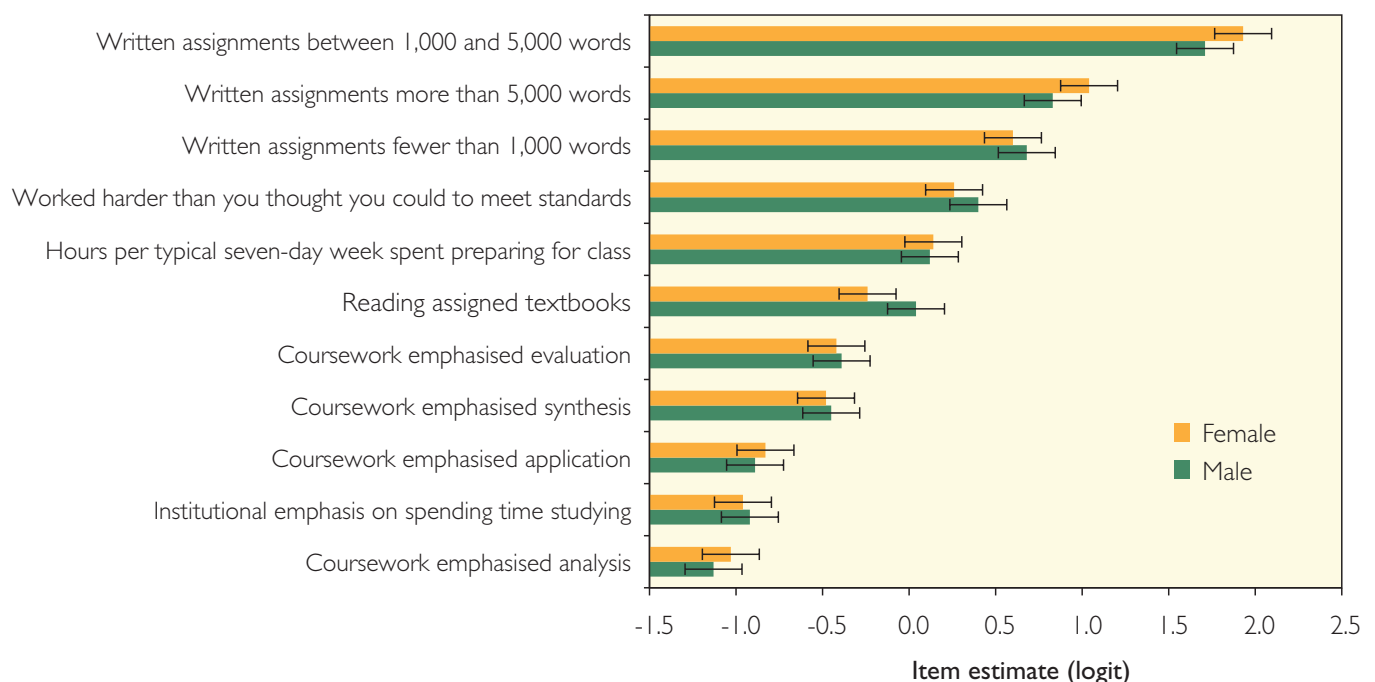

Figure 4: Comparison of male and female Academic Challenge item estimates

Table 4 Correlations between engagement scales and outcome measures

		HOT	LRN	DEV	GRD	DI	SAT
First year	AC	83	47	39	13	-5	25
	AL	31	35	33	11	-3	19
	SSI	25	31	35	3	-1	20
	EEE	28	31	34	7	-4	20
	SLE	26	50	47	11	-19	55
	WIL	27	44	34	9	-8	24
Later year	AC	81	46	38	16	-6	26
	AL	32	34	29	15	-3	18
	SSI	27	30	33	11	-3	23
	EEE	28	28	29	11	-4	18
	SLE	25	49	47	9	-21	56
	WIL	26	40	29	13	-11	23

was conducted in 2008 and reported in an earlier AUSSE Research Briefing (Coates, 2008b), which showed that the profile of distance learners matched that of on-campus learners in several respects. Analyses of differential item functioning revealed no difference in the psychometric performance of the items across campus-based or distance contexts.

In general, the SEQ appears valid and relevant for the vast majority of distance learners. As with any interpretation of survey results, however, it is imperative to take learner and educational contexts into account. This is particularly important when the focus of the investigation is students and educational interactions.

Reviewing criterion validity

Criterion validity is concerned with the relationship between performance on the target instrument and a specific criterion. This is referred to as concurrent validity when the criterion is measured contemporaneously to the target instrument, and predictive validity when the criterion is measured at a future point in time. For a range of practical and technical reasons, studies of criterion validity are relatively complex and rare (Coates, Tilbrook, Guthrie & Bryant, 2006), yet they are considered among the most important means of establishing the measurement properties instrument.

To a certain extent criterion validity has been designed into the SEQ. This is because unless an activity has been connected positively with concurrent or future indicators of learning it is not reflected in a SEQ item or scale. As Kuh (2001) writes, "the voluminous research on college student development shows that the time and energy students devote to educationally purposeful activities is the single best predictor of their learning and personal development".

Very few criterion validity studies are conducted on the surveys, tests and examination instruments used with university students. In recent years a number of studies have been conducted on high-stakes admissions tests (Coates, 2008c; Coates, Edwards & Friedman, 2010; Coates & Friedman, 2010). The SEQ contains a student identifier, and several institutions have used this to examine how engagement relates to other data they have on students. As yet, there have been no cross-institutional studies in Australasia that have examined how engagement and self-reported outcome scores relate to other relevant performance indicators. There would be value in conducting a complex and large-scale study of this kind in Australasia.

A large-scale study of criterion review was conducted in the USA (Kuh, Kinzie, Cruce, Shoup & Gonyea, 2006). This research uncovered key relationships between engagement and grades and persistence, including that:

1. Engagement has positive, modest effects on grades and persistence for students from different racial and ethnic backgrounds, even after controlling for key pre-college variables.

2. Engagement has compensatory effects on first-year grades and persistence to the second year of college at the same institution for historically underserved students.

The SEQ does measure a number of self-reported outcomes, and comparing engagement scale scores against these offers an index of criterion validity. Correlations (scaled onto a reporting metric ranging between 0 and 100) are reported in Table 5. For both first- and later-year students there is a consistently positive relationship between the engagement scales and outcome measures (Departure Intention is scored negatively, so a negative correlation provides evidence that greater engagement leads to a lower departure intention). Such results provide support for the criterion validity of the instrument.

Psychometric properties of the Student Engagement Questionnaire (SEQ)

Table 5: Scale internal consistency by field of education

	AC	AL	SSI	SLE	WIL	HOT	LRN	DEV	OVL
Sciences	68	64	71	75	69	74	84	83	76
IT	70	63	70	75	70	73	84	82	82
Engineering	70	60	69	71	65	70	83	79	75
Architecture	69	66	72	75	69	79	84	83	79
Agriculture	66	64	75	76	67	74	86	85	79
Health	68	61	70	76	68	78	84	82	80
Education	69	60	67	76	65	78	86	84	82
Business	70	65	68	76	67	78	86	84	79
Humanities	68	62	68	75	69	76	84	84	78
Creative arts	71	61	67	74	70	75	82	83	80

Internal consistency

'Internal consistency' refers to the extent to which items work together to provide reliable measurement of a target scale. There are various measures of internal consistency, the most common being the chronbach alpha statistic. The alpha statistic is a lower bound estimate of reliability due to the measurement assumptions that underpin the formula.

Reliability is a property of data rather than items. Rather than look at one static estimate of reliability it is appropriate to examine variation across contexts (Thompson, 1994; Vacha-Haase, 1998; Fan and Thompson, 2001). Accordingly, Table 5 and Figure 5 show the distribution of chronbach alpha reliability statistics for each scale across fields of education and institutions. No statistics are shown for Average Overall Grade (GRD) and Departure Intention (MOB) as these are single item indicators. In addition, no measure is provided for the EEE as this scale simply reflects a count of student participation in a range of discrete and hitherto relatively low incidence extracurricular activities. The boxes in Figure 5 represent the minimum, median and maximum values.

Can we trust student self reports?

Yes we can. Studies have shown that student perceptions gathered using questionnaires are a reliable and accurate source of information about the quality of education (Marsh, 1987, 1990; Ramsden, 1991; Hu & Kuh, 2001; Brennan, Brighton, Moon, Richardson,

Figure 5: Scale internal consistency by institution

Rindl & Williams, 2003; Bradburn & Sudman, 1988). Indeed, for a concept like student engagement collecting information from students themselves may be the only feasible method of measuring the target constructs.

Of course, certain situations could work to reduce the validity and reliability of self reports. Reports may be influenced if people are asked about matters which are high stakes to them. For certain issues, self-efficacy may play role and induce students to under-report their activities. Conversely, there may be a halo effect in which students inflate certain facets of their performance. Such factors have been explored in studies such as those listed above, which affirm the integrity and credibility of student self reports.

Integrating a range of findings, Hu and Kuh (2001) identify student ratings as valid under the following conditions:

- the information is known to the respondent;
- questions are phrased clearly and unambiguously;
- questions refer to recent activities;
- respondents take the questions seriously; and
- answering the questions does not threaten, embarrass or violate the privacy of the respondent or encourage the respondent to respond in socially desirable ways.

The influence of non-response

Non-response in large-scale surveys is a pervasive multifaceted phenomenon (Coates, 2004). It may manifest in the form of survey non-response in which a member of the sample does not return a response. It may involve skipping behaviour, in which a respondent provides a completed survey yet responses to various items are missing. Unreached non-response is a very common form of missing data. Non-response is an important phenomenon as it has the potential to reduce the reliability and validity of results in unexpected ways.

Analysis of responses to SEQ items reveals no appreciable pattern of skipped non-response. That is, students tend to provide a response to those items with which they interact. For the paper form, there is also very little unreached non-response – missing data that appears after the last item completed by respondents. Unreached non-response is a pervasive phenomenon with online instruments, however, and is associated with particular characteristics of the online instrument. The time that it takes to complete a questionnaire is an important characteristic, and it is for this reason that the SEQ is limited to 15 minutes. The number of items per page is not strongly related to item completion. The number of page skips, however, has a direct relationship

to a decline in item responses. To counteract this, there are three online versions of the SEQ, each of which presents the items in different orders.

A valid and efficient instrument

This briefing has reported key psychometric characteristics of the SEQ, the instrument used in the Australasian Survey of Student Engagement. While there are always variations in the performance of instruments deployed in large-scale surveys, read as a whole the findings affirm the SEQ's sound psychometric properties. This is not surprising given its history and the steps take for its validation.

Of course, the properties of an instrument are not static but change over time with new research insights, contexts, technologies and student profiles. Rather than a series of static insights, therefore, the evidence and insights presented in this briefing provide a foundation perspective for further review development. While the SEQ has sound psychometric properties, one of its most important characteristics may well be that these properties provide a robust basis for change. The capacity for change is one of the greatest strengths of the instrument, for it underpins the instrument's ongoing relevance and validity. To maintain the integrity of the instrument, it is imperative that such change be led in research-based, educationally informed, and practically responsive ways.

Resources

Astin, A.W. (1985). *Achieving Educational Excellence: A critical analysis of priorities and practices in higher education*. San Francisco: Jossey Bass.

Astin, A.W. (1990). *Assessment for Excellence: The philosophy and practice of assessment and evaluation in higher education*. New York: Maxwell Macmillan International.

Astin, A.W. (1993). *What Matters in College: Four critical years revisited*. San Fransisco: Jossey Bass.

Australian Council for Educational Research (ACER). (2011). *Australasian Survey of Student Engagement (AUSSE)*. Accessed 1 January 2011 from: http://ausse.acer.edu.au

Australian Council for Educational Research (ACER). (2011). *Conquest Item Response Modelling Software*. Camberwell: ACER.

Barnett, J.J. (1999). *Likert response alternative direction: SA to SD or SD to SA: Does it make a difference?* Washington: ERIC.

Bock, R.D. & Jones, L.V. (1968). *The Measurement and Prediction of Judgement and Choice.* San Francisco: Holden Day.

Bradburn, N.M., & Sudman, S. (1988). *Polls and Surveys: Understanding what they tell us*. San Francisco: Jossey Bass.

Brennan, J., Brighton, R., Moon, N., Richardson, J., Rindl, J. & Williams, R. (2003). *Collecting and Using Student Feedback on Quality and Standards of Learning and Teaching in Higher Education: A report to the Higher Education Funding Council for England*. Bristol: Higher Education Funding Council for England.

Coates, H. & Friedman, T. (2010). Evaluation of the Special Tertiary Admissions Test (STAT). *Higher Education Policy and Management, 32*(2), 117-126.

Coates, H. (2004). Analysing item non-response. *Journal of Applied Measurement*, 5(1), 70-94.

Coates, H. (2006). *Student Engagement in Campus-based and Online Education: University connections*. London: Taylor and Francis.

Coates, H. (2008a). *Attracting, Engaging and Retaining: New conversations about learning*. Camberwell: Australian Council for Educational Research.

Coates, H. (2008b). Enhancing the Engagement of Distributed Learners. *AUSSE Research Briefing, 3*. Camberwell: ACER.

Coates, H. (2008c). Establishing the criterion validity of GAMSAT. *Medical Educator, 42*, 999-1006.

Coates, H. (2010). Development of the Australasian Survey of Student Engagement (AUSSE). *Higher Education, 60*(10), 1-17.

Coates, H., Edwards, D. & Friedman, T. (2010). *Evaluation of the Student Aptitude Test for Tertiary Admission (SATTA) Pilot Program*. Canberra: Department of Education, Employment and Workplace Relations.

Coates, H., Tilbrook, C., Guthrie, B. & Bryant, G. (2006). *Enhancing the GCA National Surveys: An examination of critical factors leading to enhancements in the instrument, methodology and process*. Canberra: Department of Education, Science and Training.

Dubois, B. & Burns, J.A. (1975). An analysis of the meaning of the Question Mark Response Category in Attitude Scales. *Educational and Psychological Measurement, 35*, 869-884.

Edwards, D. & Coates, H. (2008). *Development and Piloting of the Victoria University Student Dividend.* Footscray: Victoria University.

Ewell, P.T. & Jones, D.P. (1996). *Indicators of "Good Practice" in Undergraduate Education: A handbook for development and implementation*. Colorado: National Centre for Higher Education Management Systems.

Fan, X. & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement, 61*(1), 517-531.

Hu, S. & Kuh, G. D. (2001). *Being (dis)engaged in educationally purposeful activities: The influences of student and institutional characteristics*. Paper presented at the American Educational Research Association Annual Conference.

Kuh, G.D. (2001). *The National Survey of Student Engagement: Conceptual framework and overview of psychometric properties*. Bloomington: Indiana University, Center for Postsecondary Research.

Kuh, G.D. (2004). *The national survey of student engagement: Conceptual framework and overview of psychometric properties*. Bloomington: Indiana University, Center for Postsecondary Research.

Kuh, G.D. (2008). *High-impact educational practices: What they are, who has access to them, and why they matter*. Washington: Association of American Colleges and Universities.

Kuh, G.D. (2009). The national survey of student engagement: Conceptual and empirical foundations. *New Directions for Institutional Research, 141*, 5-21.

Kuh, G.D., Kinzie, J., Cruce, T., Shoup, R., & Gonyea, R. M. (2006, July). *Connecting the dots: Multi-faceted analyses of the relationships between student engagement results from the NSSE, and the institutional practices and conditions that foster student success: Final report prepared for Lumina Foundation for Education*. Bloomington: Indiana University, Center for Postsecondary Research.

Kuh, G. D., Pace, C. R., & Vesper, N. (1997). The development of process indicators to estimate student gains associated with good practices in undergraduate education. *Research in Higher Education, 38*(4), 435-454.

Kuh, G.D., Schuh, J.H. & Whitt, W.J. (1991). *Involving Colleges*. San Fransisco: Jossey Bass.

Lecointe, D.A. (1995). *How the collapsing of categories impacts the item information function in polytomous item response theory*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Linacre, J.M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Marsh, H.W. (1987). Students evaluations of university teaching: Research findings, methodological issues and directions for future research. *International Journal of Educational Research, 11*(3), 253-388.

Marsh, H.W. (1990). Multitrait-multimethod analysis. In J.P. Keeves (Ed.), *Educational Research, Methodology and Measurement: An international handbook*. Oxford: Pergamon.

National Survey of Student Engagement (NSSE) (2011). *National Survey of Student Engagement*. Bloomington: Indiana University, Center for Postsecondary Research.

Pace, C.R. (1979). *Measuring Outcomes of College: Fifty years of findings and recommendations for the future*. San Francisco: Jossey Bass.

Pace, C.R. (1988). *Measuring the Quality of College Student Experiences: An account of the development and use of the college student experiences questionnaire.* Los Angeles: Higher Education Research Institute, University of California.

Pace, C.R. (1995). *From good practices to good products: Relating good practices in undergraduate education to student achievement*. Paper presented at the Association for Institutional Research, Boston.

Pascarella, E.T. & Terenzini, P.T. (1991). *How College Affects Students: Findings and insights from twenty years of research*. San Francisco: Jossey Bass.

Pascarella, E.T. & Terenzini, P.T. (2005). *How college affects students: A third decade of research*. San Francisco: Jossey Bass.

Preston, C. & Colman, A.M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power and respondent preferences. *Acta Psychological, 104*(2), 1-15.

Radloff, A. & Coates, H. (2010). *Doing More for Learning: Enhancing engagement and Outcomes.* Camberwell: Australian Council of Educational Research.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: the Course Experience Questionnaire. *Studies in Higher Education, 16*(2), 129-150.

Scott, G., Coates, H. & Anderson, M. (2008). *Learning Leaders in Times of Change: academic leadership capabilities for Australian higher education*. Sydney: Australian Learning and Teaching Council.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*(4), 837-847.

Tinto, V. (1993). *Leaving College: Rethinking the causes and cures of student attrition*. Chicago: University of Chicago Press.

Vacha-Haase, T. (1998). Reliability generalisation: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6-20.

Wright, B.D., & Masters, G.N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

acer.edu.au