Doing what it says on the tin? A psychometric evaluation of the Assessment Experience Questionnaire

John Batten[a], Tansy Jessop[b]* and Phil Birch[c]

[a] *Department of Sport, Exercise and Health, The University of Winchester, Winchester, UK*
[b]*Solent Learning and Teaching Institute, Solent University, Southampton, UK*
[c]*Institute of Sport, University of Chichester, Chichester, UK*

Biographical note:

John Batten is a Senior Lecturer in Sport and Exercise Psychology and Programme Leader BSc (Hons) Sport and Exercise Science at the University of Winchester. His primary research interests include expectancy effects in education, health and sport.

Dr Tansy Jessop is Professor of Research Informed Teaching at Solent University. She leads the Transforming the Experience of Students through Assessment (TESTA) project. She has published on assessment and feedback, learning spaces, and social justice in education.

Dr Phil Birch is a Senior Lecturer in Sport and Exercise Psychology at the University of Chichester. His main research interests include psychometrics and using think aloud protocols to examine cognitions in golfers.

This is an Accepted Manuscript of an article published by Taylor & Francis in ASSESSMENT AND EVALUATION IN HIGHER EDUCATION on 4 November 2018, available online: https://www.tandfonline.com/doi/full/10.1080/02602938.2018.1499867.

Doing what it says on the tin? A psychometric evaluation of the Assessment Experience Questionnaire

Abstract

The Assessment Experience Questionnaire has been widely used to measure conditions of learning from assessment. It is one of three methods used in the 'Transforming the Experience of Students through Assessment' research process, originally funded by the Higher Education Academy to explore programme assessment patterns, and now used extensively in universities in the United Kingdom. Given the growth of assessment and feedback research over the last decade, the Assessment Experience Questionnaire is ripe for revision. Critics have queried its theoretical and statistical robustness. This study investigated the psychometric properties of the Assessment Experience Questionnaire, as the first step in the process of strengthening the instrument. Specifically, we examined the validity of the questionnaire with a sample of final year undergraduate students from eight UK universities (n = 633). Results were mixed, confirming that the questionnaire has some value, but indicating that not all sub-scales possess adequate psychometric properties to underpin confident conclusions. As a result, we have embarked on a process of making conceptual modifications to the Assessment Experience Questionnaire, both to update the theoretical constructs, and to ensure stronger overall validity. This article indicates the direction of these modifications, which will be outlined in a second article.

Key words: assessment environment; student experience; deep learning; questionnaire validity

This is an Accepted Manuscript of an article published by Taylor & Francis in ASSESSMENT AND EVALUATION IN HIGHER EDUCATION on 4 November 2018, available online: https://www.tandfonline.com/doi/full/10.1080/02602938.2018.1499867.

Introduction

The context of assessment and feedback

Research has shown that assessment has a profound influence on students' learning and their study behaviour (Ramsden 2003; Gibbs & Simpson 2004; Gibbs 2006; Nicol & McFarlane-Dick 2006; Harland et al. 2015; Wass et al. 2015). Assessment outcomes have long-term effects, powerfully influencing graduate opportunities, employability and life-long learning (Boud 2000; Knight & Yorke 2003; Boud & Falchikov 2006). Until recently, research on assessment and feedback in the United Kingdom (UK) has focused on modules, for example, the Higher Education Academy Subject Centre projects examining assessment design at a modular level (Knight & Yorke 2003). The modular design of assessment contributes to a host of unintended consequences for student learning that may only be visible at a programme-level (Jessop, El Hakim and Gibbs 2014a; Harland et al. 2015). Modular degrees tend to fragment degree coherence, interfere with clear lines of progression, and place 'slow learning' at risk (Claxton 1998; Bloxham & Boyd 2007; Harland et al. 2015; Berg & Seeber 2016). There is growing sector-wide recognition that changing assessment to improve student learning requires a programme approach to assessment and feedback design (Gibbs & Dunbar-Goddet 2009; Jessop, El Hakim and Gibbs 2014a; Jessop & Tomas 2017).

Research linked to the 'Transforming the Experience of Students through Assessment' (TESTA) project (2009 -) has played a significant role in shifting sector-wide thinking towards a programme view of assessment, described as prompting a 'step change' in the sector by Professor Sue Bloxham (Personal correspondence 2016). TESTA has been used in

more than 50 UK universities to investigate the student experience of modular assessment and feedback and to provide programmatic strategies, based on well-founded assessment principles. These principles derive from established research on the value of formative assessment and feedback for student learning (Sadler 1989; Black & Wiliam 1998; Nicol & McFarlane-Dick 2006; Hattie & Timperley 2007); setting high expectations and challenging tasks that require student effort (Innis 1996; Chickering & Gamson 1987; Gibbs & Simpson 2004; Arum & Roksa 2011); the opacity of written criteria for clarifying goals and standards and the value of social practice in making complex judgements about standards (Shay 2005; Orr 2007; O'Donovan, Price and Rust 2008); and feedback that is developmental, dialogic and given in time for students to make use of it (Higgins, Hartley and Skelton. 2002; Gibbs & Simpson 2004; Nicol 2010). Overall, TESTA builds on theories about deep and surface approaches to learning (Marton & Saljo 1976). It explores the impact of too much assessment *of* learning which measures and grades students, in relation to too little assessment *for* learning, which provides opportunities for students to develop and fine-tune their work. TESTA identifies the big picture of assessment and feedback from a student perspective, providing evidence of problematic patterns, as well as strategies to address them (Jessop, El Hakim and Gibbs 2014a; Jessop & Maleckar 2016; Jessop & Tomas 2017; Jessop 2017).

The Assessment Experience Questionnaire

The Assessment Experience Questionnaire Version 3.3 (AEQ 3.3) is central to TESTA's mixed-methods approach, which triangulates data from three sources: an audit to collect data through a discussion with the programme team leader about assessment in the planned curriculum; the AEQ 3.3; and focus groups with final year students about their lived experience of assessment on the whole programme. The foundations for all three methods relate to assessment principles distilled in the literature, particularly about conditions of

assessment to improve learning (Gibbs & Simpson 2004).

The AEQ 3.3 examines the extent to which students experience various conditions of learning on whole programmes of study (Gibbs & Dunbar-Goddet 2007; 2009). It consists of 28 items across nine sub-scales linked to conditions of learning from assessment, with one *Overall Satisfaction* item. Students respond on a five-point Likert scale, ranging from strongly disagree (1) to strongly agree (5). The nine sub-scales on the AEQ 3.3 are: *Quantity of Effort*; *Coverage of Syllabus*; *Quantity and Quality of Feedback*; *Use of Feedback*; *Appropriate Assessment*; *Clear Goals and Standards*; *Deep Approach*; *Surface Approach*; and *Learning from the Examination*.

The triangulated TESTA method has high face validity and has proven its worth as "a way of thinking about assessment and feedback" (Jessop, El Hakim and Gibbs 2014b). However, given academics' criticism of the AEQ 3.3, the authors have been prompted to re-examine its validity, which has provided an opportunity to re-examine underpinning theoretical constructs, given recent developments in the field.

Why revise the Assessment Experience Questionnaire?

The quality of research depends partly on the reflexivity of researchers (Cousin 2009), particularly the ability to recognise, and where possible, address, limitations of methods, data collection and analysis, as well as the interpretations that arise from research. Reflexivity is often associated with qualitative research but its principles apply equally to quantitative research in that researchers need to constantly question: "how they know what they know" (Goodall 2000, 137). No questionnaire or research tool is the 'final word', theoretically or

This is an Accepted Manuscript of an article published by Taylor & Francis in ASSESSMENT AND EVALUATION IN HIGHER EDUCATION on 4 November 2018, available online: https://www.tandfonline.com/doi/full/10.1080/02602938.2018.1499867.

statistically, given the dynamic fields of exploration and practice, coupled with the need to update research in light of new knowledge. An iterative process of development is part and parcel of questionnaire design in higher education research, as seen, for example, in revisions to the Course Experience Questionnaire (CEQ):

> *Given the dynamic nature of higher education and our current steep learning curves about how best to measure 'teaching quality', there is little possibility that the CEQ will be immutable over time - modifications are a necessary outcome of a participatory process which seeks to address the perceptions and needs, and thus earn the confidence, of institutions (Wilson, Lizzio and Ramsden, 1997, 35).*

The concept of questionnaires 'evolving over time' is set against a desire for stability and the need for time-series data to provide trend analyses, such as in nationally administered surveys, for example, the CEQ and National Student Survey.

This study is a response to the questions of academics and experts who have criticised the validity of the AEQ 3.3 as an instrument. Criticisms of the AEQ 3.3 have fallen into three main categories. First, like many questionnaires, the AEQ 3.3 has borrowed items and sub-scales from elsewhere, especially from the Approaches to Study Inventory (ASI) and the CEQ, both of which were validated in large scale testing (Meyer & Parsons 1989; Wilson, Lizzio and Ramsden, 1997). However, the items for *Appropriate Assessment*, *Clear Goals and Standards*, *Surface Approach* and *Deep Approach* appear to have been incorporated into the AEQ 3.3 without further testing, with Cronbach's alpha reliability scores unavailable for

these borrowed scales (Gibbs and El Hakim 2011). A second criticism is that some of the sub-scales are low on items; e.g., *Quantity of Effort*. Third, academics have expressed unease about sub-scales like *Coverage of Syllabus* which, for many, run counter to the idea of a university, both suggesting textbook-like content, as well as a content-focused approach to the curriculum. For these reasons, along with the need to bring the AEQ 3.3 in line with more contemporary theories and approaches, we decided to review and evaluate it. Indeed, users of the AEQ 3.3 will only have confidence in its data if they trust the validity of the instrument and its underpinning theoretical framework.

Responding to critiques of the Assessment Experience Questionnaire

In response to our use of the AEQ 3.3 within TESTA in many universities, as well as the growing sense that the instrument needs to be reviewed and further developed, the authors embarked on a concurrent three-stage process of construct development, validation and statistical testing to check the status of the AEQ 3.3. This process is the precursor to redeveloping the AEQ 3.3. In this article, we evaluate the psychometric properties of the AEQ 3.3, with a particular focus on its distributional properties (means, standard deviation, and normal distribution), internal consistency (Cronbach's alpha coefficients) and inter-factor correlations, as well as an examination of factorial validity using Exploratory Factor Analysis (EFA). These empirical tests of validation were conducted alongside a participatory developmental process with assessment experts to update and sharpen the theoretical focus of the AEQ 3.3. A further paper will explore the theoretical constructs of the revised AEQ. The focus here is the statistical examination of the AEQ 3.3 and how this has contributed to the further development of the AEQ.

This is an Accepted Manuscript of an article published by Taylor & Francis in ASSESSMENT AND EVALUATION IN HIGHER EDUCATION on 4 November 2018, available online: https://www.tandfonline.com/doi/full/10.1080/02602938.2018.1499867.

Method

Participants

We recruited final year undergraduate students (n = 633) from eight universities from across the UK. 591 students were studying for single honours degrees, while the remaining 42 students were studying on combined honours degree programmes. The breakdown of courses represented in our sample was: n = 237 students from professional courses; n = 218 students from humanities; n = 208 students from sciences.

Sample sizes for conducting EFA are considered good if more than 300 and very good if more than 500 (Comrey & Lee 1992). Stevens (1996) suggests using five participants per variable to obtain adequate power. Given that the AEQ 3.3 consisted of 28 items, we needed a minimum sample of 140 participants. However, we also used the Kaiser-Meyer-Olkin (KMO) test to compute for sampling adequacy. The value obtained was 0.794, which according to Hutcheson and Sofroniou (1999) is good, as it is above the recommended value of 0.60 required to conduct an adequately powered EFA (Garson, 2006).

Procedure

We provided final year undergraduate students with an information sheet and consent form, demographic questionnaire, as well as a copy of the AEQ 3.3. The AEQ 3.3 was distributed to students during lectures and seminar classes, completed in the presence of a fully briefed research assistant in order for any questions to be answered. No course tutors were present. This prevented any ethical compromises around the anonymity of data. The AEQ 3.3 took approximately 10 minutes to complete. We recruited participants in different universities over

a two year period beginning in 2015. Participants were assured of confidentiality and anonymity, as well as their right to withdraw from participation at any point. Institutional ethical approval was obtained prior to data collection.

Results

Distributional properties

Prior to running EFA, we performed Kolmogorov-Smirnov and Shapiro-Wilk tests to assess the distribution of items. Means and standard deviations were calculated to complement these analyses (Table 1). Establishing whether items on the AEQ 3.3 are normally distributed is important when drawing inferences from the data. However, contrary to the recommendations in the scale development literature (Tabachnick & Fidell 2001; Byrne 2010), the results show that all AEQ 3.3 items were non-normally distributed ($p < 0.05$). Other research has used response variability as a means to assess an instrument's distributional properties whereby item standard deviations > 1 are deemed satisfactory (Hall et al. 1998; Cumming et al. 2005; Williams & Cumming 2011). Given that item standard deviations either approached or were > 1 in the present study, these findings demonstrate some preliminary support for the distributional properties of the AEQ 3.3.

INSERT TABLE 1 HERE

Internal consistency

Cronbach's alpha reliability coefficients were performed to assess the inter-correlations among items and, in turn, to estimate response consistency (Vaughn, Lee & Kamata 2012). Cronbach's alpha reliability coefficients equivalent to (α) > 0.70 are acceptable (Nunnally &

This is an Accepted Manuscript of an article published by Taylor & Francis in ASSESSMENT AND EVALUATION IN HIGHER EDUCATION on 4 November 2018, available online: https://www.tandfonline.com/doi/full/10.1080/02602938.2018.1499867.

Bernstein 1994). Table 2 illustrates that three out of AEQ 3.3's five sub-scales, namely *Use of Feedback, Appropriate Assessment* and *Learning from Examinations*, demonstrated adequate internal reliability ($\alpha > 0.70$). Two further sub-scales (*Quantity of Effort, Quantity and Quality of Feedback*) approached this threshold ($\alpha = 0.65$). Conversely, *Surface Approach, Deep Approach, Coverage of Syllabus* and *Clear Goals and Standards* demonstrated inadequate values that may be indicative of poor items and thus poor sub-scales. The negative scores are of particular concern here. However, we can confirm that there are no coding errors present, in particular linked to reverse coded items, which is a common cause of negative values in Cronbach's alpha. Yet, caution should still be exercised when evaluating the internal consistency of AEQ 3.3 items due to the low number of items making up each sub-scale (Schmitt 1996).

Table 2 compares our analysis to original Cronbach's alpha scores for the AEQ 3.3 provided for the widely used questionnaire (Gibbs and El Hakim 2011).

INSERT TABLE 2 HERE

Inter-factor correlations

We conducted Spearman's Rank Order Correlation Coefficients to assess the degree to which the sub-scales (factors) within the AEQ 3.3 are related. Questionnaires typically exhibit factors that are related to some extent, but are not so related that they measure the same concept (Byrne 2010). In this study, we conducted bivariate correlations and identified no

This is an Accepted Manuscript of an article published by Taylor & Francis in ASSESSMENT AND EVALUATION IN HIGHER EDUCATION on 4 November 2018, available online: https://www.tandfonline.com/doi/full/10.1080/02602938.2018.1499867.

evidence of multicollinearity (two factors measuring the same concept), with all correlations below 0.80 (Stevens 1996; see Table 3). However, questionnaires with sound psychometric properties typically yield weak (r = 0.3 - 0.4) to moderate (r = 0.6 - 0.7) inter-factor correlations (Tabachnick & Fidell 2001). Yet, despite correlations generally reaching significance, most of the correlations observed in this study were very weak (< 0.2: Fallowfield, Hale and Wilkinson 2005). Such correlations indicate that the factors making up the AEQ 3.3 may be more distinct than anticipated.

INSERT TABLE 3 HERE

Exploratory Factor Analysis

EFA examines the adequacy of AEQ items to measure the hypothesised factor structure and to examine which items form coherent subsets, remaining relatively independent of one another. Given that the AEQ 3.3 was underpinned by independent latent constructs, implying that factors were more independent than related by nature, we conducted Principle Components Analysis (PCA) with orthogonal (Varimax) rotation to explore the underlying factor structure of the AEQ 3.3 (Tabachnick & Fidell 2001).

Through PCA, we identified eight factors with eigenvalues > 1. Eigenvalues represent the amount of variation explained by a factor, whereby an eigenvalue > 1 represents a substantial amount of variation. In an effort to enhance the clarity of the identified eight-factor solution, a secondary EFA was conducted whereby the number of factors to be extracted was restricted to eight. The secondary eight-factor model explained approximately 60.08% of the

cumulative variance, allowing the complexity of the data set to be reduced with a loss of 40% of information.

The purpose of the EFA was to identify items that loaded onto their hypothesised factors, items that loaded onto multiple factors (cross-loading), items that loaded onto an incorrect factor (misloadings), and items that did not load satisfactorily onto any factor. Factor loadings > 0.71 are thought to be excellent, 0.63 to 0.70 very good, 0.55 to 0.62 good, 0.45 to 0.54 fair, with 0.32 to 0.44 considered poor (Comrey & Lee 1992). Table 4 shows the associated variables, as well as the rotated factor loadings of the secondary eight-factor mode l.

The *Appropriate Assessment*, *Learning from the Examination*, *Use of Feedback* and *Deep Approach* factors show promise with respect to their psychometric properties. In contrast, the *Quantity of Effort* and *Coverage of Syllabus* factors are questionable as factors. Specifically, the *Coverage of Syllabus* factor included items from the *Quantity of Effort* factor. In addition, two items (AEQ5R and AEQ11) were negative, suggesting problematic relationships with the other items in the *Coverage of Syllabus* factor. Two further items (AEQ3R and AEQ12R) cross-loaded and one item (AEQ28) mis-loaded, leaving the *Clear Goals and Standards* and *Quantity and Quality of Feedback* sub-scales with two items each. Overall, factor analysis indicated a mixed picture of the validity of the AEQ 3.3, confirming academics' anxieties about various sub-scales.

INSERT TABLE 4 HERE

Discussion

The purpose of this study was to examine the factorial validity of the AEQ 3.3. In response to critique, we were attempting to discover if there were good grounds for revising the questionnaire after more than a decade of use, in line with new developments in the research field and the iterative process of questionnaire development. The findings of the EFA provided limited support for the hypothesised factor structure of the AEQ 3.3.

The most problematic finding was the identification of two overlapping AEQ factors, namely *Quantity of Effort* and *Coverage of Syllabus*. The *Quantity of Effort* factor seeks to understand students' perceptions of time-on-task, in relation to the amount and distribution of effort in studying for their degrees (Gibbs and Simpson 2004). *Coverage of Syllabus*, on the other hand, investigates to what extent the assessment environment enables students to be selective and strategic about what they study, as well as what they can afford to neglect without significant consequences for their achievement. The EFA confirms that these two factors are not distinct from one another.

The overlap between *Quantity of Effort* and *Coverage of Syllabus* is compounded by the fact that *Quantity of Effort*, as a two-item factor, is weak anyway, and that *Coverage of Syllabus* is questionable given the implication of a 'syllabus' to be covered. Indeed, *Coverage of Syllabus* resonates strongly with a content-centred rather than a learner-centred paradigm of assessment and course design (Fink 2003). While knowledge does matter in assessing student outcomes (Wheelahan 2010; Harland & Wald 2018), knowing and understanding theory is not the same as covering content. Students need to learn critical, analytical and creative ways of knowing, as well as skills of inquiry and application at university, in order to integrate that knowledge (Fink 2003). This is very different from the language of 'covering' the syllabus,

This is an Accepted Manuscript of an article published by Taylor & Francis in ASSESSMENT AND EVALUATION IN HIGHER EDUCATION on 4 November 2018, available online: https://www.tandfonline.com/doi/full/10.1080/02602938.2018.1499867.

which is reminiscent of the banking model of education (Freire 1970). The finding that *Coverage of Syllabus* is not statistically robust is in line with academics' qualms about the factor.

On the other hand, *Quantity of Effort* is a factor which should be strengthened to reflect the importance of effort, independent study and time-on-task in student learning (Chickering & Gamson 1987; Gibbs and Simpson 2004; Arum & Roksa 2011). Assessment is a powerful driver of student effort (Ramsden 2003), but summative assessment *of* learning has limited value in distributing effort, especially on modular degrees with their tendency to encourage compartmentalisation and disposable learning (Jessop, El Hakim and Gibbs 2014a; Jessop & Tomas 2017). The extent to which the *Quantity of Effort* factor adequately represents this construct is a problem that needs to be addressed in the redevelopment of the AEQ.

Both *Quantity and Quality of Feedback* and *Appropriate Assessment* have been problematic factors for interpretation in the TESTA process. *Quantity and Quality of Feedback* has caused confusion by virtue of its double agenda – addressing both quantity *and* quality of feedback in one construct. *Appropriate Assessment* has baffled many of those trying to make sense of student scores, given it resembles *Surface Approach* items: "The staff seemed more interested in testing what I had memorised than what I understood" (AEQ10R); "Too often the staff asked me questions just about facts" (AEQ14R); "To do well on this course all you needed was a good memory" (AEQ18R). *Appropriate Assessment* only makes sense when users understand its emphasis on the assessment environment as the driver for deep or surface learning as distinct from an individual taking a particular approach to learning. Both *Quantity and Quality of Feedback* and *Appropriate Assessment* contain only three items. As two items from these factors cross-loaded or misloaded, these factors are effectively weak two-item

scales. Specifically, in *Quantity and Quality of Feedback*, item AEQ3R "I received hardly any feedback on my work" cross-loaded with the *Appropriate Assessment* factor. Surprisingly, however, the item that has cross-loaded with the *Appropriate Assessment* factor is not *Surface Approach* item, but rather a *Clear Goals and Standards* one: "It was often hard to discover what was expected of me in this course" (AEQ12R). Perhaps this reflects the fuzziness of the factor.

Problems with loading occur when items do not adequately represent their factors. Our evaluation showed that some existing AEQ 3.3 items do not fully encapsulate the core facets of the focal construct (MacKenzie, Podsakoff & Podsakoff 2011). For example, the EFA showed that AEQ3R (I received hardly any feedback on my work) and AEQ12R (It was often hard to discover what was expected of me in this course) were both inadequate representations of their factors. The structure, clarity and specificity of items is weak here, leading to mis-loading and/or cross-loading (Ibid 2011). Indeed, the exact phrasing of items can exert an influence on the construct being measured (Clark & Watson 1995). This lack of item clarity and specificity may contribute towards the limited factorial validity of the AEQ 3.3, with some items being open to misinterpretation by being quite vague, making it difficult for respondents to answer accurately.

Overall, 20 out of 25 items (80%) have very good or above relationships with their associated factors (Comrey & Lee 1992). However, while in general, the items are good representations of their respective factors, there are some problematic issues compromising the factorial validity of the AEQ 3.3. The lack of conceptual clarity between the *Coverage of Syllabus* and *Quantity of Effort* factors is one, while the low number of items on three factors (*Quantity of Effort, Quantity and Quality of Feedback, Clear Goals and Standards*) is another. Experts

agree that a minimum of three items is required to adequately represent a given factor (Clark & Watson 1995; Kline 2000; MacKenzie, Podsakoff & Podsakoff 2011). These problems cast some doubt over the conceptual coverage of the AEQ 3.3 as it stands. In addition, Cronbach's alpha scores question the internal consistency of some of the AEQ 3.3's sub-scales. Although scrutiny of 'scale if items deleted' values might help to identify particularly problematic items here, the other – and arguably more important – conceptual issues identified above suggest that there is limited value in doing this at this stage.

While we agree with the premise that a questionnaire is only as good as its capacity to measure what it says it will measure 'on the tin', factorial validity is only one part in the complex process of questionnaire design. The nature of the constructs and the strength of theory underpinning those constructs is another. Thus, there are also some modifications we propose on theoretical grounds. First, the feedback scale should be strengthened to incorporate ideas about dialogic and personal feedback that have become increasingly pertinent in the literature and in TESTA findings about students' disconnection from feedback across modules (Nicol 2010; Batten et al. 2013; Jessop, El Hakim and Gibbs 2014a; Birch, Batten and Batey 2016; Jessop & Tomas 2017; Pitt & Winstone 2018). Second, the three scales, *Appropriate Assessment*, *Deep Approach* and *Surface Approach* should be condensed into one factor, with its meaning clarified so that it investigates how students learn in a whole assessment environment, as well as incorporating authentic assessment (Meyers & Nulty 2009; Ashford-Rowe, Herrington and Brown 2013). Finally, a new factor should be established that investigates how students perceive formative assessment, in line with TESTA's main findings that modular degrees have led to high summative assessment diets across programmes, with weakly practiced formative (Jessop, El Hakim and Gibbs 2014a; Jessop & Tomas 2017; Wu & Jessop 2018).

This is an Accepted Manuscript of an article published by Taylor & Francis in ASSESSMENT AND EVALUATION IN HIGHER EDUCATION on 4 November 2018, available online: https://www.tandfonline.com/doi/full/10.1080/02602938.2018.1499867.

Although we have tested the validity of a questionnaire *in medias res* on the basis of its statistical validity, we also bring to the discussion a theoretical and experiential perspective, having worked for eight years with academics on hundreds of programmes in universities in the UK, Australia and India. Given that psychometric development of an instrument is an on-going process, further refinement of the factors and items seems necessary. Given theoretical developments in the field, the authors are convinced of the imperative to redevelop the AEQ.

Conclusion

Our study offers mixed conclusions about the statistical validity of AEQ 3.3. On the one hand, 80% of the items have good relationships with their associated factors. On the other, there are problems with items from two factors significantly overlapping, as well as some items misloading or cross-loading. Three factors are thin on items with only two items in each factor. These weaknesses point to the need for a revision. Besides, 12 years is a long time in the life of a questionnaire in a fast-moving and relatively new research field. This is especially the case as programme assessment, the assessment environment and the ecology of learning have attained greater prominence in the literature. On balance, our statistical exploration of the AEQ 3.3, together with developments in the field and its widespread use within the TESTA process, suggests that we need to take a fresh look at the questionnaire. This will enable users to obtain better data about how students are learning and studying within programme assessment environments. Without wishing to leave our readers on the edge of their seats, this fresh look will necessarily be the subject of our next article. In the meantime, researchers and practitioners using the AEQ 3.3 as a measure of conditions of learning from assessment should proceed with an element of caution.

Disclosure statement

This is an Accepted Manuscript of an article published by Taylor & Francis in ASSESSMENT AND EVALUATION IN HIGHER EDUCATION on 4 November 2018, available online: https://www.tandfonline.com/doi/full/10.1080/02602938.2018.1499867.

No potential conflict of interest was reported by the authors.

References

Arum, R., &J. Roksa. 2011. *Academically Adrift: Limited Learning on College Campuses.* Chicago. University of Chicago Press.

Ashford-Rowe, K., J. Herrington & C. Brown. 2013. "Establishing the critical elements that determine authentic assessment". *Assessment & Evaluation in Higher Education*. 39 (2) 205-222. DOI: 10.1080/02602938.2013.819566

Batten, J., Batey, J., Shafe, L., Gubby, L., & Birch, P.D.J. 2013. "The influence of reputation information on the assessment of undergraduate student work." *Assessment & Evaluation in Higher Education*, 38 (4) 417-435.

Berg, M. & Seeber, B. 2016. *The Slow Professor: Challenging the Culture of Speed in the Academy.* Toronto. University of Toronto Press.

Birch, P.D.J., Batten J., & Batey, J. 2016. "The influence of student gender on the assessment of undergraduate student work." *Assessment & Evaluation in Higher Education*, 41 (7) 1065-1080.

Black, P., & D. Wiliam. 1998. *Inside the Black Box: Raising Standards through Classroom Assessment*. London: Grenada Learning.

Bloxham, S., & P. Boyd 2007. *Developing Effective Assessment in Higher Education.* Berkshire. Open University Press.

Bloxham, S. 2016. Personal Correspondence. 5 March 2016.

Boud, D. 2000. "Sustainable Assessment: Rethinking Assessment for the Learning Society". *Studies in Continuing Education*. 22 (2): 151–167.

Boud, D., & N. Falchikov. 2006. "Aligning Assessment with Long-term Learning". *Assessment & Evaluation in Higher Educa*tion. 31 (4): 399–413.

Byrne, B.M. 2010. *Structural Equation Modelling with AMOS: Basic Concepts, Applications, and Programming* (2nd Ed.). New York. Routledge.

Chickering, A.W. & Gamson, Z.F. 1987. "Seven Principles for Good Practice in Undergraduate Education". *American Association for Higher Education Bulletin*. 3-7. March 1987. Available at: https://eric.ed.gov/?id=ED282491

Clark, L.A., & D. Watson 1995. "Constructing validity: Basic issues in objective scale development". *Psychological Assessment*. 7, 309-319.

Claxton, G. 1998. *Hare Brain, Tortoise Mind*. London: Fourth Estate.

Cousin, G. 2009. *Researching Learning in Higher Education: An Introduction to Contemporary Methods and Approaches*. Abingdon. Routledge.

Cumming, J., Clark, S.E., Ste-Marie, D.M., McCullagh, P., & C. Hall 2005. "The functions of observational learning questionnaire". *Psychology of Sport and Exercise*. 6. 517-537.

Comrey, A.L., & H.B. Lee 1992. *A First Course in Factor Analysis* (2nd Ed.). Mahwah, NJ. Lawrence Erlbaum Associates.

Fallowfield, J.L., Hale, B.J., & Wilkinson, D.M. 2005. *Using Statistics in Sport and Exercise Science Research.* Chichester: Lotus Publishing.

Fink, L.D. 2003. *Creating Significant Learning Experiences: An integrated approach to designing college courses.* San Francisco. Jossey-Bass.

Freire, P. 1970. *The Pedagogy of the Oppressed.* London. Continuum.

Garson, G.D. 2006. *Factor Analysis. Statnotes: Topics in Multivariate Analysis.* North Carolina St. University, Quantitative Research in Public Administration. Available at: http://www2.chass.ncus.edu/garson/pa765/factor.htm

Gibbs, G. & C. Simpson. 2004. Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education.* 1(1) 3–31.

Gibbs, G. 2006. How Assessment Frames Learning. In Bryan, C. and Clegg, K. (Eds). *Innovative Assessment in Higher Education*. (23-36). Abingdon. Routledge.

Gibbs, G. & H. Dunbar-Goddet. 2007. "The effects of programme assessment environments on student learning. York. Higher Education Academy.

Gibbs, G. & H. Dunbar-Goddet. 2009. "Characterising programme-level assessment environments that support learning". *Assessment & Evaluation in Higher Education.* 34 (4) 481–489.

Gibbs, G. & Y. El Hakim 2011. "Using the Assessment Experience Questionnaire to engage Course Teams in the revision of programme-level assessment regimes". Paper presented at Higher Education Surveys for Enhancement Conference 2011, The National College for School Leadership. 19 May 2011. Available at: https://www.heacademy.ac.uk/knowledge-hub/using-assessment-experience-questionnaire-engage-course-teams-revision-programme-level

Goodall, H. l. 2000. *Writing the New Ethnography*. Lanham, Maryland. AltaMira Press.

Hall, C., D. Mack, A. Paivio, & H. Hausenblas. 1998. “Imagery use by athletes: Development of the Sport Imagery Questionnaire”. *International Journal of Sport Psychology*. 29, 73-89.

Harland, T., A. McLean, R. Wass, E. Miller, & K. N. Sim. 2015. “An Assessment Arms Race and Its Fallout: High-stakes Grading and the Case for Slow Scholarship.” *Assessment & Evaluation in Higher Education*. 40 (4): 528–541.

Harland, T. & N. Wald. 2018. “Curriculum, teaching and powerful knowledge”. *Higher Education*. 1-14.

Hattie, J. & H. Timperley. 2007. “The Power of Feedback”. *Review of Educational Research*. 77:1. 81-112

Higgins, R. P. Hartley & A. Skelton 2002. “The conscientious consumer: Reconsidering the role of assessment feedback in student learning”. S*tudies in Higher Education*, 27 (1) 53-64.

Hutcheson, G., & N. Sofroniou. 1999. *The Multivariate Social Scientist*. London: Sage Publications Ltd.

Jessop, T., Y. El Hakim, & G. Gibbs. 2014a. “The Whole is Greater Than the Sum of Its Parts: A Large-scale Study of Students’ Learning in Response to Different Programme Assessment Patterns.” *Assessment & Evaluation in Higher Education* 39 (1): 73–88. doi:10.1080/02602938.2013.792108.

Jessop, T., Y. El Hakim, & G. Gibbs. 2014b. “TESTA 2014: A way of thinking about assessment and feedback. *Educational Developments*. 14:3.

Jessop, T., & B. Maleckar. 2016. “The Influence of Disciplinary Assessment Patterns on Student Learning: A Comparative Study.” *Studies in Higher Education* 41 (4): 696–711. doi:10.1080/03075079.2014.943170.

Jessop, T. 2017. "Inspiring Transformation through TESTA’s Programme Approach." In *Scaling up Assessment for Learning in Higher Education*, edited by D. Carless, S. Bridges, C.K.W. Chan, and R. Glofcheski, 49-64. Sydney: Springer.

Jessop, T., & C. Tomas. 2017. “The Implications of Programme Assessment Patterns for Student Learning.” *Assessment & Evaluation in Higher Education* 42 (6): 990–999. doi:10.1080/02602938.2016.1217501.

Innis, K. 1996. “Diary Survey: How undergraduate full-time students spend their time”. Leeds. Leeds Metropolitan University.

Kline, P. 2000. *The Handbook of Psychological Testing* (2nd Ed.). London. Routledge.

Knight, P., & Yorke, M. 2003. *Assessment, Learning and Employability.* Maidenhead: Open University Press.

This is an Accepted Manuscript of an article published by Taylor & Francis in ASSESSMENT AND EVALUATION IN HIGHER EDUCATION on 4 November 2018, available online: https://www.tandfonline.com/doi/full/10.1080/02602938.2018.1499867.

MacKenzie, S.B., P.M. Podsakoff & N.P. Podsakoff. 2011. “Construct measurement and validation procedures in MIS and behavioural research: Integrating new and existing techniques”. *MIS Quarterly*, 35, 293-334.

Marton, F., & R. Saljo. 1976. “On Qualitative Differences in Learning I: Outcome and Process.” *British Journal of Educational Psychology* 46: 4–11.

Meyer, J.H.F. & P. Parsons 1989. “Approaches to studying and course perceptions using the Lancaster inventory - A comparative study”. *Studies in Higher Education*.14:2, 137-153, DOI: 10.1080/03075078912331377456.

Meyers, N.M. & D. D. Nulty 2009. “How to use (five) curriculum design principles to align authentic learning environments, assessment, students’ approaches to thinking and learning outcomes”. *Assessment & Evaluation in Higher Education*. 34:5, 565-577, DOI:10.1080/02602930802226502

Nicol, D. J. 2010. "From Monologue to Dialogue: Improving Written Feedback Processes in Mass Higher Education." *Assessment & Evaluation in Higher Education* 35 (5):501–517.

Nicol, D. J., & D. MacFarlane-Dick. 2006. "Formative Assessment and Self-regulated Learning: A Model and Seven Principles of Good Feedback Practice." *Studies in Higher Education* 31 (2):99–218. doi: 10.1080/03075070600572090.

Nunnally, J.C., & I.H. Bernstein. 1994. *Psychometric Theory* (3rd Ed.). New York: McGraw Hill.

O'Donovan, B., M. Price, & C. Rust. 2008. “Developing student understanding of assessment standards: a nested hierarchy of approaches”. *Teaching in Higher Education*, 13 (2) 205-217.

Orr, S. 2007. “Assessment moderation: constructing the marks and constructing the students”. *Assessment & Evaluation in Higher Education*. 32 (6) 645-656.

Pitt, E. & N. Winstone 2018. “The impact of anonymous marking on students’ perceptions of fairness, feedback and relationships with lecturers”. *Assessment & Evaluation in Higher Education.* DOI: 10.1080/02602938.2018.1437594. Published online 9/2/18.

Ramsden, P. 2003. *Learning to Teach in Higher Education.* London. Routledge.

Sadler, D. R. 1989. "Formative Assessment and the Design of Instructional System." *Instructional Science* 18: 119-144.

Schmitt, N. 1996. “Uses and abuses of coefficient alpha”. *Psychological Assessment*, 8, 350-353.

Shay, S. 2005. “The assessment of complex tasks: A double reading”. *Studies in Higher Education*. 30(6). 663-679.

Stevens, J. 1996. *Applied Multivariate Statistics for the Social Sciences.* Mahwah, NJ: Lawrence Erlbaum.

Tabachnick, B.G, & L.S. Fidell. 2001. *Using Multivariate Statistics* (4th Ed.). MA: Allyn and Bacon.

Vaughn, B.K., H. Lee, & A. Kamata, A. 2012. Measurement basics, methods, and issues. In G. Tenenbaum, R.C. Eklund, & A. Kamata (Eds.), *Measurement in Sport and Exercise Psychology* (25-32). Champaign, IL: Human Kinetics.

Wass, R., T. Harland, A. McLean, E. Miller, & K.N. Sim. 2015. “Will press lever for food: behavioural conditioning of students through frequent high-stakes assessment”. *Higher Education Research and Development* 34 (6) 1324–1326.

Wheelahan, L. 2010. *Why Knowledge Matters in Curriculum: A Social Realist Argument.* Abingdon. Routledge.

Williams, S.E., & Cumming, J. 2011. “Measuring athlete imagery ability: The Sport Imagery Ability Questionnaire”. *Journal of Sport & Exercise Psychology,* 33, 416-440.

Wilson, K.L., A. Lizzio, & P. Ramsden 1997. “The development, validation and application of the Course Experience Questionnaire”. *Studies in Higher Education*. 22:1, 33-53, DOI: 10.1080/03075079712331381121

Wu, Q. & Jessop, T. 2018. “Formative assessment: missing in action in both research-intensive and teaching focused universities?” *Assessment & Evaluation in Higher Education*. Published online 15 January 2018.