Student Perceptions of Predictability of Examination Requirements and Relationship with Outcomes in High-Stakes Tests in Ireland

Jo-Anne Baird*, Daniel H. Caro & Therese N. Hopfenbeck

Oxford University Centre for Educational Assessment

ABSTRACT

Entirely predictable examinations are ones for which the questions are known in advance. Some assessments are designed this way, but in public examinations, predictability is subtler. Students familiarise themselves with the requirements broadly: likely topics that will come up, question formats and how to maximise their marks. If students can predict what they have to do, they can memorise performances, such as essays, and restrict their learning to fit only with examination requirements. The danger is that this focus could undermine curriculum aims. Further, examinations that are overly predictable might produce results that do not generalise to other performances or have predictive validity. This paper presents part of a broader project investigating whether the Higher Level Irish Leaving Certificate Examinations were too predictable. Here, the development of a rating scale for students' views of examination predictability is described. Data were collected from 1,002 Irish Leaving Certificate students taking Higher Level examinations in biology (n=536), English (n=749) and geography (n=387). Students' views on predictability of the examination could be grouped consistently across subject areas into three factors: valuable learning, predictability and narrowing of the curriculum. Belief that narrowing of the curriculum was a good examination preparation tactic had a negative relationship with examination scores and perceived learning value of examinations was positively associated with students' scores in biology and English. These findings indicate that the scoring system rewards students who believe they must study the discipline broadly.

Keywords: predictability, washback, student perceptions, examinations, structural equation modelling

*Corresponding author: jo-anne.baird@education.ox.ac.uk

Introduction

In many countries, there is concern about the effects of high-stakes testing upon the curriculum, teaching and learning (e.g Au, 2007; Berry, 2011; Frederiksen, 1984; Kirkpatrick, 2011; Mizutani, Rubie-Davies, Hattie & Philp, 2012, Kelly & Leavy 2013). A high-stakes test is one in which the results have a direct link to rewards or sanctions for students, their teachers or institutions (Madaus, 1988, p.29). The Higher Irish Leaving Certificate (LC) examinations, upon which this article focuses, are competitive, university-entrance examinations and are also used for selection for employment. Students experience a great deal of pressure in their preparation for the Leaving Certificate due to these high stakes (O'Shea, 1983). Students need points from the examinations for entry into higher education and tend to conflate the LC and the points system (Hyland, 2011). Points may be accumulated on up to six examination results. All subjects are equally weighted in the composite score, irrespective of the course being applied for, although 'bonus points' are currently being awarded for Higher Level Mathematics, on a trial basis since 2012.

In Ireland, current concerns about high-stakes testing are manifested mainly in allegations that the examinations are too 'predictable', This has also been a concern in other countries such as England (Ofqual, 2008). Predictability is not a well-defined technical assessment term. Nevertheless, it is no less serious a

concern than validity or fairness in the educational sphere in Ireland. We identified a previously published peer-reviewed article on the predictability of Leaving Certificate Accounting examinations which analysed exam papers and surveyed students' views on it. It was found that the exams were highly repetitive and predictable, motivating teachers to teach towards the examination. Since many of the questions were lacking variation in form and content, the researchers concluded that the examinations were encouraging rote learning (Byrne and Willis, 2001; Byrne & Willis, 2004:58). The broader literature on the effects of high-stakes tests upon teaching and learning and test preparation is highly relevant to, and underpins, the current work.

Three unpublished articles using the term predictability were identified, one of which set out various methods for investigating examination predictability (Murphy et al., 2012). GCSE and A-level examination papers were reviewed by subject matter assessment experts for predictability in an empirical study (Ofqual, 2008). It was concluded that formulaic wording of questions was beneficial in communicating the requirements to candidates. Syllabuses, however, were sometimes so detailed that examiners had little room to manoeuvre in setting question topics. This finding shows that predictability needs to be viewed more broadly than as an artefact of question papers, which influenced the design of our research programme.

We conducted expert reviews of the examination materials and concluded that the three subjects included in this study were not overly predictable, although there were some subject-specific issues of concern (Baird, Hopfenbeck, Elwood, Caro, & Ahmed, 2014). Predictability concerns were manifested differently in each subject. In biology, a greater emphasis upon scientific reasoning and less upon recipes for scientific methods was suggested by the expert reviewers as an improvement. Deeper engagement with the purposes of text and language used rather than so much emphasis upon personal responses (which sometimes lacked focus) were suggested in English. In geography, although again the subject was not viewed as problematically predictable overall, some topics came up too frequently. The third piece of unpublished work was an internal review of the Higher Irish Leaving Certificate conducted by the State Examinations Commission (SEC – the body responsible for the examinations). This report concluded that the choice of topics should be reviewed in some subjects, including English (SEC, 2012).

An element of examination predictability is necessary to ensure that a structured educational experience can be provided. High-stakes examinations in many countries are curriculum-related. Examination materials, including test papers, marking schemes and test-taking regulations are public documents.[1] Curriculum-test alignment should have the positive effect of

[1] http://www.examinations.ie/index.php?l=en&mc=en&sc=ep

driving instruction through the setting of clear goals (Frederiksen & Collins, 1989; Shepard, 1993). Making examination materials public removes secrecy and reduces social effects associated with proximity to the examiners, who are 'in the know'. Seen positively, predictability is transparency. Alignment between tests and the taught curriculum is 'measurement-driven instruction' in the US literature (Popham, 1987). However, the measurement-driven instruction movement has been undermined by the negative impacts of testing upon teaching and learning (e.g. Darling-Hammond, 2010; Madaus, Russell & Higgins, 2009), especially instrumental teaching to the test.

Distinct from the high-stakes testing arena is the 'Assessment for Learning' agenda, which seeks to build more valid assessments based upon classroom interaction (Black & Wiliam, 1998). From this literature, we know that sharing assessment criteria with students can have *positive* effects upon their outcomes because it enhances capacity for self-assessment (Jonsson, 2010; Kleinmann, Kuptsch & Köller, 1996; McDonald & Boud, 2003). However, research in vocational assessment settings suggested that too much transparency can undermine learner autonomy and call into question the validity of inferences made from the assessments (Torrance, 2007). There is no objective way to distinguish whether an assessment is transparent or overly predictable. Finding the right balance to deliver the curriculum objectives in a valid manner is key.

To find out whether examinations strike the right balance, we need to know stakeholders' views of the examinations, not just the views of examiners or testing agencies. How teachers and students prepare for the examinations, which is outside the realm of the examiner, is central to the relationship that students have with the examination. Of course, assessment design matters for predictability, but it is not the sole arbiter in publicly available examinations. Students might perhaps be taught or otherwise discern how to score marks on a test with pre-prepared responses even if the topics appear sufficiently distinctive between years to the examiners. For example, students could judge that there will always be a question about character development related to a Shakespeare character and prepare an essay on Richard II that could be adapted slightly (or not at all) to match whatever question is set. Readers of this article will be familiar with examination preparation techniques and might question whether this is problematic.

Organising and consolidating learning in preparation for an examination can have positive effects. Rehearsal can lead to better long-term retention of information (Craik & Lockhart, 1972). Making connections across the material helps students to form an understanding of the deep structure of a discipline and to 'chunk' packets of information. This is a foundation for building expertise (Gobet, 2005). Chunking of information frees working memory for higher order skills of analysis, evaluation and

problem solving (Ericsson & Kintsch, 1995). The foregoing assumes that revision builds upon deep learning, but examination predictability could allow students to shortcut the learning stage: cramming instead of forming deep understanding of the subject.

In Ireland, private tuition is referred to as 'grinds', after the Charles Dickens *Hard Times* headteacher, who is associated with a cold concern for facts and numbers. Only 15% of students in the current study had used grind colleges, but more students might have had private tuition on an individual basis. Smyth (2009) reported that 45% of students had received private tuition (including in grinds colleges and on an individual basis) in their last year of school and this had risen from 32% in 1994. Use of private tuition is not a direct measure of cramming behaviours, but is widely interpreted as an indicator of cramming in Ireland. Interestingly, private tuition did not significantly advantage students' grades once socio-economic effects, prior performance in examinations and attitudes to school were taken into account (Smyth, 2009).

This research set out to investigate students' views about how predictable examinations were and to establish whether there were relationships between these views and examination outcomes. Effects of predictability upon examination scores could be due to improvements in test-wiseness or domain knowledge. Predictability effects upon domain knowledge would be produced by use of the examination materials in combination with teaching.

Before turning to the aims of the current study, we introduce our definitions of levels of predictability. In ‘blatant predictability’, questions are known in advance, but no special training is associated with the pre-release. In ‘format familiarity’, the candidates have seen a version of the assessment previously and are re-tested. Finally, in ‘subtle predictability’ there is training on the examination requirements, though the exact questions are not known in advance. In this complex version of predictability, students know the assessment format well and are trained on it.

Blatant predictability: prior knowledge of the questions

Disclosure of essay topics had no significant effect in an experiment conducted with 300 prospective graduate students (Powers & Fowles, 1998). Neither were significant effects upon performance found in security breaches in an objective structured clinical examination (Wilkinson, Fontain & Egan, 2003). Reiter, Salvatori, Renfeld, Trinh & Eva (2006) investigated the effects of releasing the questions for a multiple mini-interview assessment for student admissions to a medical course. Over three studies, they found no significant impact upon judgments of applicants’ competencies. Surprisingly, knowing the questions in advance did not improve students’ results in these studies.

Format predictability: test-wiseness

Test-wiseness involves students knowing what is expected from them in the examination. Findings on re-testing are relevant to the current research because they provide some evidence on the size and type of gains when test format is experienced by students for a second time. A number of studies have been conducted on psychological tests. A meta-analysis of re-testing on a range of cognitive ability measures to investigate practice effects found a small Cohen's *d* effect size of .26 (Hausknecht, Halpert, Di Paolo & Moriarty Gerrard, 2007). Re-testing gains reduced with subsequent opportunities, but were higher with up to six retests, and people with higher cognitive ability benefitted more from re-testing (Kulik et al., 1984; Arendasy & Somer, 2013). Analytical and quantitative tasks show higher re-testing effects than verbal tasks (Hausknecht et al., 2007). Recent research indicates that re-testing effects are domain specific, as opposed to showing an increase in general intelligence, or *g* (Arendasy & Somer, 2013).

Subtle predictability: teaching to the test

Re-testing with training has been shown to impact additionally upon intelligence test scores (Freund & Holling, 2011). Anastasi (1981) distinguished three kinds of training in advance of tests: 1) a broad education, 2) test-wiseness and 3) drilling on test content. Anastasi (1981) argued that training types 1 *and* 2 increased validity, as promoting test-wiseness reduced construct-irrelevant variation amongst students. Drilling on test content

(type 3) undermines validity because students over-perform on the test in relation to their knowledge and skills. Thus, scores when students have been drilled on test content might not generalise so well to other contexts and predictive validity is reduced. Distinguishing drilling on test content from Anastasi's other two categories can admittedly be subtle, but there is an implication that students may not understand the discipline deeply or broadly, or be able to apply or evaluate their understandings due to teaching to the test (Volante, 2004).

The current research

Examination questions are not released in advance for the Higher Irish Leaving Certificate, but students are highly familiar with the test content and there is concern about teaching to the test (subtle predictability). This research looks at three subject examinations: biology, English and geography. We defined a problematically predictable examination as one in which teachers and students can anticipate the test-taking conditions, performances required, question formats, topics and scoring to the extent that undesirable effects upon the educational process are pervasive. Note that this goes beyond question spotting because knowing the kind of responses that are given credit might ultimately be more important than predicting the questions. Knowledge of the kind of answer that will be credited, the extent to which the question needs to be addressed specifically and the

likelihood of pre-prepared, formulaic responses being given credit all relate to predictability of examinations. The underlying concern about 'predictable' examinations is that they overly reward recall of knowledge at the expense of higher order thinking skills. Undesirable effects of overly predictable examinations include narrowing of the curriculum, superficial rote learning, drilling on test content and failure to develop a broad and deep understanding of a discipline. This research focuses upon alleged subtle predictability of the LC examinations.

Systemic validity involves testing that brings about curricular, instructional and learning strategy changes that foster the cognitive traits that the tests are designed to measure (Frederiksen & Collins, 1989, p.27). Overly predictable examinations lack systemic validity because they measure test preparation narrowly rather than the intended broader assessment objectives.

In this article, we report part of a large-scale evaluation of high-stakes, school-leaving examinations in Ireland. We measured students' views on the predictability of the examinations and related these to their examination outcomes.

Aims

We investigated student perceptions of the predictability of the LC examinations in English, biology and geography. First, we

constructed scales that measure perceptions of the examinations. We then related the scale ratings to the scores.

This research was part of a broader programme on *Predictability in the Irish Leaving Certificate* (Baird et al., 2014) that included analysis of the examination materials, interviews with teachers and students (Elwood et al., 2015) and collected data on a number of variables on the questionnaire not reported here. It is important to note that students had very high levels of test-wiseness, with over 86% in each discipline indicating that the examination format had been explained to them, over 77% responding that the marking criteria had been explained and over 94% had been given past papers (also available on the internet). As mentioned above, subject matter experts concluded that these three examinations were not overly predictable (Baird et al., 2014). Although subject experts' views provide a context for the current research, it is ultimately through students that any impact of examination predictability upon learning outcomes is produced. In summary, this research is on high-stakes examinations in which students have a great deal of test-wiseness and, though highly transparent, the examinations were not generally seen to be overly predictable by independent subject experts from another country. Our research questions were:

1. What are the views of students on the predictability of the examinations and how do they vary by subject area?

2. Can views on predictability be summarised and grouped into meaningful factors reflecting underlying perceptions of the examinations?
3. How do different views on predictability relate to examination outcomes?
4. Do students who believe the examinations were predictable attain higher scores?

METHOD

The Irish Leaving Certificate Examinations

The LC examination marks the end of upper secondary education in Ireland. Students normally spend two years studying for the examination. Syllabuses are produced by the National Council for Curriculum and Assessment (NCCA), and the State Examinations Commission (SEC) is responsible for the examinations. In total, 52,767 students sat the LC in 2013. The examination entrants represented well over 90% of the age cohort. Students usually sit exams in 7 or 8 subjects each year. The median age of students sitting the examinations is 18.

Examinations are available at two levels and, due to their high-stakes nature, this research focused upon the higher level LC examinations. Assessments consist of end-of-course examination papers lasting approximately three hours, with 400 maximum raw scores. Two question papers are set in the English exam. Questions are short answer and extended answer in biology and

geography exams, but all questions are extended answer in English. A practical investigation is assessed separately in geography and is worth 20% of the total final score. Scoring is conducted independently from the schools.

Participants

One hundred schools (with 7,204 higher level LC entrants) were selected randomly to participate in the survey, with students in schools participating voluntarily. The population of students who sat at least one of the higher level LC examinations in English, biology or geography was 5,578. The survey was administered in online and paper and pencil formats.

A total of 1,002 responses were collected (147 online and 855 paper and pencil surveys). Of the total student respondents, 806 gave their permission to link their exam scores to the survey. The sample was biased towards girls (N=473; 58.7%) and higher performing students (see Table 1). Results must therefore be interpreted in that context.

Table 1: Cumulative percentage at each grade: questionnaire sample compared with population (Pop)

	Biology		English		Geography	
Grade	Sample	Pop	Sample	Pop	Sample	Pop
A	22.3	14.4	14.4	9.7	15.4	8.7
B	55.5	43.6	43.6	36.4	51.6	38.1
C	77.8	81.9	81.9	76.1	85.6	75.3
D	94.3	99.1	99.1	98.3	99.1	97.2
E	99.2	99.9	99.9	99.9	100	99.8
F	100	99.9	99.9	100	100	100
NG	100	100	100	100	100	100
No.	449	23,436	624	33,279	312	19,762

Questionnaire

A survey methodology was used to gain a broad view of students' perceptions of the examinations. Background variables were collected: gender, parents' education, and information on socio-economic status (SES). Views on predictability, use of learning strategies when preparing for the exams and information about support in preparation for the exams were also collected.

The questionnaire was piloted in a cognitive interview with two former Irish higher education students who had sat the LC. In this method, the researcher asks the participants to a) read the question in a survey, b) explain what it means, c) read the answer option and choose the answer which best fits the respondent's perception, and d) explain the reason for the answer (Karabenick et al, 2007). It is a think-aloud technique.

Questionnaires were printed and sent to school examinations officers by the State Examinations Commission, with an accompanying letter explaining the importance of the study. The paper and pencil version of the questionnaire was offered to students immediately after they had sat their final LC exam.

An online version of the survey was also made available and posters with information about the online version were displayed in the examination halls. Thus, students could choose between a paper or online survey after they had finished their exams; both were offered in English or in Irish. Twenty-three students

answered Irish versions. All open responses were translated from Irish to English.

The first page of the questionnaire gave information about the purpose of the study and confidentiality of responses. Students were asked to write their exam number and for permission to link their exam score to the survey results. Students were also informed of a prize draw for five Apple iPads if they completed the survey. Access to the prize draw was not contingent upon permission to link to examination scores, but the questionnaire had to be completed for the student to be entered for the draw.

Analysis

Variables

The following variables are used in the analysis:

Examination scores. Data on student grades provided by the SEC were transformed into examination scores using the Central Applications Office (CAO) scheme.[2] Scores range from 0 to 100. For each grade (A1–D3) points are awarded to the applicant. Score points range from 0 to 100. A grade of A1 is awarded 100 points, a grade of A2 is awarded 90 points, a grade of B1 is awarded 85 points, and so on down to the lowest score of 45 for a grade of D3. No points are awarded for a grade below D3.

[2] http://www2.cao.ie/app_scoring/lcegrid.htm

The Student Experience of Exam Predictability Scale. Ten Likert-scale items (i.e., 'strongly disagree', 'disagree', 'agree', and 'strongly agree') assessing perceptions of students on the examination (see Table A1). These items are based upon the literature on the effects of high-stakes tests and preparation for them. This research was exploratory and no specific factor structure was anticipated; thus the number of items was not balanced across theoretically-driven factors. The original scale included eleven items but previous analysis indicated that one item had inconsistent behaviour across subject areas and was therefore removed (Caro & Hopfenbeck, 2014): 'I can do well in this exam even if I do not fully understand the topics'. Also, one item of this scale was part of a different question relating to preparation for the exam rather than experience of the examination (item *i*); thus it was presented in a different section of the questionnaire and had different answer categories ('almost never', 'now and then', 'often', and 'always').

Anastasi's (1981) three forms of preparation for high-stakes tests were drawn upon in the design of the items. Thus, items relating to a broad education (*a* to *e* and deleted item), test-wiseness (*f* to *h*) and a narrow focus (*i* and *j*) were devised in relation to predictability (Table A1). More specifically, item *a* was inspired by the work of Wigfield and Eccles (2000) and the theory of achievement motivation and items measuring the *value* of subjects (Eklof, 2007). Items *b*, *d* and *e* related to surface and

deep learning strategies (Biggs, 2003; Diseth & Martinsen, 2003; Entwistle & Entwistle, 1991; Entwistle, 1988; Ramsden, 1988; Shepard et al., 2005). Item *c* was derived from research findings regarding student dissatisfaction with the kind of learning assessed in high-stakes examinations (Daly et al., 2012). Items *f* to *h* were associated with prediction of the assessment requirements (Ofqual, 2008). Finally, items *i* and *j* concerned narrowing of the curriculum studied (Au, 2007; Madaus, Russell and Higgins, 2009).

Family SES. Dichotomous data on home possessions (i.e. TV, car, dishwasher, room of your own, quiet place to study, computer for school work, Internet access, iPad or other tablet, Smartphone, mobile phone, PlayStation/X-box/Wii, classic literature, and dictionary) and ordinal data on parental education and number of books were summarised into a single family SES scale using the partial credit model (Caro & Hopfenbeck, 2014).

Analytical strategy

We conducted exploratory factor analysis (EFA) to assess the underlying structure of the proposed *Student Experience of Exam Predictability Scale* and performed regression analysis to evaluate the association between the resulting EFA factors and examination scores.

Analyses were conducted in R using package *psych* for EFA (Revelle, 2014) and *lavaan* for regression analysis (Rosseel, 2012). EFA was performed with polychoric correlations rather than Pearson correlations to handle the ordinal categorical data of Likert-scale items and using varimax rotation to simplify interpretability of orthogonal factors and maximise variance explained by each factor. Regressions employed the full information maximum likelihood (FIML) method. FIML performs similarly to multiple imputation techniques for accounting for missing data uncertainty (Collins, Schafer, Kam, 2001).

Models controlled for family SES to counteract sample selection bias, since the analytic sample was selected towards higher SES students. Inasmuch as views on predictability were related to family SES, the SES control also evaluated whether associations with predictability views are independent of or actually driven by family SES characteristics.

Analytic sample

The analytic sample is concerned only with those students who sat the higher level LC examination. Samples sizes varied by subject area and for the EFA and regression models. The EFA of the *Student Experience of Exam Predictability Scale* included students with complete information in the scale items: 749 for the English analysis, 536 for biology and 387 for geography. Regression

models relating the *Student Experience of Exam Predictability Scale* to examination results included a smaller sample due to missing values in examination scores: 603 students for the English analysis, 430 for biology, and 301 for geography. These figures represented over 10% of students who sat each subject in the schools sampled, which was 1-2% of the national population of students taking these subjects.

FINDINGS

Student Experience of Exam Predictability Scale

Table A1 reports the distribution of responses on the *Student Experience of Exam Predictability Scale* by discipline. Interestingly, 73% of the students reported (item *a*: agree/strongly agree) that they believed that they would be able to use what they had learned for their exam in the future in biology, whilst only 37% believed the same about English and 57% in geography. In other words, there seems to be more positive beliefs about the relevance of the biology exam compared to the other disciplines. A considerable number of students reported that they predicted the exam questions well (item *f*: agree/strongly agree), but this varied by subject: 70% in English, 50% in geography and 32% in biology. In biology, 90% of students agreed with the statement 'To do well in this exam, I need a broad understanding of the subject, across many topics' (item *d*). This is, again, the highest

reported agreement with the item between the three disciplines, indicating that the biology exam was viewed as less predictable, examining a broad understanding and there was a perception that the learning would be useful for the future. Overall, students considered that the examinations required a broad and deep understanding of the subjects and application of knowledge.

Table 2 reports results of EFA. Factor loadings are reported separately for the sample of students taking the English, biology, and geography LC examinations. Only loadings larger than 0.3 are shown. Eigenvalues and the proportion of variance explained by each factor are reported at the bottom of Table 2.

Table 2. EFA results (standardised factor loadings)

	English (n= 749)			Biology (n= 536)			Geography (n= 387)		
F1: Valuable learning factor	F1	F2	F3	F1	F2	F3	F1	F2	F3
a) I think I will be able to use what I learned for this exam in the future	0.65			0.57			0.50		
b) To do well in this exam I need to think and adapt what I know	0.55			0.60			0.61		
c) The exam tests the right kind of learning	0.54			0.61			0.49		
d) To do well in this exam, I need a broad understanding of the subject, across many topics	0.47						0.41		
e) To do well in this exam, remembering is more important than understanding	-0.42			-0.39			-0.39		
F2: Predictability factor									
f) I predicted the exam questions well		0.66			0.64			0.63	
g) I felt I knew what the examiners wanted this year		0.54			0.62			0.54	
h) I was surprised by the questions on the exam this year		-0.51			-0.65			-0.62	
F3: Narrowing of the curriculum factor									
i) I chose not to study some topics as I thought they would not come up			0.70			0.61			0.49
j) I left a lot of topics out of my revision and still think I will do well			0.68			0.82			0.99
Eigenvalues	2.40	1.17	1.91	2.25	2.09	1.15	1.26	2.01	2.08
Proportion variance explained	0.16	0.10	0.11	0.14	0.13	0.12	0.13	0.12	0.13

Note. Loadings larger than 0.30 are reported.

The EFA solution consistently produced three interpretable factors across subject areas. Eigenvalues were larger than one only for the first three factors. Items grouped consistently into three interpretable factors with loadings larger than 0.3 and no cross-loadings. Altogether, these three factors accounted for around 38% of the variance in the item data in each subject (see Table 2).

Item indicators for the first factor are: a) ‘I think I will be able to use what I learned for this exam in the future’; b) ‘To do well in this exam I need to think and adapt what I know’; c) ‘The exam tests the right kind of learning’; d) ‘To do well in this exam, I need a broad understanding of the subject, across many topics’; and e) ‘To do well in this exam, remembering is more important than understanding’. These items are all related in that they seem to assess the extent to which students thought the learning experience of preparing for the examination was valuable. Except for biology where the loading of item d) was 0.18 (<0.30), these items consistently loaded on the first factor across subjects. We will thus refer to this factor as *valuable learning*. Note that item e) asks about the importance of memorisation and therefore loads negatively. The contribution of this factor to the total variance tended to be the largest across subjects: 16% in English, 14% in biology, and 13% in geography (see Table 2).

Item indicators for the second factor are: f) ‘I predicted the exam questions well’; g) ‘I felt I knew what the examiners wanted this year’; and h) ‘I was surprised by the questions in the exam this year’. These items reflect the extent to which students reported that they were able to predict questions in the exam. Accordingly, the factor has been labelled *predictability*. Item f) ‘I predicted the exam questions well’ has the highest loading across disciplines.

The third factor consists of only two items: i) ‘I chose not to study some topics as I thought they would not come up’ and j) ‘I left a lot of topics out of my revision and still think I will do well’. These items reflect whether students decided to narrow the curriculum while preparing for the exam. It has been named *narrowing the curriculum*.

Predictability and examination results

Table 3 reports regression results of examination scores on the scores of the three resulting EFA factors whilst controlling for family SES. Model 1 includes the EFA factors and Model 2 the family SES control additionally. Standardised coefficients, R-squares, and sample sizes are reported. Sample sizes are smaller than for the measurement model because of missing values in examination scores: not all students granted permission to use their examination results.

Table 3. Regression of examination scores (standardised coefficients)

	English (n= 603)		Biology (n= 430)		Geography (n= 301)	
	(1)	(2)	(1)	(2)	(1)	(2)
Valuable learning factor	0.08 *	0.07	0.13 *	0.13 *	-0.08	-0.05
Predictability factor	0.07	0.04	0.12 *	0.12 *	0.15 *	0.15 *
Narrowing of the curriculum factor	-0.14 *	-0.13 *	-0.20 *	-0.20 *	-0.08	-0.06
Family SES		0.22 *		0.26 *		0.29 *
R-squared	0.03	0.08	0.08	0.14	0.04	0.12

The narrowing of the curriculum factor is negatively related to the examination scores in English, biology, and geography. Coefficients of the narrowing of the curriculum factor are negative and significant in English and biology, even after controlling for family SES. That is, students who believed that they could narrow the curriculum in preparing for the examination performed worse in the exam, especially in biology where the association is strongest. Of the three factors, narrowing of the curriculum is the strongest predictor of examination scores.

The valuable learning factor is positively and significantly related to the exam scores in English and biology. In biology, the association with valuable learning remains significant even after controlling for family SES. But no evidence of a significant positive association with valuable learning was found for the geography examination.

The predictability factor is positively related to the examination scores in English, biology and geography. Associations with the predictability factor are statistically significant in biology and geography, even after controlling for family SES.

R-squared values are in general very small for Model 1 including the three EFA factors (English=0.03, biology=0.08, and geography=0.04). Explained variance increases when SES is included in Model 2 from 0.03 to 0.08 in English, 0.08 to 0.14 in biology, and 0.04 to 0.12 in geography. That is, although in some cases significant, the predictability scales explain little variance in examination scores. Comparatively, family SES explains a larger proportion of the variability in examination scores. Its inclusion as a control variable is thus justified. Overall, however, models explain little variance in examination scores, but the purpose of the analysis has been to evaluate associations with the predictability scales rather than to maximise the model's explanatory power.

The association with family SES is consistently positive across the three subjects. That is, students of higher SES families tended to perform better in the examination compared to students of lower SES families. Main results of associations with EFA factors do not change substantially after SES is controlled.

DISCUSSION

We take each of our four research questions in turn. First, perceptions of predictability differed across disciplines; with English examinations being viewed as more predictable. However, to the extent that English differentiates by performance rather than task, this is not necessarily problematic. Most students in

each discipline considered that a broad and deep understanding of the curriculum was needed to do well. However, in keeping with Daly et al. (2012), the majority of students considered that the examinations did not test the right kind of learning. Interviews with students in this research programme indicated that students felt that examinations relied too much upon memory, such that those with the best understanding of a discipline did not necessarily get the highest grades (Baird, et al., 2014). In English, only one third of students considered that they would be able to use what they had learned in the future. One possible explanation for this finding is that cultural knowledge from the language arts does not have an obvious vocational application. Also, as with many other national language examinations, some of the texts studied might not have had an immediate relevance for all students in modern, urban Ireland (e.g. *Antigone, Wuthering Heights, Macbeth, The Great Gatsby*).

Second, students' views on predictability of the examinations could be grouped into three categories: valuable learning, predictability and narrowing of the curriculum. The valuable learning category reflected views that the learning experience of preparing for the examination was valuable. The predictability category was consistent with views that it was possible to predict exam questions. The narrowing of the curriculum category captured whether students decided to narrow the curriculum while preparing for the exam.

Third, students who believed narrowing of the curriculum was a good tactic tended to perform worse in the English and biology examinations. This result was not explained by the student's SES, because narrowing the curriculum was related to poorer performance in examinations for both students from higher and lower socio-economic backgrounds. This is the first empirical study to show a negative impact upon outcomes of students' perceptions that the curriculum can be narrowed in preparation for an examination. Additionally, this study showed that students who perceived that preparing for the examination was a valuable learning experience tended to score higher in the English and biology examinations.

Fourth, students who believed examinations were predictable tended to perform better in the biology and geography examinations, but not in English. One difference between the disciplines included in this research is the extent to which there is emphasis upon knowledge and skills, with English emphasising skills more than the other two disciplines. Another difference is that differentiation between the weakest and strongest students in biology and geography is by task, with the strongest students being able to succeed on the most difficult tasks. In English, differentiation is by performance, where students are set the same tasks and the degree of success depends upon the response. In more factually-based disciplines where differentiation is by task, understanding the kinds of questions that are likely to come up

might be more important. So, even though the English examination was perceived to be more predictable than in the other two disciplines, perceived predictability did not advantage students in English. Essay-based subjects might not be so prone to effects of prior exposure (Powers & Fowles, 1998).

Limitations and further research

This research is limited to three LC examinations in Ireland in 2013 and to the sample of students who participated in the study. Reported tests of statistical significance with this sample presuppose an underlying population to which results can be generalised. Although we might anticipate generalisation of the scale's utility in other contexts, we expect that the relationships between the scale and scores to be contingent upon the discipline and the design of the examinations.

A critical limitation is that students' prior attainment is not controlled for in associations between predictability views and examination results. And it is likely that these views are affected by prior attainment. For instance, lower performing students might tend to narrow the curriculum when preparing for examinations and value less the learning experiences from examinations. If that is the case, the negative association with narrowing the curriculum views and positive association with valuable learning views could be explained by prior attainment and not by predictability views. Prior attainment data was not

available for students in our sample, but regression models controlled for family SES. Because family SES has consistently been found to be positively associated with educational attainment (Sirin, 2005), the SES control reduces bias due to omitted prior attainment in regressions. However, the possibility that results are explained by prior attainment cannot be ruled out and remains a limitation of the study.

Another limitation is the number of items per latent factor. The valuable learning factor consists of five items, the predictability factor of three items and the narrowing of the curriculum of only two items. It would be important to develop and validate a larger number of items in future studies, particularly for the narrowing of the curriculum and predictability factors. Yet another limitation is that the sample is biased towards higher-achieving students and it is possible that less able students could change the relationship found between the sub-scale on beliefs of predictability and outcomes; for example. R-squared values are very small, but models do not attempt to reveal the data generating process for examination scores. Rather, the goal was to evaluate whether EFA factors were related to examination scores significantly. From this perspective, the analysis is correlational and does not provide evidence of causation.

Further research should evaluate measurement invariance of these scales across subject areas. That is, whether the same

constructs can be measured validly with these items in different subject areas. Such results will contribute to the generalisability of this scale to other examination contexts in which there is concern about the negative washback effects of predictable examinations.

CONCLUSIONS

At the level of the examinations, these findings indicated that the scoring system credits students who believe they must study the discipline broadly. This reflects well upon the LC. We also see that thinking that you understand which questions are likely to come up generally produces a trend for higher scores in biology and geography. This is likely to reflect test-wiseness. There was some evidence that student views that the learning was valuable mattered for examination outcomes; however, not for all subjects. Ideally, an examination should show positive, significant relationships between perceived value of learning and outcomes.

This study adds to the literature on the effects of high-stakes tests, as it was shown that beliefs that the intended curriculum could be narrowed in preparation for the examination are related to poorer performance in examinations. In the case of the Irish LC examinations, students and teachers should note that overly narrowing the curriculum could have negative consequences for examination outcomes. There are, however, serious limitations of this study (e.g. lack of prior attainment

control) that prevent us from making causal statements regarding the consequences of narrowing the curriculum, as weaker students and teachers might use narrowing the curriculum strategies more often. Clearly, further research using longitudinal data or randomised experiments is required to investigate the effects of narrowing the curriculum strategies on examination results and the mechanisms at work. Previous research, however, is in line with our findings and shows that teachers in England who used question-spotting strategies to prepare students for examinations tended to have worse school results (Greatorex & Malacova, 2006).

The findings can further inform teachers about the importance of facilitating a learning environment where students are encouraged to approach the subjects broadly, since narrowing the preparation and reading for exams does not have any positive influence on the exam scores.

Acknowledgments

This project was commissioned by the State Examinations Commission in Ireland. We are grateful to them for their support of the research and for the participation of the teachers and learners in Ireland in this project.

REFERENCES

Anastasi, A. (1981). Coaching, test sophistication and developed abilities. *American Psychologist*, 36, 10, 1086 – 1093.

Arendasy, M. & Somer, M. (2013). Quantitative differences in retest effects across different methods used to construct alternate test forms. *Intelligence*, 41, 3, 181 – 192.

Au, W. (2007). High-stakes testing and curricular control: a qualitative metasynthesis. *Educational Researcher*, 36, 5, 258 – 267.

Baird, J, Hopfenbeck, T., Elwood, J., Caro, D., & Ahmed, A. (2014). *Predictability in the Irish Leaving Certificate*. Oxford University Centre for Educational Assessment Report. OUCEA/14/1. Retrieved from http://oucea.education.ox.ac.uk/research/recent-research-projects/investigation-into-the-predictability-of-the-irish-leaving-certificate-examinations/

Berry, R. (2011). Assessment trends in Hong Kong: seeking to establish formative assessment in an examination culture. *Assessment in Education: principles, policy & practice*, 18, 2, 199 – 211.

Black, P & Wiliam, D (1998). Assessment and classroom learning. *Assessment in Education: principles, policy and practice*, 5(1), 7–75.

Biggs, J.B. (2003). *Teaching for quality learning at university*. 2nd Edition. Buckingham: Open University Press. Society for Research into Higher Education.

Byrne, M. & Willis, P. (2001) The Revised Second Level Accounting Syllabus: A New Beginning or Old Habits Retained, *The Irish Accounting Review*, 8(2), 1-22.

Byrne, M. & Willis, P. (2004) Leaving Certificate accounting: Measuring students' perceptions with the course experience questionnaire, *Irish Educational Studies*, 23(1), 49-64.

Caro, D. & Hopfenbeck, T. (2014). *Predictability in the Irish Leaving Certificate Examination. Working Paper 3: Student Questionnaire*. Retrieved from http://oucea.education.ox.ac.uk/research/recent-research-projects/investigation-into-the-predictability-of-the-irish-leaving-certificate-examinations/

Collins, L. M., J. L. Schafer, Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351.

Craik, F.I.M. & Lockhart, R.S. (1972). Levels of processing: a framework for memory research.

Daly, A., Baird, J., Chamberlain, S. & Meadows, M. (2012) Assessment reform: students' and teachers' responses to the introduction of stretch and challenge at A-level. *The Curriculum Journal*, 23, 2, 139 – 155.

Darling-Hammond, L (2010). *The Flat World and Education: how America's commitment to equity will determine our future*. New York: Teachers College Press.

Diseth, Å. & Martinsen, Ø. (2003). Approaches to learning, cognitive style, and motives as predictors of academic achievement. *Educational Psychology, 23*(2), 195-207.

Elwood, J., Hopfenbeck, T. and Baird, J. (2015). Predictability in high-stakes examinations: students' perspectives on a perennial assessment dilemma. *Research Papers in Education*, advance online publication. DOI:10.1080/02671522.2015.1086015.

Entwistle, N. (1988). Motivational factors in students' approaches to learning. In R. Schmeck (Editor), *Learning Strategies and Learning Styles*. New York, US: Plenum Press, 386 pages.

Entwistle, N. & Entwistle, A. (1991). Contrasting forms of understanding for degree examinations: the student experience and its implications. *Higher Education, 22*(3), 205 – 227.

Ericsson, K.A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211 – 245.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193–202.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Freund, P. A., & Holling, H. (2011) How to Get Really Smart: Modeling Retest and Training Effects in Ability Testing using

Computer-Generated Figural Matrix Items, *Intelligence, 39*(4), p233-243

Gobet, F. (2005). Chunking models of expertise: implications for education. *Applied Cognitive Psychology, 19*(2), 183 – 204.

Greatorex, J. and Malacova, E. (2006) Can different teaching strategies or methods of preparing pupils lead to greater improvements from GCSE to A Level performance? *Research Papers in Education, 21*(3), 255-291.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373-385.

Hyland, A (2011). *Entry to Higher Education in Ireland in the 21st Century*. Retrieved from http://www.transition.ie/files/Entry_to_Higher_Education_in_Ireland_in_the_21st_Century%20.pdf.

Jonsson, A. (2010). The use of transparency in the 'Interactive examination' for student teachers. *Assessment in Education: principles, policy & practice, 17*(2), 183 – 197.

Karabenick, S.A., Woolley, M.E., Friedel, J.M., Ammon, B.V., Blazevski, J., Bonney, C.R., de Groot, E., Gilbert, C., Musu, L., Kempler, T.M. & Kelly, K.L. (2007). Cognitive processing of self-report items in educational research: do they think what we mean? *Educational Psychologist, 42*(3), 139 – 151.

Kelly, A. E. & A. Leavy (2013). The design space of student learning: who is accountable and accountable for what? *Irish Educational Studies*, 32: 1, 1-6.

Kirkpatrick, R. (2011). The negative influences of exam-oriented education on Chinese high school students: backwash from classroom to child. *Language Testing in Asia, 1*(3), 36 – 45.

Kleinmann, M., Kuptsch, C. & Köller, O. (1996). Transparency: a necessary requirement for the construct validity of assessment centres. *Applied Psychology: An International Review, 45*(1), 67 – 84.

Kulik, J.A., Bangert-Drowns, R.L., & Kulik, C-L. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*(2), 179 – 188.

Madaus, G.G. (1988). The distortion of teaching and testing: High - stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29-46. DOI: 10.1080/01619568809538611

Madaus, G F, Russell, M K, & Higgins, J (2009). *The Paradoxes of High Stakes Testing: how they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age Publishing Inc

McDonald, B. & Boud, D. (2003). The impact of self-assessment on acheivement: the effects of self-assessment training on performance in external examinations. *Assessment in Education: principles, policy & practice, 10*(2), 209 – 220.

Mizutani, S., Rubie-Davies, C., Hattie, J. & Philp, J. (2012). Do beliefs about NCEA and its washback effects vary depending on subject? *New Zealand Journal of Educational Studies, 46*(2), 47 – 59.

Murphy, R, Stobart, G, Baird, J & Winkley, J (2012) *Investigating the predictability of GCSE examinations*. Pearson Internal Report.

Ofqual (2008). *Predictability studies report on GCSE and GCE level examinations*. Office of the Qualifications and Examinations Regulator. http://dera.ioe.ac.uk/id/eprint/9242

O'Shea, M. (1983) A study of leaving certificate students' perceptions of teachers' attitudes to leaving certificate, *Irish Educational Studies, 3*(2), 256 – 272.

Popham, J. (1987). Two-plus decades of educational objectives. *International Journal of Educational Research, 11*(1), 31–41.

Powers, D.E. and Fowles, M.E. (1998). Effects of preexamination disclosure of essay topics. *Applied Measurement in Education, 11*(2), 139 – 157.

Ramsden, P. (1988). *Improving Learning: New perspectives*. New York: Nichols Publishing Company.

Reiter, H.I., Salvatori, P., Rosenfeld, J., Trinh, K. & Eva, K.W. (2006). The effect of defined violations of test security on admissions outcomes using multiple mini-interviews. *Medical Education, 40*(1), 36 – 42.

Revelle, W. (2014) *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psych Version = 1.4.8.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36.

State Examinations Commission (2012*). Draft Report of the SEC Working Group on Predictability in the Leaving Certificate Examination*. Unpublished.

Shepard, L (1993). The place of testing reform in educational reform: a reply to CiZek. *Educational Researcher, 22*, 10–14

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F., Baratz Snowden, J., Gordon, E., Gutierrez, C., et al. (2005). Assessment. In L. Darling-Hammond and J. Bransford (Editors), *Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do* (275 – 321). San Fransisco: John Wiley & Sons, Inc.

Sirin, S.R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453.

Smyth, E (2009) Buying your way into college? Private tuition and the transition to higher education in Ireland. *Oxford Review of Education, 35*(1), 1–22

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in

post-secondary education and training can come to dominate learning. *Assessment in Education: principles, policy & practice, 14*(3), 281 – 294.

Volante, L. (2004). Teaching to the test: what every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy*, 35. Retrieved from http://files.eric.ed.gov/fulltext/EJ848235.pdf.

Wigfield, A., & Eccles, J. S. (2000) Expectancy-value theory of achievement motivation. *Contemporary Education Psychology*, *25*, 68-81.

Wilkinson, T.J., Fontaine, S. & Egan, T. (2003). Was a breach of examination security unfair in an objective structured clinical examination? A critical incident. *Medical Teacher*, *25*(1), 42 – 46.

Table A1. Students' perceptions about the examination
(percentages)

Items	English (n= 749)				Biology (n= 536)				Geography (n= 387)			
	Strongly disagree	Disagree	Agree	Strongly agree	Strongly disagree	Disagree	Agree	Strongly agree	Strongly disagree	Disagree	Agree	Strongly agree
Valuable learning factor												
a) I think I will be able to use what I learned for this exam in the future	38%	26%	27%	10%	12%	14%	40%	33%	18%	25%	45%	12%
b) To do well in this exam I need to think and adapt what I know	4%	13%	47%	36%	6%	21%	43%	30%	5%	13%	50%	32%
c) The exam tests the right kind of learning	33%	32%	29%	6%	29%	25%	35%	11%	30%	28%	32%	11%
d) To do well in this exam, I need a broad understanding of the subject, across many topics	7%	23%	45%	25%	2%	8%	36%	54%	4%	11%	40%	46%
e) To do well in this exam, remembering is more important than understanding	20%	33%	22%	25%	16%	27%	24%	32%	12%	25%	28%	35%
Predictability factor												
f) I predicted the exam questions well	7%	24%	47%	23%	24%	44%	23%	9%	14%	36%	35%	15%
g) I felt I knew what the examiners wanted this year	7%	29%	52%	11%	16%	36%	39%	9%	6%	33%	44%	16%
h) I was surprised by the questions on the exam this year	18%	50%	21%	11%	7%	20%	35%	38%	10%	42%	31%	18%
Narrowing of the curriculum factor												
i) I chose not to study some topics as I thought they would not come up	28%	27%	26%	19%	41%	29%	17%	14%	23%	31%	23%	24%
j) I left a lot of topics out of my revision and still think I will do well	20%	41%	32%	7%	31%	39%	24%	6%	20%	35%	35%	11%

Note. Item *i* is part of a different question with different response labels ('almost never', 'now and then', 'often' and 'always').