GCSEs and AS/A levels are currently being reformed, with many of the new reformed qualifications available for certification from June 2017. One of the main changes being introduced is the return to linear assessments. As a result, the JCQ has been recently working towards a common approach for how to calculate estimated marks in the new linear qualifications. Alternative methods such as the ones looked at in this research (e.g., z-score, percentile and regression imputation) have been considered in a variety of different research projects carried out by the different UK awarding bodies. The outcomes from such research did not show an outstanding method, but rather very small differences between them (this research shows just a marginal preference for regression imputation, with the performance of the z-score and percentile methods very similar). As the majority of the UK awarding bodies already use the z-score method for unitised specifications, it was agreed by the JCQ that it should be used for the new linear specifications from 2017 onwards.

References

Cheung, C.P. (2009). *Investigating different methodologies for calculating missing marks – examples using data from GCE AS new specifications (Economics and French)*. Internal Report. Cambridge: Oxford, Cambridge and RSA.

Espinoza, J. (2015). A-level results: exam boards 'guesstimating' students' grades. *The Telegraph*. (2015, August 01). Retrieved from: http://www.telegraph.co.uk/education/secondaryeducation/11777405/A-level-results-exam-boards-guesstimating-students-grades.html

JCQ (2016). *Access Arrangements and Reasonable Adjustments 2016–2017*. London: Joint Council for Qualifications. Available online at: https://examining.jcq.org.uk/exams-office/access-arrangements-and-special-consideration/regulations-and-guidance/access-arrangements-and-reasonable-adjustments-2016-2017

Linning, S. (2015). British exam boards forced to 'estimate' hundreds of A-level results each year as students' papers are 'lost in the post'. *Mail Online*. (2015, August 01). Retrieved from: http://www.dailymail.co.uk/news/article-3182362/British-exam-boards-forced-estimate-hundreds-level-results-year-students-papers-lost-post.html

On the reliability of applying educational taxonomies

Victoria Coleman Research Division

Introduction

Educational taxonomies are classification schemes which organise thinking skills according to their level of complexity. They provide a unifying framework alongside common terminology that can be used by educationalists. Primarily, most educational taxonomies focus upon thinking skills that fall within the cognitive domain, although some have also included other domains. Educational taxonomies can have a variety of different applications (Marzano, 2001). First, they can be used to analyse existing educational materials such as learning objectives, curriculum plans, lessons and assessments to ascertain which levels of thinking skills they encompass. Secondly, it is possible to use them as a framework when designing educational materials to ensure that the desired cognitive levels are targeted. They can also be adapted to form an assessment tool themselves, for example, by asking markers whether students have exhibited the required level of thinking during assessment activities. Finally, they can be used to ascertain whether corresponding curriculum objectives and assessment materials align, or whether there is a mismatch in the thinking levels that they are targeting. This can be done both in the context of designing new educational materials and in analysing pre-existing ones. Educational taxonomies can be applied in this way to a broad variety of educational contexts, being adapted according to the specific topic under investigation. This literature review will outline research investigating educational taxonomies and their use in terms of reliability.

Reliability

When discussing educational taxonomies and their application, it is important to consider reliability, as the value of different educational taxonomies is somewhat impacted by reliability constraints. There are various different types of reliability. In the subsequent literature review both inter- and intra-rater reliability are discussed[1]. The amount of consideration given to the rater reliability of educational taxonomies is highly variable and studies specifically investigating it in this context are sparse. There are three broad categories of techniques for assessing rater reliability: consensus estimates, consistency estimates, and measurement estimates (Stemler & Tsai, 2008). Within these there are a number of statistical methods that can be used in order to calculate reliability, and the technique which is selected depends on a number of factors such as the type of reliability being assessed and the nature of the data (McHugh, 2012; Stemler & Tsai, 2008). For example, a greater level of inherent dissimilarity between the items being categorised within a taxonomy will tend to lead to higher values of correlation-based reliability coefficients. However, if nearly all items being assessed fall into one or two categories, then there is little distinction between items and so less chance for raters to display a high correlation between their judgements. Conversely, if nearly all items are within a single category, simple measures (e.g., the percentage of times raters agree with one another) will appear high, as even random placement will result in a high level of agreement.

Within educational taxonomy research there has been a great deal of variation in the statistical measures used and in how the resulting reliability statistics have been interpreted. It must be noted that many of the research articles reviewed did not provide full details of the method used to calculate reliability, which limits our interpretation to some extent. Table 1 summarises the methods used by the studies in this review to calculate reliability.

1. Rater reliability is frequently referred to using different terms in the literature including coder, assessor and judge in place of rater, and terms such as consistency and agreement in place of reliability. However, for the purposes of this review, the term inter-rater reliability will be used.

Table 1: Reliability measures used by studies discussed in this article

Method	*Definition*	*Type of Reliability*	*Interpretation*
Percentage Agreement (PA)	Percentage agreement is a measure which is calculated as the percentage of times that two raters (or possibly groups of raters) gave the same rating.	Inter- and intra-rater reliability	≥70% acceptable reliability
Percentage Universal Agreement (PUA)	A variation on PA that measures the percentage of times that all raters gave the same rating.	Inter-rater reliability	≥70% acceptable reliability
Percentage Majority Agreement (PMA)	A variation on PA that measures the percentage of times for which the majority of raters gave the same rating.	Inter-rater reliability	≥70% acceptable reliability
Percentage of Partial Agreement (PPA)	Very similar to PMA. This is the percentage of instances in which there was partial agreement, defined as instances where 50% or more of the raters agreed. In this case exactly half of raters agreeing is counted as a positive outcome whereas for PMA it is not.	Inter-rater reliability	≥70% acceptable reliability
Kappa	Kappa coefficients include both Cohen's and Fleiss' kappa. Cohen's kappa can be used to assess the degree of consensus between two raters and whether it is above the level of agreement that would be expected to arise by chance alone. Fleiss' kappa is a similar measure which can be used when there are more than two raters.	Inter- and intra-rater reliability	Cohen's kappa (Cohen, 1960) < 0 poor agreement; 0.01–0.20 slight; 0.21–0.40 fair; 0.41–0.60 moderate; 0.61–0.80 substantial; 0.81–1.00 almost perfect (Landis & Koch, 1977). Fleiss's kappa <0.40 poor; 0.40–0.75 fair to good; >0.75 as excellent. (Fleiss, 1981).
Krippendorff"s alpha	Krippendorff's alpha can be used when calculating reliability for multiple raters with multiple possible ratings.	Inter-rater reliability	$\alpha \geq .800$ is good. $\alpha \geq .667$ is the lower limit for acceptable agreement (Krippendorff, 2004).
Correlations	Standard correlation coefficients such as Pearson's r and Spearman's Rho measure the association between two independent raters (or in some instances two groups of raters). In essence, they do not require that raters agree precisely on ratings, only that they place items in a similar rank order.	Inter- and intra-rater reliability	Values greater than 0.70 are typically considered acceptable levels of inter-rater reliability (Stemler & Tsai, 2008).
Cronbach's alpha	Estimates the expected correlation between the sum of scores across all raters and the (hypothetical) sum of scores across another group of raters of the same size.	Inter-rater reliability	Values greater than 0.70 are typically considered acceptable levels of inter-rater reliability (Stemler & Tsai, 2008).
Intraclass correlation coefficients (ICC)	ICCs attempt to overcome the limitations of other consistency estimates by taking into account both consistency and agreement of ratings. Essentially they measure the percentage of the variance across all ratings that is attributable to which item is being assessed (rather than which rater is doing the assessing).	Inter-rater reliability	Values greater than 0.70 are typically considered acceptable levels of inter-rater reliability (Stemler & Tsai, 2008).

Summary of the taxonomies mentioned in this review

This article presents an overview of reliability findings reported across a number of studies applying educational taxonomies. The concept of educational taxonomies was first introduced in 1956 with Bloom's Taxonomy of Educational Objectives, which provides a comprehensive system for classifying levels of thinking. It includes six categories of cognition, which are presented in a hierarchy of increasing complexity: knowledge, comprehension, application, analysis, synthesis and evaluation. Each category includes several subcategories. Bloom's taxonomy is also accompanied with examples of test items that belong to the different categories. Since its original introduction, it has been adapted and refined most notably by Anderson and Krathwohl (2001), whose revised taxonomy is widely used. Whilst the six cognitive categories remain in the revised version, some have been relabelled and it has moved away from the idea of a cumulative hierarchy and instead evolved into a two-dimensional framework with four knowledge categories added. Whilst a large number of alternative taxonomies have been developed, the vast majority of studies discussed in this review utilised Bloom's taxonomy, or adaptations of it (see Moseley et al., 2004 for a summary and review of educational taxonomies). Besides Bloom, there are numerous other educational taxonomies – those where research studies considering their reliability were found are also discussed and so are briefly outlined.

Another taxonomy that has been developed is the Structure of Observed Learning Outcomes (SOLO) taxonomy by Biggs and Collis (1982). Based on the stages outlined in Piaget's theory of cognitive development (1950), it includes five categories of understanding in a hierarchy of increasing complexity: prestructural, unistructural, multistructural, relational and extended abstract. An adaption of the SOLO taxonomy has divided the multistructural and relational categories into three subcategories each, resulting in nine SOLO levels in total (Burnett, 1999; Chan, Tsui, Chan, & Hong, 2002). A reflective thinking instrument developed by Kember (1999) can be used as an educational taxonomy to assess students' reflection and critical thinking skills. It is divided into two categories: non-reflective and reflective thinking. Non-reflective thinking is divided into habitual action and thoughtful action whilst reflective thinking is divided into reflection and critical reflection.

One of the studies in this review used Porter's taxonomy (Porter & Smithson, 2001a, 2001b). This was designed to enable standards and assessments in Mathematics and Science to be assessed. It includes three dimensions: topics, expectations of students' performance, and the modes of presentation. Each contains several subcategories. For example, the dimension of topics is a list of content areas within Mathematics and Science, with no hierarchical progression from one to the next. The final taxonomy considered is that of Marzano and Kendall (2006). This taxonomy comprises two dimensions: knowledge and mental processing. There are three knowledge domains: information, mental procedures and psychomotor procedures, with no hierarchy between these domains. Within the dimension of mental processing there are three hierarchical systems grouped into six levels. At the top is the self system, followed by the metacognitive system, with both comprising one level in the hierarchy. Following this is the cognitive system which is made up of four hierarchical levels: knowledge utilisation, analysis, comprehension and retrieval.

Literature examining the reliability of educational taxonomies

We found twenty-one studies[2] which considered reliability in the use of educational taxonomies – they are summarised in Table 2. These studies have utilised various educational taxonomies in a range of contexts and subject areas as well as employing several different measures of reliability.

The majority of studies in this review found evidence of moderate to high reliability when using educational taxonomies. The main exceptions to this were the research by Näsström (2009) and Karpen and Welch (2016). Näsström (2009) highlighted potential problems in both inter-

2. The articles were found through a Google Scholar literature search, with a list of established taxonomies searched alongside terms such as 'reliability', 'rater reliability' and 'rater consistency'. The reference lists of the initial papers that were found were then searched in order to find further relevant studies.

Table 2: Summary table of research studies examining the reliability of educational taxonomies

Study	*What was assessed*		*Which taxonomy*	*Raters*	*Inter-rater reliability (method in parenthesis)*	*Intra-rater reliability (time gap is in **bold**)*
Chan et al. (2002)	Term paper reports from 17 students		Modified nine category SOLO taxonomy Bloom's taxonomy Kember's Reflective Thinking Measurement Model (RTMM)	2 trained raters	Modified SOLO 0.60 (correlation between raters) Bloom r = 0.93 (correlation between raters) RTMM r = 0.87 (correlation between raters)	n/a
	Responses of 11 students to case study problems		As above – but with original 5 category SOLO taxonomy	2 trained raters	SOLO 0.66 (correlation between raters) Bloom's 0.68 (correlation between raters) RTMM 0.082 (correlation between raters)[a]	n/a
Crowe, Dirks, and Wenderoth (2008)	500 Life Science questions		Blooming Biology Tool–rubric based on adaptation of Bloom's taxonomy	3 raters	At least 2/3 agreed on 91% of the questions	n/a
	51 Life Science questions			36 students	98%(PMA) Additionally >80% agreed on 31/51 questions	n/a
Ebadi and Shahbazian (2015)	49 Iranian high school final exam questions		Bloom's taxonomy	2 panels of 2 researchers	0.87 (Cronbach's alpha)	**1 month later** 0.94 (Cronbach's alpha[b] for the first panel)
Edwards (2010)	Physics and Chemistry curriculum content and corresponding assessment papers		Revised Bloom's taxonomy	2 raters	0.97 for Physics curriculum objectives 0.98 for Chemistry curriculum objectives 0.88 for Physics assessment papers 0.92 for Chemistry assessment papers (method not specified)	n/a
Ewing, Foster, and Whittington (2011)	Classroom session in agricultural college	Professor discourse	Florida Taxonomy of Cognitive Behaviour – adaptation from Bloom's taxonomy	First rater: researcher Second rater: expert in cognition research	0.94 (method not specified)	**9 weeks later** 0.91 (method not specified
	Videotapes used for second rater and intra-rater reliability	Each professor question that elicited student engagement	Bloom's taxonomy	First rater: researcher Second rater not specified	0.93 (method unclear in article)	**3 weeks later** 0.84 (method not specified)
		Questions asked by students	Bloom's taxonomy	Second rater not specified	0.90 (method unclear in article)	**3 weeks later** 0.88 (method not specified)

a. The researchers stated 'the inter-rater reliability for Study 2 (the one which applied the modified version of SOLO with sub-levels) was higher than that of Study 1' (Chan et al., 2002, p.515). They suggested that this indicates that adding sub-levels increased inter-rater reliability. However, this seems to either be a misinterpretation or a reporting error as the modified SOLO was actually stated as being used in Study 1 and the unmodified SOLO showed the higher inter-rater reliability.

b. This does not particularly make sense as a measure of intra-rater reliability (in effect estimates the correlation of the sum of both measures with hypothetical set of two separate measures by the same individual). However, it is what was recorded by the author.

Table 2: Summary table of research studies examining the reliability of educational taxonomies *(continued)*

Study	*What was assessed*	*Which taxonomy*	*Raters*	*Inter-rater reliability (method in parenthesis)*	*Intra-rater reliability (time gap is in **bold**)*
	Course objectives	Bloom's taxonomy	First rater: researcher Second rater: expert in writing course objectives and cognition	0.98 (method not specified)	**3 weeks later** 0.92 (method not specified)
FitzPatrick and Schulz (2015)	165 educational outcomes statements and 182 corresponding statements from 2 units of Science curriculums from 4 jurisdictions	Revised Bloom's taxonomy	2 raters (An additional rater who was not included in reliability analysis)	80.4% for the outcomes (PA) 81.4% for the assessments (PA)	n/a
Karpen and Welch (2016)	Six questions from a teacher resource website, each targeted at a specific level of Bloom's taxonomy	Bloom's taxonomy	21 Pharmacy faculty members	0.25 (Krippendorff's alpha)	n/a
Leung (2000)	Responses from 79 students on an open ended DT item	SOLO taxonomy	1 researcher and 1 DT teacher, pre-marking meeting was held	0.49 (correlation between raters)	**Unknown time gap** 0.71 (correlation between the researchers marking and remark)
Mizbani and Chalak (2017)	57 speaking and listening activities from Iranian EFL Textbook Prospect 3	Bloom's revised taxonomy	Second rater for inter-rater reliability	0.92 (PA) on 14 of the activities	**2 weeks later** 0.98(PA) on random selection of 30 activities
Näsström and Henriksson (2008)	102 Swedish Chemistry upper secondary standards 58 assessment questions for upper secondary Chemistry	Bloom's revised taxonomy Porter's taxonomy (excluding modes of presentation domain)	2 raters	*Standards* 0.45 for Bloom's taxonomy (kappa) 0.07 for Porter's taxonomy (kappa) *Assessments* 0.36 for Bloom's taxonomy (kappa) 0.30 for Porter's taxonomy (kappa)	n/a
Näsström (2009)	35 Mathematics objectives for upper secondary schools in Sweden	Bloom's revised taxonomy	Panel of 4 assessment experts	26% (PUA, first occasion) 14% (PUA, second occasion) 46% (PMA, both occasions) 0.47 (kappa, first occasion) 0.41 (kappa, second occasion)	**2 to 3 months later** 51% (Average PA per judge) 12% SD 0.43 (kappa)
			Panel of 4 teachers	3% (PUA, first occasion) 11% (PUA, second occasion) 29% (PMA, both occasions) 0.15 (kappa, first occasion) 0.24 (kappa, second occasion)	**2 to 3 months later** 25% (Average PA per judge) 7% SD 0.18 (kappa)
Palmer and Devitt (2007)	33 MEQ's and 50 MCQ's from clinical undergraduate programme	Bloom's taxonomy	2 raters who then discussed and agreed a final rating	0.7 and 0.8 (kappa between each rater and the final agreed level MEQs) 0.7 and 0.8 (kappa, between each rater and the final agreed level MCQs)	n/a
Parham, Chinn, and Stevenson (2009)	84 statements in 24 transcripts from students' verbalisation when solving a Computer Science problem	Bloom's taxonomy	3 raters	89% (method not specified)	n/a
Plack et al. (2007)	308 reflective writing journal entries of medical students. These were assessed in terms of the highest level of cognitive processing that was displayed.	Three-level modified version of Bloom's taxonomy	3 raters	0.52–0.58 (kappa between pairs) 0.79 (ICC)	n/a
Razmjoo and Kazempourfard (2012)	One unit from Interchange EFL textbooks	Bloom's taxonomy	4 PhD student raters	0.972 (correlation of average rating of the PhD students with the researcher's rating)	**3 weeks later** 0.979 (correlation of average rating across all judges)
Rezaee and Golshan (2016)	41 questions in nationwide English exams in Iran	Bloom's taxonomy	2 raters	0.87 (correlation between two raters)	**Unknown time gap** 0.96 (1 rater's correlation with previous ratings)

Table 2: Summary table of research studies examining the reliability of educational taxonomies *(continued)*

Study	*What was assessed*	*Which taxonomy*	*Raters*	*Inter-rater reliability (method in parenthesis)*	*Intra-rater reliability (time gap is in **bold**)*
Riazi and Mosalendejad (2010)	Curriculum from 4 Iranian high school EFL textbooks	Bloom's taxonomy	No information provided	0.91 (method not specified)	**Unknown time gap** 0.98 (method not specified)
Teodorescu, Bennhold, Feldman, and Medsker (2013)	80 assessment questions from Physics textbooks	Physics adaptation of Marzano and Kendall's taxonomy	2 panels of 3 raters; each including 1 graduate student and 2 professors	0.75–0.85 (kappa between pairs on first panel) 0.70–0.82 (kappa between pairs on second panel)	**10 months later** 0.70–0.92 (kappa, between individuals on second panel)
Valcke, De Wever, Zhu, and Deed (2009)	282 messages as part of a collaborative learning group discussion task in Mathematics	Bloom's taxonomy	2 raters	0.95 (kappa)	n/a
van Hoeij, Hararhuis, Wierstra, and van Beukelen (2004)	179 short essay questions from 2 modules of a Veterinary course	Bloom's taxonomy	5 subject matter experts on first module	16% (PUA) and 57% (PPA) 34–57% (PA between each pair) <0.4 (kappa, between each pair)	n/a
			4 subject matter experts on second module	44% (PUA) and 28% (PPA) 61–77% (PA between each pair) 0.28–0.60 (kappa, between each pair)	n/a
			3 non-subject matter experts	*Module 1* 42% (PUA) and 49% (PPA) 50–71% (between pairs) <0.4 (kappa, between pairs) *Module 2* 50% (PUA) and 48% (PPA) 66–67% (between pairs) 0.55 (kappa, between pairs)	n/a
			All of the panels	*Inter-Group Reliability (non-experts vs experts)* Calculated 'modal taxonomic level' of each item for each panel group *Module 1* 65% (PA) 0.43 (kappa) *Module 2* 73% (PA) 0.63 (kappa)	n/a
Zheng, Lawhorn, Lumley, and Freeman (2008)	585 Biology exam questions (from Advanced Placement, undergraduate course, the MCAT, Graduate Record Examination and first year medical courses)	Bloom's taxonomy	3 experts in Biology education	0.53 (kappa) 0.68 (ICC)	n/a

and intra-rater reliability when using Bloom's taxonomy to assess the cognitive level of educational objectives, with none of the reliability findings showing more than moderate agreement. The findings are strengthened by the use of several different reliability measures, which have found consistent results. In particular, they found that across all of the measures, the teachers demonstrated lower reliability than the experts, which suggests there are differences as a function of the composition of the rating panels. In terms of differences between the two groups, the lower inter- and intra-rater reliability for the teachers may be related to the fact that they utilised the categories in Bloom's revised taxonomy to a greater extent and multi-categorised (categorising a single educational objective into multiple cognitive levels) to a lesser extent compared to the experts. This study involved the panels having group discussions about Bloom's taxonomy with examples presented to them before commencing their ratings. This is interesting given that training and practice was highlighted by many of the researchers as a potential avenue to improve reliability in the application of educational taxonomies. That said, conclusions about the impact of training and practice cannot be drawn from this study given that it was not examined experimentally.

Karpen and Welch (2016) also found low reliability when asking a panel of 21 faculty members to classify 6 exam questions. The researchers highlighted how this has implications for the use of Bloom's taxonomy and suggested that training of staff could be used to improve inter-rater reliability. That low reliability was found, when these questions had been purposefully written as examples of questions at specific levels in Bloom's taxonomy, is particularly concerning and perhaps also highlights challenges in using taxonomies to write questions at specific cognitive levels. It should also be noted that this study used a much greater number of raters than the other studies in this review. This therefore potentially raises the question as to whether the number of raters used impacts upon inter-rater reliability. That said, only six questions were rated and so this limits how much can be inferred from these results more generally.

Where reliability was investigated using standardised metrics such as kappa, alpha, correlation or ICC, the results tended to indicate acceptable to high reliability. However, our ability to draw conclusions about the reliability of taxonomies more generally is greatly limited by the fact that many of the studies reviewed did not specify the method that had been used to calculate reliability. Nevertheless, the high values reported by the majority of these studies do indicate a good level of reliability regardless of what measures were used to calculate them. Although, with the exception of Näsström (2009), all of the studies using PA (or variations such as PMA) found evidence of moderate to high reliability, these measures are limited in that they do not indicate how much agreement we could expect to find by chance alone. Consequently, given that there was a great deal of variability in the reported reliability found by studies using this measure, these findings must be interpreted with some caution when contributing to our overall conclusions about the reliability of educational taxonomies.

Overall, the majority of the studies provide evidence of moderate to good reliability when using educational taxonomies. In terms of inter- and intra-rater reliability, all of them considered inter-rater reliability with nine also examining intra-rater reliability. All of the studies examining intra-rater reliability found high reliability, with the exception of Näsström (2009). Whilst there were more measures of inter-rater reliability, these findings were more variable across the different research studies.

The majority of studies in this review used Bloom's taxonomy or adaptations of it, with just four including other taxonomies (Chan et al., 2002; Leung, 2000; Näsström & Henriksson, 2008; Teodorescu et al., 2013). Whilst this is unsurprising given the influence of Bloom in the field of education, the extent to which findings about the reliability of Bloom's taxonomy can be generalised to taxonomies more broadly is unclear. Consequently, other taxonomies would benefit from research being conducted to establish their reliability.

Areas for improving reliability

It is also useful to consider the way in which inter- and intra-rater reliability can be improved. The aforementioned studies have highlighted factors that influence reliability and which therefore offer a potential avenue for improvement.

Training and practice

The impact of training and practice on both inter- and intra-rater reliability was considered in a number of the research studies although none specifically examined their impact on reliability. Many of the studies either included some form of training (Chan et al., 2002; Näsström, 2009), or highlighted it as a potentially useful strategy for boosting reliability in future research (Karpen & Welch, 2016; Plack et al., 2007; van Hoeij et al., 2004). Training can be provided so as to ensure that raters are familiar with the taxonomy; it can also involve raters being given the opportunity to practise applying the taxonomy to sample materials, and having a group discussion to come to a consensus about how to interpret and apply the levels. Reliability can also be improved through providing examples of learning objectives of assessments that would fit into each taxonomic level. Some studies have demonstrated how a rubric with specific examples relevant to the topic area can be provided and used as a tool to support the application of educational taxonomies to both assessing and designing educational materials (Crowe et al., 2008; Lee, 2010). Whilst for some taxonomies, such as Bloom's, verb lists have been created to guide practitioners when applying them; evidence suggests that there is a great deal of variation in which verbs are aligned to specific levels and that individuals may interpret the meaning of different verbs at different levels (Stanny, 2016). Therefore it may be beneficial to include a group discussion as part of training, so that raters are able to develop a consensus in their interpretation and application of educational taxonomies. Thus, training and practice are potential ways in which reliability can be increased. As part of this, it is important to ensure that educational taxonomies and examples are clearly worded so as to reduce ambiguity and enhance reliability.

Rater variables

Characteristics of the rater also emerged as another factor which can impact upon rates of reliability, with differences found between experts and non-expert raters (e.g., Näsström, 2009). The number of raters used may influence inter-rater reliability, and there may be an optimum number of raters for achieving sufficient reliability whilst being practical in terms of constraints such as costs, time and finding sufficient number of raters, particularly where expertise is required. The number of raters also interacts significantly with characteristics of the raters, as the characteristics of the additional raters will impact upon the homogeneity of the group, with a homogenous group perhaps likely to show greater inter-rater reliability.

Taxonomy variables

The number of categories within an educational taxonomy was suggested as a potential factor impacting upon reliability by Chan et al. (2002), who suggested that adding sublevels to the SOLO taxonomy could increase inter-rater reliability by reducing ambiguity. Whilst it appears that their conclusion that subcategories increased reliability may be incorrect and based on a misinterpretation of the data (see footnote a), it would be useful for further research to investigate this and see if the number of categories and subcategories used impacts upon reliability. Although, of course, any conclusions on this matter would be hugely dependent upon the way in which reliability is defined.

Conclusion

Whilst very few studies were found which specifically examined the reliability of educational taxonomies, many studies did examine reliability to some extent. Although it is not possible to directly compare and summarise the findings across the studies, due to the different measures for assessing reliability, the majority of the studies discussed have provided evidence of at least moderate reliability, with evidence of poor reliability found only in a small number of instances – although of course studies showing poor reliability are less likely to get published. Inter-rater reliability has been looked at to a greater extent than intra-rater reliability. In the few studies that did consider intra-rater reliability, all but one found evidence of high intra-rater reliability. That said, many of these studies provided insufficient information about how reliability was calculated, such as failing to include information regarding which measure of reliability was used, and the time that elapsed between coding sessions. Consequently, the meaning and quality of the findings produced is sometimes unclear. Furthermore, this inconsistency in measurement places limitations on how far it is possible to compare reliability findings of different studies. Finally, whilst it seems that high reliability can be achieved using Bloom's taxonomy, and it can be hypothesised that high reliability can also be achieved when using other taxonomies, especially if

appropriate training and materials are used, there is insufficient research evidence to support or refute this hypothesis. In order to prove that research using other educational taxonomies can provide a sound evidence base for qualifications evaluation, comparability and development, further targeted studies will be necessary.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives*. New York: Longman.

Biggs, J., & Collis, K. F. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. New York.

Burnett, P. C. (1999). Assessing the structure of learning outcomes from counselling using the SOLO taxonomy: an exploratory study. *British Journal of Guidance & Counselling*, *27*(4), 567–580. Available online at: doi:10.1080/03069889908256291

Chan, C. C., Tsui, M. S., Chan, M. Y. C., & Hong, J. H. (2002). Applying the Structure of the Observed Learning Outcomes (SOLO) Taxonomy on Student's Learning Outcomes: An empirical study. *Assessment & Evaluation in Higher Education*, *27*(6), 511–527. Available online at: doi:10.1080/0260293022000020282

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297–334.

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's Taxonomy to Enhance Student Learning in Biology. *CBE-Life Sciences Education*, *7*(4), 368–381. Available online at: doi:10.1187/cbe.08-05-0024

Ebadi, S., & Shahbazian, F. (2015). Exploring the Cognitive Level of Final Exams in Iranian High Schools: Focusing on Bloom's Taxonomy. *Journal of Applied Linguistics and Language Research*, *2*(4), 1–11. Available online at: http://jallr.com/index.php/JALLR/article/view/58

Edwards, N. (2010). An analysis of the alignment of the Grade 12 Physical Sciences examination and the core curriculum in South Africa. *South African Journal of Education*, *30*(4), 571. Available online at: http://www.scielo.org.za/scielo.php?pid=S0256-01002010000400005&script=sci_arttext&tlng=en

Ewing, J. C., Foster, D. D., & Whittington, M. S. (2011). Explaining Student Cognition during Class Sessions in the Context Piaget's Theory of Cognitive Development. *NACTA Journal*, *55*(1), 68–75. Available online at: http://www.jstor.org/stable/pdf/nactajournal.55.1.68.pdf?seq=1#page_scan_tab_contents

FitzPatrick, B., & Schulz, H. (2015). Do Curriculum Outcomes and Assessment Activities in Science Encourage Higher Order Thinking? *Canadian Journal of Science, Mathematics and Technology Education*, *15*(2), 136–154. Available online at: doi:10.1080/14926156.2015.1014074

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Karpen, S. C., & Welch, A. C. (2016). Assessing the inter-rater reliability and accuracy of pharmacy faculty's Bloom's taxonomy classifications. *Currents in Pharmacy Teaching and Learning*, *8*(6), 885–888. Available online at: doi:10.1016/j.cptl.2016.08.003

Kember, D. (1999). Determining the level of reflective thinking from students' written journals using a coding scheme based on the work of Mezirow. *International journal of lifelong education*, *18*(1), 18–30. Available online at: doi:10.1080/026013799293928

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *International Biometric Society*, *33*(1), 159–174. Available online at: doi:10.2307/2529310

Lee, H. A. (2010). *Thinking Levels in Christian Publishers' Elementary Reading Textbook Questions*. (Doctor of Education), Columbia International University.

Leung, C. F. (2000). Assessment for Learning: Using Solo Taxonomy to Measure Design Performance of Design & Technology Students. *International Journal of Technology and Design Education*, *10*(2), 149–161. Available online at: doi:10.1023/a:1008937007674

Marzano, R. J. (2001). *Designing a New Taxonomy of Educational Objectives*. California, USA: Corwin Press, Inc.

Marzano, R. J., & Kendall, J. S. (2006). *The New Taxonomy of Educational Objectives*. Thousand Oaks, CA: Corwin Press.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. Available online at: http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=132393

Mizbani, M., & Chalak, A. (2017). Analyzing Listening and Speaking Activities of Iranian EFL Textbook Prospect 3 Through Bloom's Revised Taxonomy. *Advances in Language and Literary Studies*, *8*(3). Available online at: https://journals.aiac.org.au/index.php/alls/article/view/3527

Moseley, D., Baumfield, V., Higgins, S., Lin, M., Miller, J., Newton, D., & Gregson, M. (2004). *Thinking Skill Frameworks for Post-16 Learners: An Evaluation. A Research Report for the Learning and Skills Research Centre*. Retrieved from Regent Arcade House, 19–25 Argyll Street, London: http://files.eric.ed.gov/fulltext/ED508442.pdf

Näsström, G. (2009). Interpretation of standards with Bloom's revised taxonomy: a comparison of teachers and assessment experts. *International Journal of Research & Method in Education*, *32*(1), 39–51. Available online at: doi:10.1080/17437270902749262

Näsström, G., & Henriksson, W. (2008). Alignment of standards and assessment: A theoretical and empirical study of methods for alignment. *Electronic Journal of Research in Educational Psychology*, *6*(3), 667–690. Available online at: http://repositorio.ual.es/bitstream/handle/10835/565/Art_16_216_eng.pdf?sequence=1

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education*, *11*(11), 324. Available online at: doi:10.1186/1472-6920-7-49

Parham, J., Chinn, D., & Stevenson, D. E. (2009). *Using a Bloom's Taxonomy to Code Verbal Protocols of Students Solving a Data Structure Problem*. Paper presented at the 47th Annual Southeast Regional Conference, New York, USA.

Piaget, J. (1950). *The Psychology of Intelligence*. London: Routledge & Kegan Paul.

Plack, M. M., Driscoll, M., Marquez, M., Cuppernull, L., Maring, J., & Greenberg, L. (2007). Assessing Reflective Writing on a Pediatric Clerkship by Using a Modified Bloom's Taxonomy. *Ambulatory Pediatrics*, *7*(4), 285–291. Available online at: doi:http://dx.doi.org/10.1016/j.ambp.2007.04.006

Porter, A. C., & Smithson, J. L. (2001a). Defining, Developing and Using Curriculum Indicators. *CPRE Research Reports*. Available online at: http://repository.upenn.edu/cpre_researchreports/69

Porter, A. C., & Smithson, J. L. (2001b). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. H. Furhmans (Ed.), *From the Capitol to the classroom: Standards-based reforms in the States* (pp.60–80). Chicago: National Society for the Study of Education, University of Chicago press.

Razmjoo, S. A., & Kazempourfard, E. (2012). On the Representation of Bloom's Revised Taxonomy in Interchange Coursebooks. *Journal of Teaching Language Skills*, *31*(1), 171–204. Available online at: http://www.sid.ir/En/VEWSSID/J_pdf/13112012660407.pdf

Rezaee, M., & Golshan, M. (2016). Investigating the Cognitive Levels of English Final Exams Based on Bloom's Taxonomy. *International Journal of Educational Investigations*, *3*(4), 57–68. Available online at: http://www.ijeionline.com/attachments/article/53/IJEI.Vol.3.No.4.06.pdf

Riazi, M. A., & Mosalendejad, N. (2010). Evaluation of Learning Objectives in Iranian High-School and Pre-University English Textbooks Using Bloom's Taxonomy. *The Electronic Journal for English as a Second Language*, *13*(4). Available online at: http://www.tesl-ej.org/wordpress/issues/volume13/ej52/ej52a5

Stanny, C. (2016). Reevaluating Bloom's Taxonomy: What Measurable Verbs Can and Cannot Say about Student Learning. *Education Sciences*, *6*(4), 37. Available online at: doi:10.3390/educsci6040037

Stemler, S. E., & Tsai, J. (2008). 3 Best Practices in Interrater Reliability Three Common Approaches. In J. Osborne (Ed.), *Best practices in quantitative methods*. Thousand Oaks: SAGE Publications, Inc.

Teodorescu, R. E., Bennhold, C., Feldman, G., & Medsker, L. (2013). New approach to analyzing physics problems: A Taxonomy of Introductory Physics Problems. *Physical Review Special Topics – Physics Education Research*, *9*(1). Available online at: doi:https://doi.org/10.1103/PhysRevSTPER.9.010103

Valcke, M., De Wever, B., Zhu, C., & Deed, C. (2009). Supporting active cognitive processing in collaborative groups: The potential of Bloom's taxonomy as a labeling tool. *The Internet and Higher Education*, *12*(3–4), 165–172. Available online at: doi:http://dx.doi.org/10.1016/j.iheduc.2009.08.003

van Hoeij, M. J. W., Hararhuis, J. C. M., Wierstra, R. F. A., & van Beukelen, P. (2004). Developing a Classification Tool Based on Bloom's Taxonomy to Assess the Cognitive Level of Short Essay Questions. *European Veterinary Education: Structuring Future Development*, *43*(3), 261–267. Available online at: doi:http://dx.doi.org/10.3138/jvme.31.3.261

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's Taxonomy Debunks the "MCAT Myth". *Science*, *319*(5862), 414–415. Available online at:doi:10.1126/science.1147852

How much do I need to write to get top marks?

Tom Benton Research Division

Introduction

'How much am I supposed to write?' must be one of the most frequent questions students ask themselves when faced with an essay task. I remember this question being asked by someone in the class nearly every time such a task was set for homework at school, and my own children invariably ask me the same question every time I am encouraging them to do their homework. Despite the ubiquity of the question, clear answers are hard to come by. Teachers at my school would reply (rather unhelpfully) "how long is a piece of string?" whilst my response to my own children is rather more determined by how much I know they will be able to write before they start seriously complaining of fatigue than by any strong educational evidence.

There are good reasons not to answer this question. First and foremost is the fact that the quality of a response is not determined by the quantity of writing. For example, no published mark scheme for GCSEs will specify the amount that candidates are supposed to write but rather will rightly point markers towards the skill the assessment is supposed to be measuring; for example, in the case of English Literature[1], the extent to which candidates have identified the key features of the text they are studying and are able to communicate effectively. With these points in mind it is understandable if teachers want to make sure the student's efforts are focussed on producing a high-quality answer to the question and not on meeting some arbitrary target in terms of how much to write.

However, whilst this article is in no way arguing against the overriding importance of high-quality content, it is reasonable for students to want some guide to how much is expected in terms of length. An older *BBC Bitesize guide to English Literature GCSE* suggested that for a 45-minute examination students might have a target of roughly 450 words[2] – whilst also providing some more specific advice around time management and practice in structuring an essay. This article will supplement this advice by showing the amount of writing produced on average by candidates awarded different grades.

The relationship between the length of responses and the marks awarded to them has long been established within the field of automatic essay scoring. To take one example, Murray and Orii (2012) describe their own attempts to build a statistical model to achieve accurate essay scoring as part of a machine-learning competition. As a baseline comparator to their own technique, they present the correlation between predictions from a model based on essay length alone (both word count and character count) and the marks awarded to students. Across 9 different essay tasks, these correlations were all strongly positive, ranging from 0.50 to 0.82. Indeed, the extent to which automatic essay scoring algorithms can rely upon essay length has been criticised in research literature. For example, Perelman (2014, p.104) stated that "Automated Essay Scoring engines grossly and consistently over-privilege essay length in computing student writing scores" showing that, for the essays in this same competition, estimated scores from seven commercial vendors of automatic essay scoring were far more strongly related to word counts than was the case when human marking was used. However, there is no existing research linking the length of handwritten responses in GCSE examinations to the grades achieved by students.

Other research within the UK has investigated the average speed at which students can write under typical exam conditions. Such research is important for the purpose of determining the physical speed of writing below which a student may require further support by means of special considerations such as extra time or the facility to submit a typed (rather than a handwritten) essay as part of their examination. A review of this research is provided in Waine (2001). She reviewed 2 small-scale studies showing that in a free-writing task, where students had to decide what to write rather than simply copy it, students wrote on average between 14 and 18 words per minute. She also conducted her own study where, under examination conditions, 152 Year 10 (age 15) students were asked to write on the subject of 'My Life History' for a period of 30 minutes. Her results indicated that the mean writing speed of Year 10 students was 15 words per minute and that speeds between 10 and 20 words per minute were within the typical range. Similar research published by Patoss[3] (the professional association of teachers of students with specific learning difficulties) shows that, in a 20-minute free writing task, Year 10 students write at an average of 16 words per minute which rises to 17 words per minute for Year 11 students. Other research shows that when 16-year-old students are simply copying text they can write considerably even faster; at over 20 words per minute on average whilst writing neatly for 2 minutes, and at over 30 words per minute when writing as fast as possible (Barnett, Henderson, Scheib, & Schulz, 2009).

Overall, therefore, previous research has shown that the length of

1. See for example http://www.ocr.org.uk/Images/236719-mark-scheme-unit-a662-02-modern-drama-higher-tier-june.pdf (Retrieved 28 June 2017).
2. http://www.bbc.co.uk/schools/gcsebitesize/english_literature/prosejaneeyre/4prose_janeeyre_sprev1.shtml (Retrieved 28 June 2017).
3. https://www.patoss-dyslexia.org/SupportAdvice/InformationSheets/2012-09-02/Handwriting-Assessment/. (Retrieved 28 June 2017).