Educational Research and Evaluation: An International Journal on Theory and Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nere20

Assessing the ease of use in the environment and markers' acceptance of on screen marking: a Rasch measurement perspective

Zi Yan[a] & David Coniam[a]

[a] Department of Curriculum and Instruction, The Hong Kong Institute of Education, Tai Po, Hong Kong

To cite this article: Zi Yan & David Coniam (2013): Assessing the ease of use in the environment and markers' acceptance of on screen marking: a Rasch measurement perspective, Educational Research and Evaluation: An International Journal on Theory and Practice, 19:5, 461-483

To link to this article: http://dx.doi.org/10.1080/13803611.2013.793604

Educational Research and Evaluation, 2013
Vol. 19, No. 5, 461–483, http://dx.doi.org/10.1080/13803611.2013.793604

Assessing the ease of use in the environment and markers' acceptance of on screen marking: a Rasch measurement perspective

Zi Yan* and David Coniam

Department of Curriculum and Instruction, The Hong Kong Institute of Education, Tai Po, Hong Kong

(*Received 28 December 2012; final version received 22 March 2013*)

This study extends investigations into 2 areas: ease of use in the environment and markers' acceptance of on screen marking (OSM) in the Hong Kong public examination context. In contrast to previous studies where there was a single focus, this study comprises a heterogeneous approach. The sample contains scripts from three subjects (English, Chinese, and Liberal Studies). Scripts comprised essays and short-answer questions, as well as scripts written in either English or Chinese. Two scales assessing the ease of use and markers' acceptance of OSM were investigated from a Rasch measurement perspective; with both scales showing good psychometric properties. The findings revealed that markers generally had a high level of perceived ease of use in the environment and the overall acceptance of OSM was positive. Differences of person measures across language, question type, and subject were compared, and the implications of the two scales for future research are briefly discussed.

Keywords: on screen marking; Rasch measurement; ease of use in the OSM environment; acceptance of OSM; differential item functioning

Introduction

This study furthers the research conducted into on screen marking (OSM) in the context of a single jurisdiction – Hong Kong – where the Hong Kong Diploma of Secondary Education (HKDSE) examination is now marked on screen. As has been elaborated upon elsewhere (Coniam, 2009a, 2009b, 2011), three major issues are involved in the uptake of OSM. The first concerns the comparability of scores awarded on screen with those awarded for paper-based marking and marker discrepancy between the two modes. Studies conducted in the Hong Kong context (Coniam, 2009a) as well as internationally (Johnson, Nádas, & Bell, 2009; Powers et al., 1997) have confirmed the comparability of scores; this issue will consequently not be discussed further in the current article.

The second issue concerns ease of use in the OSM environment, which includes markers' technological proficiency. If markers feel the system is difficult to use, not user friendly, and remain unconvinced about anything other than paper-based marking, their marking may not be as accurate as one might hope. This issue, which Coniam (2013) refers to as "consumer validity", echoes research into perceived ease of use being an important factor in influencing user acceptance of information technologies. Research in the Hong

*Corresponding author. Email: zyan@ied.edu.hk

Kong context shows that markers regard themselves as technically competent, and their degree of self-assessed competence in this area has been increasing over time (Coniam, 2011).

The third area which the Hong Kong public examinations body, the Hong Kong Examinations and Assessment Authority (HKEAA), has been taking seriously concerns the issue of marker acceptance of OSM. This issue has been the subject of considerable research since Davis's (1989) pioneering technology acceptance model. At that time, marker reactions to OSM were initially mixed. In Powers et al.'s (1997) study, the OSM system was received "relatively positively" by most markers. In contrast, Zhang, Powers, Wright, and Morgan (2003) reported mixed reactions in their study, as did Fowles and Adams (2005). Twing, Nichols, and Harrison (2003) reported anxiety about marking on screen due to lack of computer proficiency among markers. Reactions towards OSM have, however, been improving with the passing of time, and there is a growing body of evidence to support the increasing acceptance of OSM by markers. In addition to studies by Coniam and Yeung (2010), the reader is also referred to Shaw (2008) and Shaw and Imam (2008).

The current study serves as a coda to previous studies in the Hong Kong context investigating marker attitude towards OSM, which has showed an increasing but slow acceptance of OSM in a number of ways. While both qualitative and quantitative measures have been used to gauge acceptance of the OSM system, recent investigations have involved examining reactions through a post-marking questionnaire survey (see Coniam, 2011; Coniam & Yeung, 2010).

The current study has three main purposes. First, it uses the Rasch model (1960/1980) as the approach to statistical analysis to investigate the psychometric properties of two scales, namely *Ease of Use in the OSM Environment* and *Acceptance of OSM.* This contrasts with previous studies that have, in the main, approached data analysis using classical test theory (CTT). A discussion of the advantages of Rasch over CTT is presented below.

Second, the current study extends the nature of the marker sample and type of questions marked. In previous studies (Coniam, 2009a; Coniam & Yeung, 2010), the focus has been on a single subject (e.g., English or Liberal Studies [LS]); in contrast, the current study takes a more heterogeneous approach in that it includes three subjects, namely, English, Chinese, and LS. Apart from the Coniam (2013) study, which involved the marking of short-answer questions, previous studies have all focused on extended essays. The current study differs from previous studies in that the questions being marked comprise both essays and short-answer questions. Furthermore, previous studies have tended to focus on marking in the English language, while the current study extends previous work by including examinations that are written in Chinese (Chinese language and LS) and marked on screen.

A final purpose of the current study is preparing the groundwork for the analysis of the final piece of the OSM jigsaw puzzle – a composite analysis of *all* subjects. In 2012, all 19 major subjects were marked on screen. The current study seeks to establish a robust instrument (via Rasch) for use in a future large-scale study encompassing all 19 subjects, the analysis of which, it is anticipated, will illustrate how markers across all subjects perceive the OSM system – which will provide the HKEAA with definitive information regarding future developments and improvements to the OSM system. The results of a large-scale study will have implications for any jurisdiction where all subjects are marked on screen.

To provide a background for the current study and to orient the reader, previous research regarding language and question type issues in on screen reading and marking will now be discussed. As will have been noted, there are a range of issues which may affect how markers react in terms of reading English or Chinese on screen and how markers mark extended essays and short-answer questions.

In the context of reading on screen in English or Chinese, the nature of the display appears to make a difference. Chen and Chien (2005) investigated how the type of screen affected reading comprehension accuracy with smaller devices such as laptops and mobile phones. They note that different devices optimally require different screen display types, and that reading accuracy can differ between English and Chinese depending on the display type. Chan, So, and Tsang (2011) report an investigation into on screen proofreading in the Hong Kong Chinese language context. The nature of proofreading generally involves micro-level rather than macro-level (or global) reading which is the case with extended essays. Given this, Chan et al.'s study may be seen to be broadly comparable with the marking of short-answer questions in the Chinese language examination in the current study because marking short-answer responses involves micro-level marking rather than the more macro-level marking required for extended essays. Noting that, for English, character size affects reading speed, with smaller characters producing faster English reading speed, Chan et al. propose to investigate the influence of display factors on subjects' proofreading speed and accuracy in Chinese Hong Kong Certificate of Education Examination (HKCEE) proofreading tests. Yen, Tsai, Chen, Lin, and Chen (2011) investigated typographic variables – character size, character spacing, and font type – on eye-movement measures in reading Chinese on screen. Thus, given Yen et al.'s conclusion that reading on screen in Chinese and English may have different characteristics, there appears to be ample justification for including the language variable in the current study.

Regarding question type, previous studies into OSM have had a single focus. This has been either the marking of extended essays (Coniam, 2009a), where the entire script needs to be pondered upon in order to evaluate the content (Shaw, 2008), or the focus has been on the marking of short-answer questions, where marking is more "mechanical", requiring only "expeditious reading strategies" (Shaw, 2008, p. 268), with items only needing to be marked right or wrong. The marking of short-answer questions is nonetheless demanding in a different way in terms of number of items to be marked and the time pressure that the markers experience. While the marking of the English and Chinese language papers in the current study requires less processing than does the reading of extended essays, considerable concentration and attention is required if approximately 6,000 items are to be processed and scored accurately in a single marking session.

The current study purposely includes a mix of question types: English language and Chinese language, where marking consists of short-answer questions, and LS, where marking involves reading extended essays. There is also a mix of languages in the marking in that while English language and Chinese language papers were only marked in English or Chinese, LS papers were answered and marked in both English and Chinese.

The Rasch model

As mentioned, previous studies have in the main used CTT to analyse data – specifically the survey data which the current study extends. While the use of CTT enables statistical significance to be examined, there are inherent weaknesses in this method. First, analytical techniques in CTT require linear, interval scale data input (Wright, 1997). Raw data collected through Likert-type scales, however, are usually ordinal since the categories of Likert-type scales indicate only ordering without any proportional levels of meaning. Applying conventional analysis on ordinal raw data can therefore lead to potentially misleading results (Bond & Fox, 2007; Wright, 1997). Second, CTT uses total score to indicate respondent ability levels. This results in person ability estimates being item dependent; that

is, although person abilities may be the same, person ability estimates are high when items are easy but low when items are difficult. Similarly, item difficulty estimates are similarly sample dependent; that is, even though item difficulties themselves are invariant, item difficulty estimates appear high when respondents' competence is low but low when respondents' competence is high.

The use of the Rasch model (1960/1980) enables different facets (person ability and item difficulty in the current instance) to be modelled together. First, in the standard Rasch model, the aim is to obtain a unified and interval metric for measurement. The Rasch model converts ordinal raw data into interval measures which have a constant interval meaning and provide objective and linear measurement from ordered category responses (Linacre, 2006). This is not unlike measuring length using a ruler, with the units of measurement in Rasch analysis (referred as the "logit") evenly spaced along the ruler. Second, once a common metric is established for measuring different phenomena (test takers and test items being the most obvious), person ability estimates are independent from the items used, with item difficulty estimates being independent from the sample recruited because the estimates are calibrated against a common metric rather than against a single test situation (for person ability estimates) or a particular sample of test takers (for item difficulty estimates). Third, Rasch analysis prevails over CTT by calibrating persons and items onto a single unidimensional latent trait scale (Bond & Fox, 2007; Wright, 1992). Person measures and item difficulties are placed on an ordered trait continuum by which direct comparisons between person measures and item difficulties can be easily conducted. Consequently, results can be interpreted with a more general meaning. Further, as the Rasch model provides a great deal of information about each item in a scale, its use enables the researcher to better evaluate individual items and how these items function in a scale (Törmäkangas, 2011).

The Rasch model has been widely applied in educational research, especially in the field of large-scale assessment (Schulz & Fraillon, 2011; Wendt, Bos, & Goy, 2011). It provides better assessments of performance, enhances the quality of measurement instruments, and provides a clearer understanding of the nature of the latent trait (Bos, Goy, Howie, Kupari, & Wendt, 2011). There is an extensive literature concerning the advantages of the Rasch model over traditional psychometric analysis. Using both traditional and Rasch analyses, Muís, Winne, and Edwards (2009), for example, investigated the psychometric properties of two instruments widely used in educational research settings – the Achievement Goals Questionnaire (AGQ) and the Patterns of Adaptive Learning Scale (PALS). The AGQ and PALS appeared to be reliable and valid instruments within the framework of traditional psychometric analysis. Rasch analysis, however, identified several problems which had been neglected in the traditional analysis but which had important empirical implications, namely, mismatch between the distribution of respondent abilities and item difficulties, low person reliability, and poor category function at the lower end of the scale.

Background to the study

The Hong Kong secondary school curriculum underwent significant restructuring in 2009 when the British 5+2 secondary system (with examinations at Years 11 and 13) was replaced by a 3+3 year system (similar to the Chinese and Australian education models). Secondary education now lasts 6 years with a single public examination (HKDSE) at the end of Year 12 (age 18), the annual candidature for which is in the region of 75,000.

In addition to the structural changes to the education and examination system, a major operational change accompanying the introduction of the HKDSE is that, from 2012, paper-based marking has been replaced by OSM in the core and major elective subjects.

The HKEAA, Hong Kong's public examination body, has, for some time, been investigating a variety of procedures and processes related to computerisation, including a dedicated OSM system. In late 2005, the Hong Kong SAR Legislative Council (Legislative Council Panel on Education, 2005) allocated approximately US$25 million towards the IT modernisation of the HKEAA.

The first step towards the adoption of OSM was taken in 2007 when all Year 11 Hong Kong Certificate of Education Examination (HKCEE) English Language and Chinese Language scripts were marked on screen. To investigate and validate the adoption of OSM, a series of studies were subsequently conducted with key subjects – English language and LS – to compare the two modes of marking. Coniam (2009a, 2011) and Coniam and Yeung (2010) reported quantitatively and qualitatively on marking comparability issues and marker attitude for the two different subjects.

The OSM marking process in Hong Kong is as follows. After each examination, scripts are delivered to a dedicated scanning centre where scripts are scanned and images saved. Markers then need to go to special OSM centres (of which Hong Kong has seven) at strategic locations where they mark at dedicated workstations using a purpose-built system.

The OSM system requires markers to progress first through *training* and *qualifying (*demonstrating that their marking has attained the appropriate standard) before finally moving to independent *live marking*. During live marking, "control scripts" (with scores previously agreed by chief examiners) are regularly issued to markers so that their marking standard can be monitored and instant feedback provided. A further reliability-enhancing feature of the OSM system is that markers may view their own marking statistics (marking speed, mean score, mark distribution spread) and observe how their control script marking statistics compare with those of the chief examiners. For a more detailed discussion of OSM with regards to procedures, benefits and drawbacks, the reader is referred to Coniam (2009a).

The study

This section describes the data that make up the study, the research questions, and the methods used to analyse the data.

The data for the current study are drawn from three subjects: the 2009 Hong Kong Advanced Level Examination (HKALE) LS examination (candidature 3,307), the 2012 HKALE Use of English, Paper C (candidature 39,807), and the 2012 HKDSE Chinese Language Paper 1 (candidature 71,284).

The majority of Hong Kong public examinations are marked by experienced teacher markers, due to the obvious need for subject expertise. There are, however, sections of certain papers in some subjects, generally comprising "mechanical" short-answer questions, which are able to be marked by undergraduate students in the relevant discipline. To mark such short-answer questions in the English and Chinese language examinations, marking panels comprising undergraduate language or linguistics major students from local Hong Kong universities are engaged. These marking panels mark the English and Chinese language examinations in a single day, at a single marking centre. The undergraduate student markers are paid for a day's work – approximately 6 hr worth of marking: from 9 a.m. to 4 p.m. with a break for lunch (and rest breaks as necessary). In contrast, for teacher markers, the marking period usually extends to 2 weeks, with markers typically marking for a 3-hr stretch at a marking centre of their choice.

In the current study, after marking was complete, markers filled in a short questionnaire (see Appendix 1). This comprised – in addition to background demographic data – two sections to gauge (a) their perceived ease of use in the OSM environment and (b) the level of their acceptance of OSM. The Ease of Use in the OSM Environment scale, consisting of 10 items, tapped OSM centre comfort issues, markers' computer proficiency, their competency in manipulating the mouse, enlarging and scrolling the screen image, as well as ergonomic issues such as desktop height and screen resolution. The Acceptance of OSM scale, consisting of 9 items, tapped issues such as how accurate they felt their on screen/on-paper marking was, how tired their eyes became through marking in the two modes, and how often they needed to take a break while marking. It also enquired about their preference as to marking mode, that is, OSM or paper-based marking.

Questions were posed on a 6-point Likert scale, with "1" indicating a positive response or agreement and "6" a negative response or disagreement. For all items, a lower score represents a higher level on the latent trait under investigation; that is, an overall lower score by a respondent indicates a higher level of perceived ease of use or acceptance of OSM.

Research questions

The sample in the current study contains the entire set of scripts from three subject areas: Liberal Studies (LS), English language, and Chinese language. LS scripts comprise extended essays, while English language and Chinese language scripts comprise the sections requiring short-answer questions. In addition, there is a language factor, with candidate scripts written in English or Chinese.

With this heterogeneous background – and with the overriding objective of calibrating a robust instrument for use in a future large-scale study – there are two linked sets of hypotheses in the current study: one set for ease of use in the OSM environment and another for acceptance of OSM. These are as follows:

(1) Concerning ease of use in the OSM environment, the item difficulties will not differ for markers with regards to:
   - the language (English or Chinese) that scripts are marked in;
   - the type of question (extended essays or short-answer questions);
   - the subject (English or Chinese or LS).

(2) Concerning acceptance of OSM, the item difficulties will not differ for markers with regards to:
   - the language (English or Chinese) that scripts are marked in;
   - the type of question (extended essays or short-answer questions);
   - the subject (English or Chinese or LS).

The statistical approach in the current study will involve, as mentioned, the use of Rasch measurement. While such an approach will enable the research questions above to be explored, it will also permit a robust scale with sound psychometric properties to be established from the items in the questionnaire.

Data analysis

The software used to conduct the Rasch analysis was Winsteps version 3.7 (Linacre, 2011). In Rasch analysis, the mean difficulty of all items is initially set at zero. Therefore, for the

scales used in this study, person measures lower than zero indicate a positive response while person measures higher than zero indicate a negative response.[1]

Considering the heterogeneity of the sample and the fact that there are different groups of markers across language, question type, and subject, a method of analysis pertinent to the current study is differential item functioning (DIF), which refers to differences in the functioning of items across groups. DIF analysis serves to check construct equivalences across groups and to examine whether different groups have a different interpretation of or perspective on the items (Wang, 2000). DIF exists when subjects from two or more groups with the same level of latent trait have a different probability of answering an item correctly. In other words, an item exhibiting DIF has different item difficulties for different groups. In the view of Rasch measurement, an item with DIF does not measure the same construct as other items in a scale – contradicting the principle of objective measurement: that a valid measurement should measure only one trait at one time (Bond & Fox, 2007). The purpose of DIF analysis is, therefore, to investigate lack of invariance of item difficulty and to deal with a possible threat to internal validity (Zumbo, 2007). As a rule of thumb, a difference equal to or larger than 0.64 logits (Educational Testing Service, cf. Zwick, Thayer, & Lewis, 1999) and with statistical significance ($p < .05$) will be regarded as a sign of substantial DIF.

Four criteria other than DIF were also used to investigate the psychometrical properties of the scales, as well as to provide measurement information for items, specifically: Rasch reliability, item fit statistics, the amount of variance explained by measures, and category functioning.

Rasch measurement provides both person reliability and item reliability indices. Rasch person reliability indicates the consistency of person ordering along the latent trait scale, with Rasch item reliability referring to the replicability of item placements along the trait continuum if the same set of items were administered to another, similar sample (Bond & Fox, 2007).

Item fit statistics, including outfit and infit mean square error (MNSQ), indicate the extent to which the data match the specifications of the Rasch model. The values of outfit and infit MNSQ range from 0 to positive infinity with 1.0 indicating a perfect fit to the Rasch model. It has been suggested by researchers (e.g., Anshel, Weatherby, Kang, & Watson, 2009; Linacre, 2006; Weigle, 2002) that a range of 0.5 to 1.5 indicates acceptable model fit; this range is therefore generally taken as an indicator of productive measurement.

Variance explained by Rasch measures refers to the proportion of variance in the observed data that is able to be explained by the item difficulties, person abilities, and rating scale structures (Linacre, 2006). A higher proportion of variance means that the Rasch model better predicts both items and persons.

Questions were posed on a 6-point Likert scale. Category functioning is, therefore, another important aspect to be checked to ensure the quality of the scale. If the categories function well, no "threshold disordering" (Bond & Fox, 2007; Linacre, 2002) will be observed. That is, the threshold calibrations should advance monotonically with category, indicating that higher performance categories correspond to higher measures of the latent trait.

After the scales were examined via Rasch analysis and it was confirmed that all remaining items fit the Rasch model and were DIF free, person measures on the Ease of Use in the OSM Environment and Acceptance of OSM scales were then calibrated. Those person measures – as interval data – were then analysed using *t*-test and analysis of variance (ANOVA) statistics to compare the respondents' perceived ease of use and acceptance of OSM across variables including *language*, *question type*, and *subject*.

Table 1. Language and question type for each subject.

	English	Chinese	Liberal Studies
Language	English ($N = 301$)	Chinese ($N = 34$)	Chinese ($N = 37$); English ($N = 9$)
Question Type	Short answer ($N = 301$)	Short answer ($N = 34$)	Extended essay ($N = 46$)

Language, question type, and subject of scripts

Questionnaires were retrieved from 301 English language markers, 34 Chinese language markers, and 46 LS markers. Of these 381 subjects, 310 (81.4%) marked examination scripts in English, while 71 (18.6%) marked in Chinese. Regarding question type, the majority (335, 87.9%) marked short-answer questions, while 46 (12.1%) marked essays. Table 1 summarises the language, question type, and subject of examination scripts analysed in the current study. It should again be noted that although the sample sizes were different, they did represent *the entire set* of markers and scripts for each subject.

Results and discussion

The two scales, namely Ease of Use in the OSM Environment and Acceptance of OSM, were analysed separately. The psychometric properties of different versions of scales from Rasch measurement perspective were compared. Since deleting items may affect the accuracy of the remaining scale and the error rate of model estimates (Hart & Wright, 2002), the identified misfitting or DIF items were removed from the scale one at a time, with Rasch analysis re-applied until all remaining items showed sufficient fit to the Rasch model and were DIF free. This is standard procedure for Rasch analysis.

Ease of Use in the OSM Environment scale

Before calibrating items in the Ease of Use in the OSM Environment scale, the fit statistics of all markers were checked since, as argued by Bond and Fox (2007), underfitting persons (MNSQ fit statistics much higher than 1.0) are detrimental to calibrating a measurement scale. One way to improve the calibration of a measurement scale is to exclude temporarily some respondents whose performances do not fit the Rasch model (Verhelst & Glas, 1995). In this case, markers were excluded from the scale calibration if both their outfit MNSQ and infit MNSQ were higher than 2.0. Consequently, a total of 42 cases were excluded. The Rasch calibration of the 11 items based on the data from the remaining 339 markers identified Item 16 ("What is your preference for mark input? Mouse vs. keyboard") as showing underfit (both infit and outfit MNSQ being higher than 1.5), suggesting that Item 16 might be measuring a different construct from other items in the scale.

DIF emerged on three *language* items. On Items 11 ("How comfortable were you reading off the screen?") and 15 ("How easily could you input marks using the keyboard?"), markers marking in Chinese had lower scores, while on Item 16 ("What is your preference for mark input?"), markers marking in English had lower scores. These findings indicate that, given that they are at the same level of perceived ease of use, markers marking in Chinese felt more comfortable reading off the screen, finding it easier to input marks using the keyboard than markers marking in English. Further, not surprisingly, markers marking in Chinese preferred to input marks using the keyboard, while markers marking in English preferred to use the mouse.

For the factor *question type*, DIF emerged for four items. On Items 11 and 15, markers marking extended essays had lower scores, while on Items 14 ("How easily could you input marks using the mouse?") and 16, markers marking short-answer questions had lower scores. These results showed that markers with the same level of perceived ease of use felt more comfortable reading off the screen, and found it easier to input marks using the keyboard when marking extended essays. In contrast, markers marking short-answer questions found mark input easier using the mouse and, therefore, preferred using the mouse for mark input.

Four items demonstrated DIF across *subject*. On Items 11 and 15, LS markers had lower scores than did English language/Chinese language markers, while on Items 14 and 16, English language/Chinese language markers had lower scores than LS markers. To an extent, this result echoes the DIF across question type, since LS consists of extended essay questions, while the English/Chinese tests consists of short-answer questions.

According to Hart and Wright's (2002) suggestion, the identified misfitting or DIF items were removed from the scale one at a time. Item 11 appeared to be DIF free on *language* and to have marginal DIF (the difference of item difficulties across groups being 0.65 logits) on *question type* and *subject* when Items 14, 15, and 16 were removed from the scale. Given that Item 11 examines an important aspect of the ease of use in the OSM environment, this item was retained in the scale. Consequently, items showing misfit to the Rasch model (16) and items with DIF (14, 15, and 16) were removed from the scale. The final version of the Ease of Use in the OSM Environment scale therefore comprises 7 items. As can be seen from Table 2, however, the shortened (7-item) scale has coefficients – in terms of person reliability, item reliability, and variance explained – comparable to the original 10-item scale. Item reliability fell slightly from 0.99 to 0.95, and variance explained from 62.0% to 53.6%, while person reliability increased from 0.77 to 0.80. Such similarity of psychometric properties of two versions of scale supports the adoption of the reduced version. It was, therefore, concluded that these 7 items fit the expectations of the Rasch model quite well. The 10 items in the original scale were possibly assessing more than one dimension; with the removal of the misfitting or DIF items, the remaining 7 items could now be taken to assess a single latent trait – ease of use in the OSM environment.

Table 2. Rasch results for the Ease of Use in the OSM Environment and Acceptance of OSM scales.

	Rasch person reliability	Rasch item reliability	Variance explained	Misfitting items	DIF items
Ease of Use in the OSM Environment scale					
10-item scale	0.77	0.99	62.0%	16	11, 14, 15, 16
7-item scale*	0.80	0.95	53.6%	N/A	N/A
Acceptance of OSM scale					
9-item scale	0.79	0.98	48.3%	N/A	19, 21
7-item scale	0.77	0.99	54.3%	17	17
6-item scale	0.80	0.99	59.8%	N/A	N/A

*Item 11 was retained since it appeared to be DIF free when Items 14, 15, and 16 were removed from the scale.

According to the response structures of the items, a grouped rating scale model was used in the calibration for the Ease of Use in the OSM Environment scale. Items 12 and 13 shared the same rating scale, but the other items had their own rating scales. The threshold calibrations, or the step difficulties, for each item are presented in Table 3. It can be seen that the rating scales for all items function well, and there is no "threshold disordering". All the threshold calibrations advance monotonically with category, indicating that higher performance categories represent higher measures of ease of use. The calibrations for Threshold 5 (the intersection point between Categories 5 and 6) for Items 08 and 09 are not available since there is only one marker who selected Category 6 for these items. Categories 5 and 6 for these two items were therefore combined, which resulted in a 5-category response structure without threshold disordering. Collapsing the categories in this manner had very little impact on the scale calibration since the psychometric properties of the scale remained unchanged.

Acceptance of OSM scale

Regarding the Acceptance of OSM scale, after the removal of the 32 extremely underfitting markers (both their outfit MNSQ and infit MNSQ were higher than 2.0) from the calibration, no item was identified as misfitting in the Rasch analysis. There is now no item demonstrating DIF across *language*.

Two items showed DIF on the factor *question type*. On Item 19 ("How often did you need to take a break while marking on screen?"), markers marking extended essays had lower scores, while on Item 21 ("How much support and feedback did you receive from the OSM system?"), markers marking short-answer questions had lower scores. Markers marking extended essays took less frequent breaks, but reported receiving less support and feedback from the OSM system than markers marking short-answer questions. HKEAA personnel were subsequently approached and interviewed as to why this should be the case. It transpired that the current study involved one lengthier period of time working on screen for marking short-answer questions (for English and Chinese). Specifically, the marking of the entire paper was completed in a single day. This required markers to work on screen for approximately 6 hours in that single day, each marker marking

Table 3. Threshold calibrations for items.

	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5
Ease of Use in the OSM Environment scale					
Item 07	–4.33	–0.74	0.27	0.55	4.25
Item 08	–3.57	–0.38	1.37	2.58	N/A
Item 09	–3.39	0.07	1.26	2.06	N/A
Item 10	–3.96	–0.73	1.06	1.50	2.13
Item 11	–4.92	–1.58	0.55	1.68	4.26
Items 12 & 13	–3.78	–1.38	–0.01	0.90	4.27
Acceptance of OSM scale					
Item 18	–3.45	–0.71	–0.16	1.24	3.08
Item 20	–3.32	–1.01	–0.57	1.07	3.84
Item 22	–4.22	–1.70	0.25	1.54	4.12
Item 23	–4.43	–1.22	0.77	1.19	3.69
Item 24	–3.80	–0.73	1.18	0.70	2.65
Item 25	–2.67	–0.53	0.63	0.83	1.74

approximately 6,000 short items. When marking extended essays, LS markers marked for a 3-hour stretch, with the marking period lasting 2 weeks.

DIF emerged on the same items across *subject*. For Item 19, markers marking Chinese/LS had lower scores than did the English markers, while for Item 21, markers marking English had lower scores than did LS/Chinese markers. These findings echo the results reported above on the DIF findings across *question type*.

After removing the DIF items one at a time, Item 17 ("How much training did you receive?") misfitted the Rasch model, with DIF emerging on *question type* (lower scores for short-answer questions) and *subject* (English/Chinese being lower than LS). Markers marking short-answer questions (i.e., English and Chinese) felt they received more training than did markers marking extended essays (i.e., LS).

On the basis of the rationale laid out above, Item 17 appears not to be measuring the intended construct, that is, acceptance of OSM. This item was consequently removed – resulting in a 6-item scale with all items showing fit to the Rasch model and being DIF free. While the 6-item scale is again comparatively short, its psychometric properties are better than those of the longer versions, as Table 2 illustrates. Person and item reliabilities slightly increased from 0.79 to 0.80 and 0.98 to 0.99, respectively, whereas the amount of variance explained by measures in the 6-item scale increased from 48.3% to 59.8% – suggesting that the 6-item scale provides a better prediction of item and person performance. The conclusion arrived at was that the original 9-item scale was possibly also assessing more than a single dimension. After removing 3 misfitting or DIF items, the remaining 6 items fit the expectations of the Rasch model, that is, that the single latent trait – acceptance of OSM – was being assessed.

A partial credit model was used in the calibration for the Acceptance of OSM scale since each item had a unique rating scale. It can be seen from Table 3 that the threshold calibrations for all items, except Item 24, advance monotonically with category. "Threshold disordering" exists for Item 24 on Threshold 5 (the intersection point between Categories 5 and 6), which has a lower calibration (0.70 logits) than that of Threshold 4 (the intersection point between Categories 4 and 5) (1.18 logits). This implies less frequent observation on Category 4, and this category is never a modal category in these data. As a response to this threshold disordering, Categories 4 and 5 were combined, and a 5-category rating scale was used for Item 24. This resulted in subtle changes to the psychometric properties of the scale. The Rasch person reliability decreased slightly from 0.80 to 0.79, and the variance explained by measures decreased from 59.8% to 58.2%.

Person measures on the Ease of Use in the OSM Environment and Acceptance of OSM scales

The Rasch model calibrates person measures and item difficulty on the same scale. Figure 1 and Figure 2 are the item-person maps for Ease of Use in the OSM Environment and Acceptance of OSM scales, respectively. As presented in the maps, the continuum to the left side of the ruler indicates the markers' measures on each scale. The markers with lower levels of perceived ease of use/acceptance (higher scores on the scales) are placed at the top of the scale, and those with higher levels of perceived ease of use/acceptance (lower scores on the scales) are placed at the bottom. In a similar way, the items with higher difficulty levels (on which markers tend to have low scores) are placed at the top, and those with lower difficulty levels (on which markers tend to have high scores) are placed at the bottom of the middle continuum.

The mean estimate of markers' perceived ease of use is –1.95 (*SD* = 1.68) logits, which is much lower than the mean estimate of item difficulty (zero), indicating quite a high level of perceived ease of use. The item difficulties range from –0.54 logits (Item 09) to 0.76 logits (Item 13). Figure 1 shows that markers spread across a quite wide range of the scale while the items tend to cluster at the top of the scale. The majority of the markers

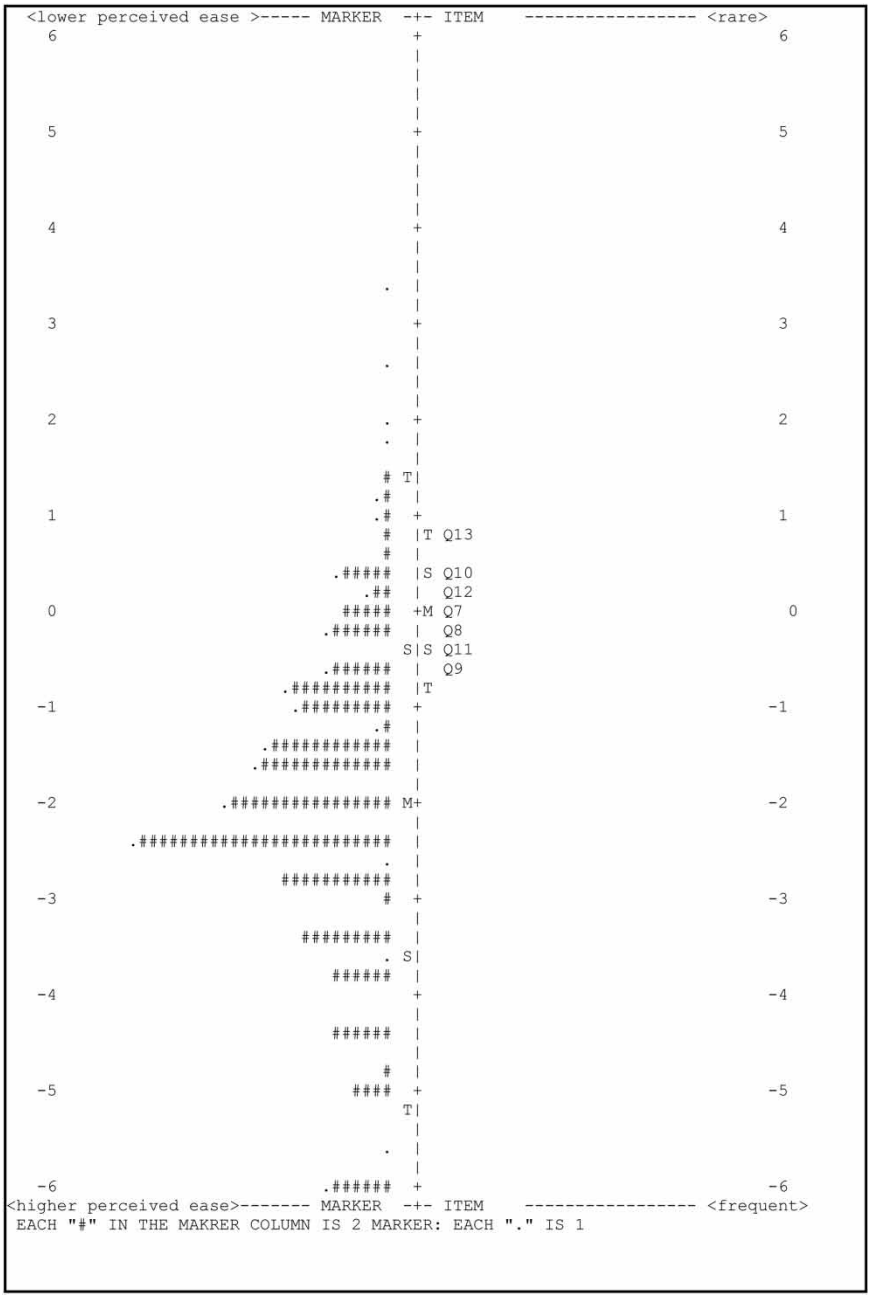

Figure 1. Item-person map for the Ease of Use in the OSM Environment scale.

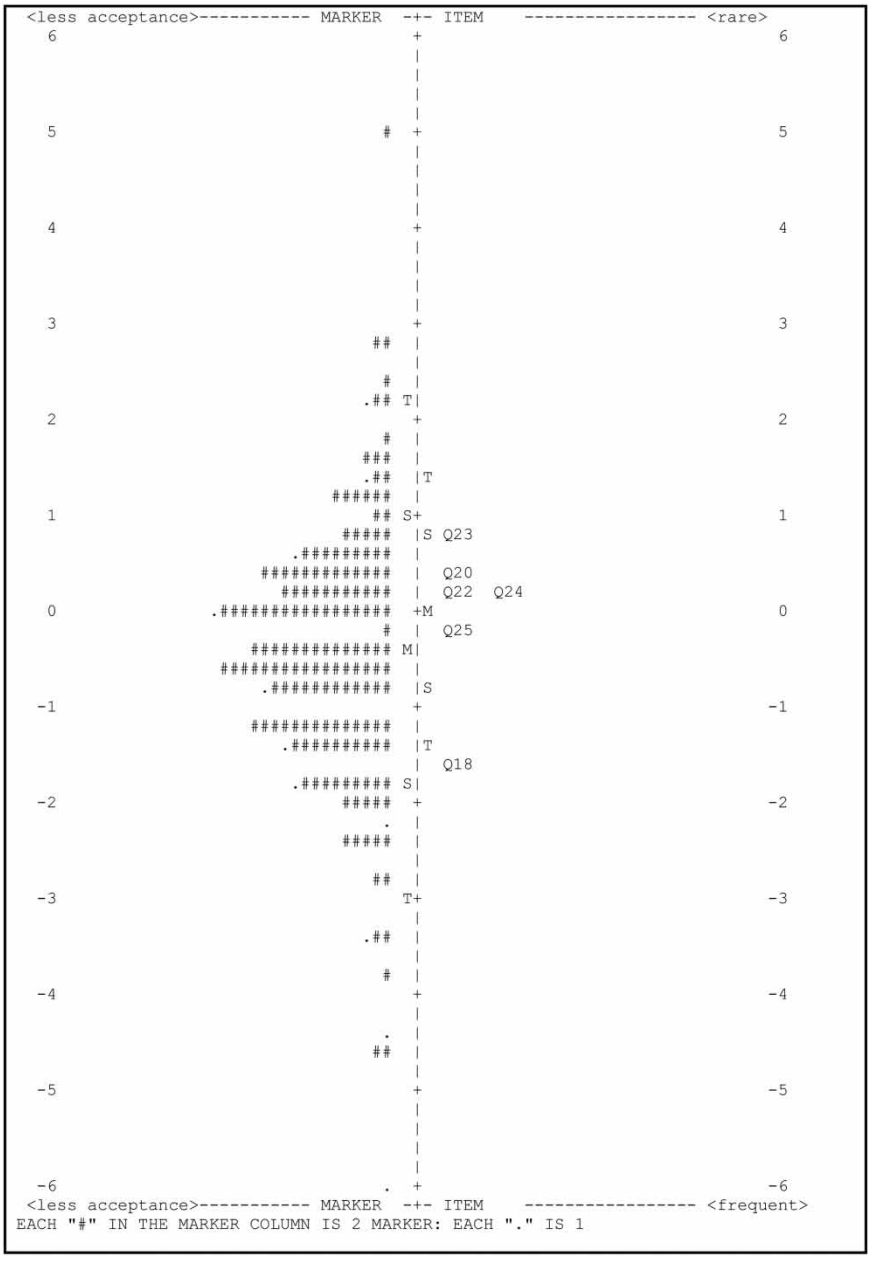

Figure 2. Item-person map for the Acceptance of OSM scale.

tend to have low scores (i.e., a high level of perceived ease of use) on items – illustrated by their being located below the items on the ruler.

The mean estimate of markers' acceptance of OSM is –0.42 (*SD* = 1.39) logits, which is close to the mean estimate of item difficulty (zero). The item difficulties range from –1.51 logits (Item 18) to 0.87 logits (Item 23). The results presented in Figure 2 suggest that, overall, the acceptance of OSM is positive, although a small number of markers remain slightly negative.

Comparisons of mean person measures

As mentioned earlier, ordinal data collected through Likert-type scales can be transformed, via Rasch calibration, into interval measures which are appropriate for CTT inferential statistics. Rasch-calibrated person measures for all 381 markers on the two scales were subsequently compared across *language*, *question type*, and *subject*. The results are presented in Table 4.

Regarding ease of use in the OSM environment, *t*-test results showed that markers marking English had significantly lower scores (i.e., higher ease of use) ($p < .05$) than markers marking Chinese; but no significant differences emerged between markers marking short-answer questions and those marking extended essays. On ANOVA, the main effect of *subject* was statistically significant ($p < .05$); post-hoc comparisons, however, indicated that only the difference between English and Chinese was significant ($p < .05$). Markers for English had significantly lower scores (i.e., higher ease of use) than markers for Chinese. An explanation for this discrepancy between the English and Chinese markers may be observed from the written comments that markers were invited to contribute – on any aspect of the OSM system or their OSM experience – in the final section of the post-marking questionnaire. Of the 280 English language markers, comments were proffered by 106 (37.9%) of them. Of these, 22 comments (20.7%) were positive, if not complimentary about the OSM system, showing a clear shift in attitude towards OSM. This contrasts with previous studies (e.g., Coniam & Yeung 2010), where additional comments have only been negative. Most negative comments from the English language markers related to the general environment – over cold air-conditioning, not enough female toilets, and so forth. Only 8 comments (7.5%) related to technical issues, where matters such as screen resolution, problem with zooming, and scrolling functions were mentioned. In contrast to the English language markers, all comments received from the 15/34 (44.1%) Chinese language markers who responded were negative. Four markers (26.7%) commented on the fact that the current OSM system forces markers to use one particular system for inputting Chinese language characters; they requested that the system allow different Chinese input methods to be used. Three respondents (20%) also commented on the fact that the system requires them to use the mouse more than the keyboard, and consequently felt hampered by this restriction. It may therefore be that the lower ease of use reported for markers marking in Chinese may in part be a result of certain limitations in the current OSM system itself.

Table 4. Comparisons of mean person measures (in logits).

	Ease of Use in the OSM Environment	Acceptance of OSM
Language		
English ($N = 310$)	–1.67	–0.50
Chinese ($N = 71$)	–1.25	+0.08
t-test results	$t = –2.012$, $df = 379$, $p = .045$	$t = –3.857$, $df = 379$, $p = .000$
Question type		
Short answer ($N = 335$)	–1.63	–0.54
Extended essay ($N = 46$)	–1.27	+0.65
t-test results	$t = –1.431$, $df = 379$, $p = .153$	$t = –6.859$, $df = 379$, $p = .000$
Subject		
English ($N = 301$)	–1.70	–0.54
Chinese ($N = 34$)	–1.03	–0.47
Liberal Studies ($N = 46$)	–1.27	+0.65
ANOVA results	$F(2, 378) = 3.772$, $p = .024$	$F(2, 378) = 23.329$, $p = .000$

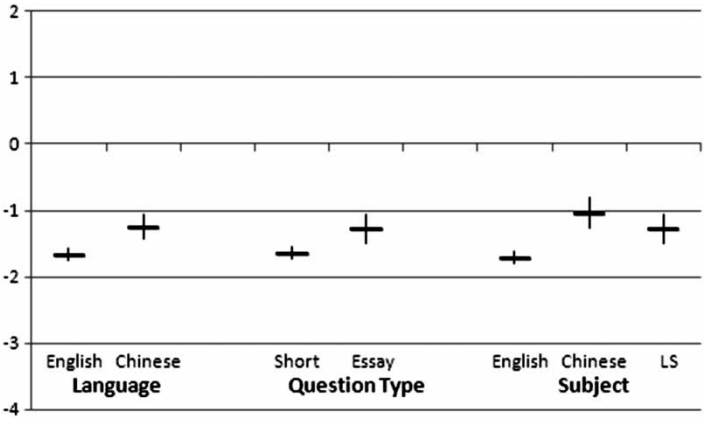

Figure 3. Comparisons of mean person measures on the Ease of Use in the OSM Environment scale ($M \pm 1$ *SE*).

With regard to markers' acceptance of OSM, results showed that markers marking in English exhibited a higher level of acceptance of OSM ($p < .01$) than those marking in Chinese; markers marking short-answer questions were more positive towards OSM ($p < .01$) than those rating extended essays. Markers for English exhibited the highest acceptance of OSM, followed by markers for Chinese, with markers for LS holding the lowest level of acceptance. Post-hoc comparisons, however, indicated that only the differences between LS and English/Chinese were significant ($p < .01$). A possible explanation for this lies in the time frame of the studies. The data for LS markers were collected in 2009, whereas the data collection for Chinese/English markers took place in 2012. It can thus be hypothesised that with the passage of time, markers are being more positive about the OSM system.

To more clearly illustrate the mean differences in ease of use in the OSM environment and acceptance of OSM across language, question type, and subject, graphics summarising the results are presented in Figures 3 and 4. In the figures, mean person measures are represented by the horizontal lines, with plus or minus one standard error indicated by the vertical lines.

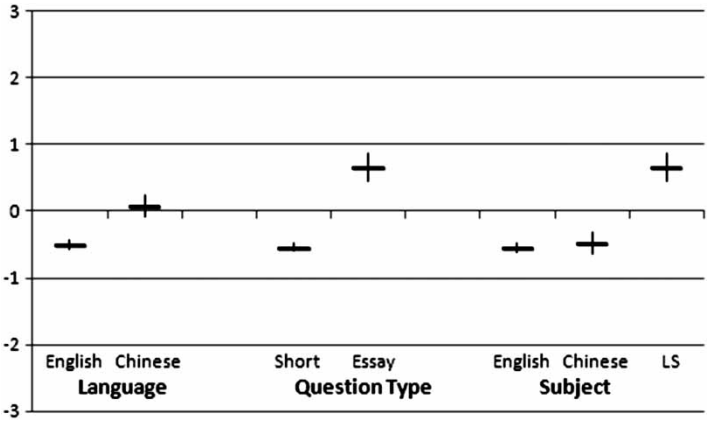

Figure 4. Comparisons of mean person measures on the Acceptance of OSM scale ($M \pm 1$ *SE*).

As the graphics in Figure 3 illustrate, on the Ease of Use in the OSM Environment scale, markers are clustered within a narrow range, affirming that none of the three variables impinges on the ease of use. In contrast, Figure 4 shows how, on the Acceptance of OSM scale, Chinese language markers are less accepting of OSM than English language markers, extended essays evoke less acceptance than short-answer questions, and LS markers are less accepting of OSM than Chinese language and English language markers.

Conclusion

This article has reported an investigation into the psychometric properties of two scales, namely, Ease of Use in the OSM Environment and Acceptance of OSM, using Rasch analysis. The purpose of the study has been to establish two robust scales in order to provide a global perspective on ease of use in the environment and on markers' acceptance of the OSM system which Hong Kong has now adopted across the board, that is, for all subjects. The scales were examined within the context of two linked sets of three hypotheses: that the item difficulties in these two scales would not differ for markers with regards to:

(1) the language (English or Chinese) that scripts were marked in;
(2) the type of question (extended essays or short-answer questions);
(3) the subject (English or Chinese or LS).

As has been illustrated, the three hypotheses were mostly accepted. Seven items out of 10 in the Ease of Use in the OSM Environment scale and 6 items out of 9 in the Acceptance of OSM scale appeared DIF free across variables including *language*, *question type*, and *subject*.

Regarding the Ease of Use in the OSM Environment scale, Rasch analysis identified one item showing misfit to the Rasch model and three items exhibiting DIF. These items were therefore removed, one by one, resulting in a 7-item scale which had comparable psychometric properties to the original 10-item scale. As for the Acceptance of OSM scale, one item was identified as misfitting the Rasch model and three items exhibiting DIF. These items were removed from the scale – resulting in a 6-item scale which showed even better psychometric properties than the original 9-item scale. The Rasch person reliabilities for the two scales are 0.80, which might be sufficiently accurate for diagnosing individuals. This result accomplished one of the purposes of the current study: to establish a robust instrument for use in a future large-scale study encompassing all 19 subjects. Although these two scales have been previously reported (Coniam & Yeung, 2010), the current study furthered the investigation from two perspectives. First, Coniam and Yeung's (2010) study involved only 49 markers who marked LS; the current study substantially extended the sample – to 381 markers marking three subjects, namely, Chinese, English, and LS. Second, the previous study limited itself to reporting in a CTT framework; the current study utilised Rasch analysis, adhering more closely to a major principle of objective measurement: that a valid measurement should measure only one trait at one time (Bond & Fox, 2007). Rasch analysis ensures that all items in a scale measure the same construct by providing mechanisms such as item fit statistics and DIF analysis to identify items that do not act consistently with other items and measure different dimensions. The remaining items, 7 for the Ease of Use in the OSM Environment scale and 6 for the Acceptance of OSM *s*cale, satisfied the prior requirements of the Rasch model and showed sufficiently good psychometric properties for use in future studies. The finalised scales are presented

in Appendix 2. A note of caution must, however, be offered to the effect that the misfitting or DIF items were removed not because they had nothing to do with ease of use or acceptance of OSM, but because they did not behave – in a measurement sense – in the same way as the remaining items. Such an interpretation of the data should be taken into account in further scale development.

From a methodology perspective, the current study provided an example of achieving objective measurement based on traditional Likert-type raw scores. Unlike true score statistical techniques and more general IRT models that adopt a "the model fits the data" approach and manipulate different parameters to accommodate the idiosyncrasies of a given data set, the Rasch model requires that "the data fit the model" (Andrich, 2004) and offers a list of indices to check whether this goal has been achieved. This is one of the key differences between Rasch-based studies and other quantitative studies in the human sciences (Yan & Bond, 2011), and it embraces an approach which can be applied to any test or survey data in studies which aim at objective measurement.

In the current sample, most markers tended to have low scores on the Ease of Use in the OSM Environment scale items, with the mean estimate of marker measure (–1.79 logits) being much lower than the mean estimate of item difficulty (zero), indicating that, in general, markers had quite a high level of perceived ease of use. This finding echoes that of previous studies (e.g., Coniam, 2009b, 2011) where markers perceived themselves technically competent regarding OSM. No significant differences were found between markers marking short-answer questions and those marking extended essays. However, markers marking in English language had a significantly higher level of perceived ease of use ($p < .05$) than markers marking in Chinese language, and markers for the English subject stated that they had a significantly higher level of perceived ease of use ($p < .05$) than markers for the Chinese subject. A possible reason put forward by them was that there are approximately six or seven Chinese language input systems currently in use in Hong Kong, with several Chinese input methods provided by the HKEAA as part of the OSM system. It may well be the case, however, that the proliferation of different input systems for Chinese is a factor in the lower level of perceived ease of use. This issue is one that the HKEAA will need to consider when OSM system upgrades are being looked at. It may be necessary for the HKEAA, after appropriate discussion and consultation, to decide on a system that the majority prefer and, possibly, to provide extra training for those who find the decided-upon system alien.

The mean estimate of markers' acceptance of OSM (–0.44 logits) was close to the mean estimate of item difficulty (zero). It was revealed that the overall acceptance of OSM was positive, although a small number of markers were negative about OSM. Results indicated that markers marking in English exhibited a higher level of acceptance of OSM ($p < .01$) than those marking in Chinese; markers for Chinese/English marking short-answer questions exhibited significantly higher acceptance of OSM ($p < .01$) than markers for LS rating extended essays. This is possibly due to the time frame of the different studies. With the passing of time, markers have become more accepting of and more positive towards OSM.

It was mentioned earlier that in 2012, HKDSE examinations for the core and major elective subjects were marked on screen. In some subjects, papers consist of extended essays; in other subjects, short-answer questions; and in others a mixture of both. After marking, markers of all papers in all subjects completed a post-marking questionnaire similar to that presented in Appendix 1. Consequently, it is now felt that the two calibrated scales presented in this article provide a robust basis for further investigation across all papers in all subjects to give a comprehensive picture of OSM in a jurisdiction where all

examinations are being marked on screen. The current study is therefore the stepping stone in the analysis of the final piece of the Hong Kong OSM jigsaw puzzle, which, it is anticipated, will give a definitive picture of marker acceptance of OSM for policy makers worldwide.

Acknowledgements

We would like to thank the Hong Kong Examinations and Assessment Authority – and in particular Christina Lee, the General Manager for Assessment Development – for support on the project regarding access to markers and for data collection.

Notes

1. "Person" (as in "person measure", "person reliability", etc.) is a commonly used term in Rasch analysis for "subjects", "test takers", "raters", and so forth. In the current article, it refers to "markers".
2. The page numbers for the Coniam 2013 reference were not assigned at the time of writing.

Notes on contributors

Zi Yan is an Assistant Professor in the Department of Curriculum and Instruction at The Hong Kong Institute of Education. His main research interests are in Rasch measurement, scale development, and large-scale assessment.

David Coniam is a Chair Professor in the Department of Curriculum and Instruction at The Hong Kong Institute of Education, where he is a teacher educator, working with teachers in Hong Kong secondary schools. His main publication and research interests are in language assessment, language teaching methodology, and corpus linguistics.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*(Supplement 1), 1–7.

Anshel, M. H., Weatherby, N. L., Kang, M., & Watson, T. (2009). Rasch calibration of a unidimensional perfectionism inventory for sport. *Psychology of Sport and Exercise*, *10*, 210–216.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Bos, W., Goy, M., Howie, S. J., Kupari, P., & Wendt, H. (2011). Editorial: Rasch measurement in educational contexts Special issue 2: Applications of Rasch measurement in large-scale assessments. *Educational Research and Evaluation*, *17*, 413–417.

Chan, A. H. S., So, J. C. Y., & Tsang, S. N. H. (2011). Developing optimum interface design for on-screen Chinese proofreading tasks. In *Proceedings of the 1st international conference on Human interface and the management of information: Interacting with information - Volume Part II* (pp. 3–10). Berlin, Germany: Springer-Verlag.

Chen, C.-H., & Chien, Y.-H. (2005). Effect of dynamic display and speed of display movement on reading Chinese text presented on a small screen. *Perceptual and Motor Skills*, *100*, 865–873.

Coniam, D. (2009a). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation*, *15*, 243–263.

Coniam, D. (2009b). Validating onscreen marking in Hong Kong. *Asia Pacific Education Review*, *11*, 423–431. http://www.springerlink.com/content/l6663m43jn501317/

Coniam, D. (2011). A qualitative examination of the attitudes of Liberal Studies markers towards onscreen marking. *British Journal of Educational Technology*, *42*, 1042–1054.

Coniam, D. (2013). The increasing acceptance of onscreen marking – the "tablet computer" effect. *Journal of Educational Technology & Society, 16*(3).[2]

Coniam, D., & Yeung, S.-C. A. (2010). Markers' perceptions regarding the onscreen marking of Liberal Studies in the Hong Kong public examination system. *Asia Pacific Journal of Education*, *30*, 249–271.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*, 319–340.

Fowles, D., & Adams, C. (2005, September). *How does assessment differ when e-marking replaces paper-based marking*? Paper presented at the 31st International Association for Educational Assessment Conference, Abuja, Nigeria.

Hart, D. L., & Wright, B. D. (2002). Development of an index of physical functional health status in rehabilitation. *Archives of Physical Medicine and Rehabilitation*, *83*, 655–665.

Johnson, M., Nádas, R., & Bell, J. F. (2009). Marking essays on screen: An investigation into the reliability of marking extended subjective texts. *British Journal of Educational Technology*, *41*, 814–826.

Legislative Council Panel on Education. (2005). *Grant to support the modernization and development of the Hong Kong Examinations and Assessment Authority's examination systems* (LC Paper No. CB(2)323/05-06(01)). Retrieved from http://www.legco.gov.hk/yr05-06/english/panels/ed/papers/ed1114cb2-323-1e.pdf

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85–106.

Linacre, J. M. (2006). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2011). *WINSTEPS: Rasch measurement computer program*. Chicago, IL: Winsteps.com.

Muís, K. R., Winne, P. H., & Edwards, O. V. (2009). Modern psychometrics for assessing achievement goal orientation: A Rasch analysis. *British Journal of Educational Psychology*, *79*, 547–576.

Powers, D., Kubota, M., Bentley, J., Farnum, M., Swartz, R., & Willard, A. (1997). *A pilot test of on-line essay scoring* (ETS Report RM-97-07). Princeton, NJ: Educational Testing Service.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago, IL: University of Chicago Press.

Schulz, W., & Fraillon, J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation*, *17*, 447–464.

Shaw, S. (2008). Essay marking on-screen: Implications for assessment validity. *E–Learning*, *5*(3), 256–274.

Shaw, S., & Imam, H. (2008, September). *On-screen essay marking reliability: Towards an understanding of marker assessment behaviour*. Paper presented at the International Association for Educational Assessment Conference, Cambridge, UK.

Törmäkangas, K. (2011). Advantages of the Rasch measurement model in analysing educational tests: An applicator's reflection. *Educational Research and Evaluation*, *17*, 307–320.

Twing, J., Nichols, P., & Harrison, I. (2003, June). *The comparability of paper-based and image-based marking of a high-stakes, large-scale writing assessment*. Paper presented at the International Association for Educational Assessment Conference, Manchester, UK.

Verhelst, N. D., & Glas, C. A. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 215–237). New York, NY: Springer.

Wang, W. C. (2000). Modeling effects of differential item functioning in polytomous items. *Journal of Applied Measurement*, *1*, 63–82.

Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Wendt, H., Bos, W., & Goy, M. (2011). On applications of Rasch models in international comparative large-scale assessments: A historical review. *Educational Research and Evaluation*, *17*, 419–446.

Wright, B. D. (1992). IRT in the 1990s: Which models work best? *Rasch Measurement Transactions*, *6*, 196–200.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45.

Yan, Z., & Bond, T. G. (2011). Developing a Rasch measurement physical fitness scale for Hong Kong primary school-aged students. *Measurement in Physical Education and Exercise Science*, *15*, 182–203.

Yen, N.-S., Tsai, J.-L., Chen, P.-L., Lin, H.-Y., & Chen, A. L. P. (2011). Effects of typographic variables on eye-movement measures in reading Chinese from a screen. *Behaviour & Information Technology*, *30*, 797–808.

Zhang, Y., Powers, D., Wright, W., & Morgan, R. (2003). *Applying the online scoring network (OSN) to advanced program placement program (AP) Tests* (ETS Research Report RR-03-12). Princeton, NJ: Educational Testing Service.
Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.
Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*, 1–28.

Appendix 1. 2012 HKALE/HKDSE – On Screen Post-Marking Questionnaire

Now that you have marked, please fill in the questionnaire below – reflecting upon your experience of the on screen marking process. All information collected is for research purposes only, will be kept in the strictest confidence, and will not be released to any other party**.**

Section 1: Personal and Marking Details
Please enter all your details and responses on the attached computer sheet. Fill in the ovals, or enter detail as appropriate.

01. Name:	02. Marker number:	
03. HKID number:		
04. Subject Marking Panel:	05. Paper:	
06. I marked scripts in ….. 1 English	2 Chinese	3 both English and Chinese

Section 2: Computer issues

07. How comfortable was the marking area in the assessment centre (general ambience, space, lighting, air-con, etc.)?	very comfortable	1	2	3	4	5	6	very uncomfortable
08. How would you rate your computer proficiency?	very proficient	1	2	3	4	5	6	not proficient at all
09. Was the desktop at the right height for you?	exactly right	1	2	3	4	5	6	very bad indeed
10. How was screen resolution?	very good	1	2	3	4	5	6	very poor
11. How comfortable were you reading off the screen?	very comfortable	1	2	3	4	5	6	not comfortable at all
12. How easily could you enlarge the screen image?	very easily	1	2	3	4	5	6	with much difficulty
13. How easily could you scroll the screen image?	very easily	1	2	3	4	5	6	with much difficulty
14. How easily could you input marks using the mouse?	very easily	1	2	3	4	5	6	with much difficulty
15. How easily could you input marks using the keyboard?	very easily	1	2	3	4	5	6	with much difficulty
16. What is your preference for mark input?	mouse	1	2	3	4	5	6	keyboard

Section 3: Your On screen Marking (OSM) Experience

17. How much training did you receive?	too much	1	2	3	4	5	6	too little
18. How tired did your eyes get marking on screen?	not tired at all	1	2	3	4	5	6	very tired
19. How often did you need to take a break while marking on screen?	never	1	2	3	4	5	6	very frequently
20. How much pressure did you feel, knowing that your marking performance was being constantly monitored?	no pressure at all	1	2	3	4	5	6	a lot of pressure
21. How much support and feedback did you receive from the OSM system?	a great deal	1	2	3	4	5	6	none at all
22. How helpful did you find the support and feedback from the OSM system?	very helpful	1	2	3	4	5	6	not helpful at all
23. Overall, how would you rate your on screen marking experience?	very good	1	2	3	4	5	6	very bad
24. How do you now feel about the move from paper-based to on screen marking?	a good move	1	2	3	4	5	6	a bad move
25. Would you prefer to mark on screen or on paper?	on screen	1	2	3	4	5	6	on paper

Would you be available for a short follow-up interview? (arranged at a location and time that suits you)	If YES, leave your mobile number: and your email:

Appendix 2. Finalised instrument

Ease of Use in the OSM Environment scale

01. How comfortable was the marking area in the assessment centre (general ambience, space, lighting, air-con, etc.)?
02. How would you rate your computer proficiency?
03. Was the desktop at the right height for you?
04. How was screen resolution?
05. How comfortable were you reading off the screen?
06. How easily could you enlarge the screen image?
07. How easily could you scroll the screen image?

Acceptance of OSM scale

01. How tired did your eyes get marking on screen?
02. How much pressure did you feel, knowing that your marking performance was being constantly monitored?
03. How helpful did you find the support and feedback from the OSM system?
04. Overall, how would you rate your on screen marking experience?
05. How do you now feel about the move from paper-based to on screen marking?
06. Would you prefer to mark on screen or on paper?