IAEA Brisbane 2009

Annotating essays on-screen: the influence of reading environment on annotative practice and assessor comprehension building

Stuart Shaw and Martin Johnson, Cambridge Assessment

Abstract

This paper considers how annotation practices influence the cognitive processes of assessors whilst they make judgements about the qualities of written compositions. Many academic tests use written evidence as an indicator of performance, therefore making it important to consider the role of assessors' comprehension building when reading candidates' textual responses.

Some literature suggests that reader annotation practices might perform an important function in mediating reader workload and enhancing comprehension, although relatively little theoretical or empirical work exists to shed light on such functions in the context of educational assessment practices. In contrast to discussions about the formal role of annotations in the accountability structures of large scale educational assessment, this paper discusses the role of annotation practices on comprehension building in the context of essay marking, suggesting that effective annotation links to the flexible characteristics of annotating practices making it able to respond to the varying features of a reading environment.

The paper then draws on empirical literature to suggest that some reading environments might hinder flexible annotation practices, potentially compromising reader comprehension and therefore undermining the validity of assessments of written composition.

Introduction

In assessment research the study of examiner annotation practices has been largely overlooked. This is perhaps surprising given the clear connections between annotating and reading activities that is reported in other literature (Anderson and Armbruster, 1982; O'Hara, 1996). This lack of research might link to the fact that in some formal examination systems annotating serves a clearly prescribed transmissive function. This certainly appears to be the case in England, Wales and Northern Ireland where official regulatory documents have tended to emphasise the requirement for examiners to use annotation as a tool to communicate the rationale for their judgments to other examiners or stakeholders involved in the assessment process (QCA, 2005; 2007). This paper reflects a growing interest in another crucial function of annotation. Crisp and Johnson (2007) found that annotation had an important cognitive dimension, that of facilitating examiners' thinking processes whilst marking. A corollary of this function was that annotating also tended to be a highly individualistic act; reflecting the complex process of comprehension building that occurs when a reader actively engages with a text. From this perspective annotations represent outwardly tangible evidence of the internal comprehension building processes that occur when a reader brings their existing understanding to a text.

The recognition that annotations have a cognitive as well as a communicative function has potentially significant contemporary implications for educational assessment. This is because recent technological developments carry with them the potential to influence a variety of traditional assessment practices. Digital technology affords the opportunity to transfer considerable amounts of information between individuals very efficiently. Agencies involved in large scale assessment are increasingly employing such technology to deliver digital images of exam scripts to examiners for marking, potentially leading to a change in the ways that examiners interact with those scripts when reading and marking them.

Reinking, Labbo and McKenna, (2000) caution that comparing working practices across technological boundaries might be too simplistic since conceptualising new technology in the same frame as old technology leads to expectations that tasks should be conducted in similar ways with similar strategies employed. They go on to argue that 'seeing new technologies through the lens of existing conceptions…is a natural first, yet transient, stage of development' (p 112) and that once new technology is 'assimilated' comparative studies which are based on old conceptualisations of technology will appear naïve. On the other hand, we would argue that a clear understanding of the practices that reside within the context of 'old technology' is also required in order to be able to situate the practices that evolve within a new technological environment and evaluate their true value.

The arguments in this paper are structured around four hypotheses. In order to illuminate the processes involved when assessors judge protracted textual information this paper focuses on interactive models of reading comprehension. The first hypothesis is that reader annotation might support comprehension building through a link with the monitor mechanism during reading activity. The second hypothesis is that annotating behaviours can be systematically influenced by the environment in which reading occurs. The third hypothesis suggests that one such environmental influence is the mode (i.e. either screen or paper) in which reading occurs. The final hypothesis is that readers tend to annotate less when reading from screen compared with paper. The implications of these hypotheses are then discussed in terms of potential threats to the validity of the assessment process.

Hypothesis 1: Annotation supports assessor comprehension building

Johnson and Shaw (2008) assert that the cognitive processes involved in reading can play a significant role in assessment judgements. They consider how annotation serves to influence reader comprehension building at an informal personal level whilst simultaneously fulfilling other more formal functions within assessment processes. A critical link is hypothesized between annotating and reading activities. An important aspect of this relationship is associated with reader comprehension building. Annotating activity is essentially an informal and potentially highly individualistic activity, influenced by the interaction of a variety of particular factors at a given time (Crisp and Johnson, 2007; Johnson and Shaw, 2008; Shaw 2008a; 2008b). They contend that the medium in which a text is presented can systematically influence reading processes to reduce comprehension through interfering with annotating practices.

Johnson and Shaw (2008) outline a model of reading comprehension which is currently shared by researchers in the fields of psycholinguistics, cognitive psychology, and language assessment and which applies to both first language (L1) and second language (L2) reading ability. The Interactive model of reading comprehension is a synthesis of *Bottom-up* and *Top-down* processing (Weir and Khalifa, 2008). *Bottom-up* processing portrays proficient readers as those who process a written text by working their way up the scale of linguistic units starting with identification of letters, then words, then sentences and finally text meaning, with the requirement that lower-level units should be mastered before higher-level ones can be acquired or developed. *Top-down* processing describes how comprehension takes place when readers integrate incoming information with their existing 'schemata' (their knowledge structures). Meaning is constructed as the readers integrate what is in the text and what they already have, that is, their existing linguistic, content and cultural knowledge. In this case, higher-level processes direct the flow of information through the lower levels: readers make use of their background knowledge to predict what lies ahead in the text.

Interactive models of reading comprehension expect both directions of processing to proceed simultaneously as well as to interact and influence each other (Hudson 1991, p83). In this model, reading involves the simultaneous application of elements such as context and purpose along with knowledge of grammar, content, vocabulary, discourse conventions, graphemic knowledge, and metacognitive awareness in order to develop an appropriate meaning.

Within the context of interactive reading models Weir and Khalifa (2008) present a cognitive processing approach to defining reading comprehension which builds on the earlier work of Kintsch and van Dijk (1978, 1983), Just and Carpenter (1980, 1987) and Urquhart and Weir

(1998). The approach comprises three key features: the *goal setter*; the *processing core*, and the *monitor*.

- The overall purpose or goal of reading activity is determined by the *goal setter* which also identifies and selects the most suitable strategy for reading which is most likely to realise that goal.
- The *central processing* core characterises a sequence of reading behaviours which represent successively higher-order levels of processing:
    - First Level: *Visual recognition* which entails word recognition and lexical decoding.
    - Second Level: S*yntactic parsing* concerned with the assembling of words into larger textual units thereby establishing propositional (core) meaning at clause and sentence level.
    - Third Level: *Inferencing* where additional information is brought to the text by the reader in an effort to make the text more meaningful.
    - Final Level: *Creating a text-level structure* in which a discourse-level structure is constructed for the entire text.
- The *monitor* is a mechanism which provides the reader with feedback regarding the efficacy of the selected reading process.

Research has identified a range of written support activities commonly related to reading (Anderson and Armbruster, 1982; O'Hara, 1996). These activities are habitually generated by the reader, often operating concurrently and being seamlessly integrated with reading activity.. Annotation has been identified as an activity which supports reading comprehension (Bramley and Pollitt, 1996; O'Hara and Sellen, 1997; Marshall, 2001).

In the context of annotation practices, the central processing core in the interactive processing model appears to be of specific interest. The construction of a mental model of a text involves the integration of visual textual information with the readers' world knowledge involves self-monitoring, and this necessarily entails the use of working memory. Annotation involves the active integration of a reader's present understanding with new information encountered within the text, and might perform an important function in mediating reader workload and enhancing comprehension.

A body of research explores how annotating facilitates textual encoding during the reading process (Hsieh et al. 2006; Hartley and Davies, 1978). Textual encoding - the basic perceptual process of converting a sensory input into a subjectively meaningful experience, plays an integral role in reading comprehension. A further crucial aspect of encoding involves spatial encoding. Reading is a spatial activity with the reader's eyes moving from one fixation location to the next to identify spatially distributed visual information and processing positional information (Piolat, Roussey & Thunin, 1997; Fischer, 1999). An implication of spatial encoding is that the act of reading involves the mental spatial tagging of ideas and concepts in a text rather than the tagging of the location of words alone. These observations reinforce Kennedy's (1992) 'spatial coding hypothesis' in that readers consider texts to behave as physical objects providing the reader with spatial code in addition to lexical information. A tangible outcome of the spatial encoding hypothesis are studies which highlight how reader information recall correlates positively with increased reader annotation (Hartley and Davies, 1978; Hartley, 1983; Khan, 1994).

Annotating might also perform an important metacognitive function during reading. Self-monitoring is a complex metacognitive operation which provides the reader with feedback about the success of their reading processes. McMahon and Dunbar (2003) suggest that annotation functions as a form of self-monitoring, a phenomenon also observed in a study by Crisp and Johnson (2007). There appears to be a link to research observations which suggest that annotations might support such a metacognitive function by aiding working memory in a retrospective manner. For example, Marshall (1997) notes that readers' annotations can be used as a visible trace of their attention and suggests that annotations serve as place markings which aid the annotator's memory. Thus annotation can function as a storage bank of information external to individual working memory. In a recent on-screen marking pilot, Shaw (2008a) notes that examiners use annotations in order to marshall their thinking: in this sense, annotations act as place markings for aiding memory. Pilot examiners suggested that

annotation " … enables me to mentally build-up credit in a long answer" and "When marking an extended response, annotation serves to form one's judgement initially; as an aide memoir to assist one in the process of marking, then to confirm professional judgements in accordance with the grade related criteria."

The limited extant empirical literature relating to the role of annotation in the context of educational assessment has a number of parallels with the wider literature. The functions of annotating outlined by Wolfe and Neuwirth (2001) appear pertinent to assessment, particularly notions that annotations can facilitate assessors' reading, eavesdrop on the insights of assessors and call attention to topics in important text passages. Crisp and Johnson (2007) found that annotating supported both individual and public functions. Studies by Bramley and Pollitt (1996) and Shaw (2005) have also highlighted the way that annotations can interact with examiner confidence.

These studies also provide evidence that annotating activities in large scale assessment systems serve multiple and unequally weighted functions. Annotation is more often referenced as a tool for satisfying accountability functions, through facilitating transparent communication, rather than for its ability to support comprehension building at the individual examiner level. Annotating is expected to play an important communicative role in the quality control process in terms of accountability, ensuring that information passes between examiners of different seniority levels in effective ways. Transparent communication between different markers is a key aspect of accountability structures in large scale examination systems; a function considered to be all the more important in expanding examinations system such as that of the UK (Williamson, 2003).

Hypothesis 2: Annotative behaviour is systematically influenced by the environment in which reading occurs

The premise underlying this paper is that readers' annotation practices can function as a very important support mechanism for comprehension building processes. Moreover, this function is highly individualistic in character; reflecting the particular context of a reading activity where a reader with a unique level of understanding interacts with a text containing specific aspects of information.

A key traditional characteristic of annotating relates to its fluidity as a practice, being both situation-responsive and relatively effortless in nature. This combination of characteristics supports comprehension through affording pertinent reflections without exacting a great deal of manual/technical effort which could potentially encroach on working memory load. We argue that effective annotating, measured in terms of the extent to which it supports reader comprehension, links closely to its flexibility and its ability to respond to the varying features of a reading environment.

A consequence of this observation is that features of a reading environment which constrain annotation practices also threaten to hinder reader comprehension processes. Constraints on annotation can be understood in two ways. An environment might restrict annotation opportunities by denying access to annotation tools. An environment might also constrain annotation through expecting readers to employ only particular annotations. Empirical evidence in support of this argument is difficult to find since there have been few studies that have looked at this aspect of reading behaviour, especially in relation to assessment. This has led us to review interview data from our own research studies to substantiate such arguments.

When asked to consider the role of annotating on their practice examiners are often very explicit about its role in their ability to make judgements about the qualities of an essay or protracted text. All of the pilot markers who participated in an on-screen study of the writing component of an international diagnostic testing service, which provides standardised assessments for mid-secondary school pupils aged around 14, similar to the Key Stage 3 tests in the UK, were emphatic in their belief that the use of annotation positively influences comprehension and supports textual understanding. Moreover, the markers noted that a paper-based marking environment enabled them to identify more closely with the candidate's answer as a physical object, an identification that is absent when screen marking.

Some English examiners in another marking mode study (Johnson and Nádas, 2009) suggested that their reduced levels of annotating led them to make less considered judgements when marking essays in one of the modes. This in turn resulted in examiners reporting reduced levels of confidence in their own judgments, for example, 'I felt I was making much more snap judgements on-screen, I think mainly because I use my own annotation as a source of thoughts for my judgement and because that was so much more limited I felt that I wasn't marking as well as I mark on paper'. A concern about judgemental confidence resonates with the findings of Bramley and Pollitt (1996) as well as additional interview data from the Crisp and Johnson study (2007). In that project a Business Studies examiner reflected on the challenges of an earlier assessment exercise, where they had tried to mark longer passages of text without annotating, 'The view that we had about marking without annotations was that we could speed up, but we were not very comfortable…because then you're relying on your memory, virtually, to actually come to an overall judgement. Without any annotation…it became very difficult, and at best a guess; impression marking…the annotation forces you to make judgements'.

Another constraint on annotation practice is where they are stipulated, leading to a lack of personal meaning for the reader. Examiners in the Johnson and Nádas (2009) study discuss the problems of using pre-specified annotations, suggesting that a lack of personal ownership of any annotations employed led to annotations functioning ineffectively. In this study a group of 12 experienced examiners marked a set of GCSE English Literature essays on screen. The marking software included a set of 10 annotations that were pre-defined as being relevant to the essay mark scheme by a senior examiner. Examiners observed 'I did use the annotations but I felt sometimes I was using them just for the sake of it and it's not what I would have written, I would have written something else, and I didn't really feel that that was a good thing', and, 'I felt I was just using ticks and question marks and they weren't going to mean anything to anybody or necessarily to me reading through it again'. The use of pre-specified annotations appears to have had a potentially important effect on some examiners, increasing the difficulty of recalling where particular information was located in texts. This view is illustrated by another examiner reflection, 'I was not so confident because I went back more because I didn't have my own annotation to guide my thoughts towards what [mark] band I would be awarding'.

Hypothesis 3: Screen and paper reading environments can influence annotating behaviours differentially

The literature suggests that the mode of reading has a systematic effect on the annotation practices of the reader. Key to understanding how the practice of reading scripts is affected by mode is the notion of affordance (Gibson, 1979). The concept of affordance recognises that the environment influences subjects' behaviour, with some environments facilitating or inhibiting certain types of activity. Gibson (1979) suggested that part of the success of human evolutionary development has been a consequence of their ability to identify and exploit the affordances of different environments. He claimed that humans perceive affordance properties of the environment in a direct and immediate way and they perceive possibilities for action, i.e. surfaces for walking, handles for pulling, space for navigation, and tools for manipulating. In the same sense, the modes of paper and computer exist as environments within which activity is carried out, and each has its own affordances.

Sellen and Harper (2002) identify and compare the affordances of paper and digital reading environments. They note that paper affords flexible navigation, cross document use, annotation while reading, and the interweaving of reading and writing. The key affordances of digital reading environments they describe as being their the ability to store and access large amounts of information, to display multimedia documents, to enable fast full-text searching, to allow quick links to related materials, and to allow content to be dynamically updated or modified. These digital affordances are potentially very important for educational assessment systems involving large numbers of individuals; not only through the efficient transfer of digital texts between markers but also through providing the opportunity for ongoing quality assurance measures to be integrated into assessment procedures through online marker training facilities (Knoch et al., 2007; Elder et al., 2007; Hamilton et al., 2001). In respect of annotation in particular, Sellen and Harper (2002) contrast the way that paper supports

annotation while reading whilst on screen such activity becomes cumbersome and difficult without altering the original. O'Hara and Sellen (1997) also allude to the relative effortlessness of annotating on paper compared with on screen. These observations are substantiated by findings from empirical educational research studies. Price and Petrie (1997) and Greatorex (2004) report mode-related influences on annotating practices, with on-screen annotations differing in quality and quantity from paper-based annotation profiles.

Literature suggests that the quality of annotations might influence their effectiveness. Note-taking research has a certain degree of overlap with annotation research since it also considers the degree to which such an activity might influence the encoding and comprehension of written or verbal information. Some literature suggests that the qualitative form of notes taken will affect cognitive activity. Bretzing and Kulhavy (1979) suggest that notes which paraphrase and summarise stimulate deeper semantic processing, whereas notes which transcribe verbatim do not. The consequence of this finding is that annotation tools which restrict the ability of a reader to freely summarise points encountered in a text might hinder deep processing. Literature also points to the effect of tool use on information processing. Kobayashi's (2005) meta-analysis of the effectiveness of note-taking techniques suggests that the mechanical demands of note-taking can hinder information encoding. This point has parallel implications for annotating, with on-screen annotation tools being sometimes more distracting than paper-based annotation tools (O'Hara and Sellen, 1997). This leads Marshall (1997) to argue that annotation should interrupt reading activity as little as possible. Shaw (2007) observed that the inability to apply a full-range of annotations on-screen plays an important role in terms of examiner self-assurance. Use of a restrictive annotative palette engendered general dissatisfaction amongst examiners.

Hypothesis 4: The influence of reading on-screen is to obstruct annotating behaviours and impede comprehension building

The narrow empirical literature suggests that the mode in which a text is presented can affect reader annotation in a number of ways. O'Hara and Sellen (1997) suggest that paper-based annotation is a moderately effortless procedure which factors automatically into the meaning construction process during reading. Computer-based annotation practices, however, can be impeded by a lack of authentic annotation tools. For example, Shaw (2007) comments on the frustration of examiners not being able to underline parts of a candidate response in early, less sophisticated software marking applications. The ability to underline parts of an answer to emphasize is especially important for composition marking particularly if the comments remain visible so that they can be easily read when scanning back over an answer. Underlining, serves as a permanent record for subsequent adjudication thus reinforcing the prevailing belief that annotation performs a communicative function between examiners. Price and Petre (1997) found that 'emphasizing' annotations were used less when marking on-screen than with paper based marking.

Price and Petre (1997), Greatorex (2004) and Shaw (2007; 2008a; 2008b) found that mode influenced some annotation practices with assessors using different annotation conventions on screen compared with paper. Shaw (2007) observed that the physical effort expended to annotate on screen compared with the seemingly effortless task of performing the same function on paper intrudes upon authentic examiner interaction with extended candidate answers. Inability to apply annotations quickly constantly engendered frustration amongst examiners despite the fact that with practice the process became easier (Price and Petre, 1997, report similar findings). Several examiners, for example, commented on the cognitive and physical constraints that annotating on-screen imposed: "It was awkward to annotate on-screen and very time-consuming … I didn't make as many as I would have done on paper e.g. little comments or errors to link faults", "I wrote very few annotations on screen, whereas I would write several on script", and, "On paper I would have added a lot more comments e.g. generous/too vague/just about ok".

Other studies suggest emotive and physical dimensions in relation to computer annotating. Greatorex (2004) reports teacher frustration when moderating electronic portfolios. Shaw (2008) observes that examiner concentration is adversely affected when assessing on-screen. Shaw notes that not being able to replicate paper and pen practice when applying annotations on screen is a predominant concern. More generally, the physical process of

selecting and applying annotations on screen has significant implications for examiner concentration as the following quotes illustrate: ”Normally I need all my concentration on a script. My right hand annotates, but I do not watch it. It is automatic, and my hand knows the symbols and where to put them. When you annotate on-screen, you lose concentration as you have to find the symbol and drag it into position. This is not an argument against annotation, which is important, but it might be against marking online”, ”It’s natural to mark/tick/highlight linguistic errors with one hand but be thinking about content/style etc with the brain. On-line marking requires more concentration on the technical aspect of annotation (or if this is left out, the overview is not as accurate in my opinion” . Shaw (2008) concludes that awkward application of the certain annotative facilities serve to protract assessment and detrimentally affect examiner attentiveness.

Further empirical study has gathered evidence about the specific characteristics of mode-related influence on annotation practices. Johnson and Nádas (2009) investigated mode-related reading for assessment behaviours by analysing 720 scripts from 12 English Literature GCSE examiners marking on screen and on paper. Their analyses showed that mode affected examiners’ annotation patterns, with examiners annotating less on screen.

When working in their usual paper environment examiners on average made 23.98 annotations per essay, compared with 18.62 annotations per essay on screen.

Table 1 Mean number of annotation types per script

Mean number of annotation types per script for each mode			
Annotation	*Meaning*	Paper	Screen
✓	*Creditworthy point*	17.89	15.41
Comment	*Free text*	3.88	Not Available
?	*Question mark*	0.49	0.67
VGD	*Very Good*	0.04	0.38
EXC	*Excellent*	0.00	0.14
SUPP	*Support for point made*	0.35	1.21
DEV	*Point developed*	0.25	0.56
NARR	*Drifting towards narrative*	0.04	0.11
REP	*Repetition*	0.05	0.09
^	*Missing information*	0.23	Not Available
[UNDERLINE]	*Drawing attention to text*	0.38	0.03
X	*Incorrect*	0.01	0.02
[ARROW]	*Linking text*	0.17	Not Available
[SIDELINE]	*Drawing attention to text*	0.22	Not Available
[CIRCLE]	*Drawing attention to text*	0.01	Not Available
Total		23.98	18.62

Table 1 illustrates some of these annotation differences, with the ‘comment’ annotation representing the greatest difference, being used on average nearly four times per paper script. These comments generally included sets of phrases directly linked to evidence found in the text for a variety of purposes, sometimes bringing together subtle reflections (e.g. “possibly”), holistic and/or tentative judgment (e.g. “could be clearer”; “this page rather better”), internal dialogue or dialogue with the candidate (e.g. “why?”), or taking note of particular features or qualities found (e.g. “context”; “clear”).

Whilst the authors acknowledge that some of the difference was accounted for by the demands of the on-screen environment which predetermined to some extent the types of annotations available, even when comparing only those annotations available for use in both modes (and therefore taking out an aspect of the potentially unfair mode-related comparison in the research design) there was still evidence of a mode-related effect. Examiners annotated more on paper (19.48 annotations per essay) compared with screen marking (18.62 annotations per essay) with some significant mode-related differences found for particular types of annotations. ANOVA analyses showed significant mode-related differences between the mean number of paper and screen annotations for four different annotation categories. “Underlining” ($F(1, 22) = 7.87$, $p = 0.01$) was used more heavily on paper whilst

"Very Good" ($F(1, 22) = 4.78$, $p = 0.04$), "Excellent" ($F(1, 22) = 4.68$, $p = 0.04$) and "Support" ($F(1, 22) = 5.28$, $p = 0.03$) annotations were used significantly more frequently on screen. T-test analyses also showed that examiners were significantly more likely to use ideographic annotations to make links between chunks of text on paper (e.g. circling and sidelining; $t(5) = 2.66$, $p < 0.05$), whereas screen annotations tended to allow examiners to label the existence of discrete qualities found in the text.

Johnson and Nádas' analysis suggested that several factors might have contributed to the relative lack of annotation on screen. Firstly, the choice of annotations available on screen was not pertinent for the examiners on some of the occasions that they wanted to apply them, and secondly, the physical and mental effort of annotating on screen was greater than on paper. Johnson and Nádas also argue that the mode-related annotation differences were symptomatic of examiners' reading behaviours being different across the modes. A variety of examiners alluded to this during interview and observation sessions; 'I still feel it was easier for me to navigate back and forwards on paper, even within the script, possibly because my handwritten comments are likely to be more individual and easier to identify a point in the script', 'I didn't have my own annotations to guide my thoughts towards what band I would be awarding and therefore…I found it more difficult to navigate my way to a band on screen.', and, '[The Examiner] says that the normal procedure at the end of a script was to make a comment and then quickly check back over the script before awarding a mark, but she didn't go back on screen because she was worried about losing the comment'.

"(2007; 2008) also notes that pilot markers tended to annotate less rigorously and less censoriously when marking on-screen; "The method of adding annotations is very time consuming and potentially causes RSI. I can only do one hour at a time. So I do reduce keystrokes as much as possible" and, "I used the annotations less …. I annotated a lot less – pressure of time … I used fewer on line".

More sparing use of annotation on screen led to more judicious selection of annotations; 'Given that the online interface had limited symbols – and swapping from one to the other is time-consuming – and RSI threatening – I tended to go for being more concise. (Many spelling errors and grammar /syntax errors were mentally noted but not indicated by available annotation icons)'.

Johnson and Nádas noted that changes in annotation patterns appeared to diminish examiners' perceptions about their marking consistency, although there was no significant evidence of any actual relationship between these factors. The nature of the relationship between examiner annotation and marking outcome was confounded by the finding that mode had very little influence on the practice of examiners writing summative comments at the end of each script. The stability of this practice across modes suggested that it might have been a key factor in the consistency of examiner judgments found across modes.

Johnson and Nádas also observed that there was some tentative evidence that the types of annotations provided could afford the opportunity to represent quality in different ways. Ideographic annotations such as underlining, circling and sidelining, were used freely on paper and unavailable on screen. On the other hand, annotations that were available in both modes and described discrete qualities (e.g. 'Good'; 'Excellent'; 'Support') were used significantly more frequently on screen. It is possible to suggest that ideographic annotations offer the potential for a reader to make effective conceptual links between chunks of textual information whilst single word annotations seem more appropriate for tagging discrete pieces of information.

Conclusion: annotating environment and its implications for comprehension building and scoring validity

By bringing together literature about linguistics and annotation practices, both empirical and theoretical, we have suggested that a critical link exists between annotating and reading activities. Moreover, an important aspect of this relationship is associated with reader comprehension building. It is perhaps significant that empirical study into annotating in assessment contexts is very limited and this helps to explain why the extent to which annotating candidate responses influence or affect assessor comprehension is neither known

nor fully understood. This is an important observation since the arguments advanced in this paper suggest that such an influence is tangible and has potential implications for validity issues.

Crisp and Johnson (2007) have suggested that one of the two functions annotations serve in educational assessment is a justificatory communicative role in the quality control process in terms of accountability; assisting with transparent communication between different markers. Accountability is widely recognised to be a multifaceted and complicated concept (Day and Klein, 1987) and 'assumes the requirement to answer to the broader social community' (Kogan, 1986). In an educational context, examination boards offering high-stakes assessments are required to account for or justify certain assessment actions and behaviour for a range of potential community stakeholders. Thus, the notion of accountability is closely related to responsibility, as those who have been given responsibilities - the assessment practitioners - are asked to account for their assessments. If the conventional accountability processes are influenced by the introduction of a new reading environment then both the reliability of test scores and the validity of the assessments are potentially compromised.

Validity can be thought of as the degree to which an assessment generates an outcome (e.g. a test score or grade) which is an accurate representation of a test taker's ability. The extent to which the inferences made on the basis of the outcome are meaningful, useful and appropriate is seen as a vital aspect of validity (AERA/APA/NCME Standards, 1985, p.9). According to this definition, validity resides in the scores on a particular administration of a test rather than in the test per se.

Annotations have a direct link with validity through the way that they connect a score, the interpretation of the score, and any ensuing actions based on such an interpretation. It appears that one key purpose of examiner annotations is to communicate how the features of a performance connect with the features of a mark scheme. In this context annotations support valid interpretations about the way that a performance has been assessed when they communicate the way that an examiner has considered their judgement in relation to a performance (Johnson and Shaw, 2009). Annotations are not only a tool for communication within a script which might reflect how an examiner has applied a mark scheme in the context of a particular performance, but they can also contain tacit features that support examiner thinking. In this context annotations have a high degree of validity, clearly communicating the way that an examiner has considered their judgement in relation to a performance.

This paper has provided evidence from a variety of sources to suggest that the medium in which reading occurs can influence reading and, in particular, annotating behaviours. We suggest that a key characteristic of paper-based annotation is its flexibility, allowing it to reflect the context-specific features of a reading episode. Besides affording an opportunity to communicate meanings between different readers (or examiners) another crucially important function is its ability to aid comprehension building during reading activity. Expanding on interactive models of reading comprehension we suggest that annotating has an important metacognitive monitoring function, enabling readers to iteratively reflect on the qualities that they identify as they move through a text. A concern expressed in this paper is that reading in digital environments leads to reduced annotating behaviours. This also implies that certain qualities are not being outwardly represented through annotation and therefore not being reinforced to the reader during comprehension building.

The accountability challenge for assessment is that the reading environment affords assessors the opportunity to consider the qualities that they perceive to be important when they encounter them, and that this facility should be the same if an examiner chooses to read a text on paper or on screen. Furthermore, if there is a stipulation that all texts should be read only on screen to overcome any potential mode-related bias then it is important that the scoring validity is equivalent on either side of the technological shift, or else there are potential issues around the emergence of differential standards within the system.

To overcome the concerns raised in this paper it is important that continuing developments in technology focus on trying to afford readers access to flexible annotation tools when engaged in reading on screen, and where possible avoiding the restriction of annotation opportunities

by denying access to annotation tools or through expecting readers to employ only particular annotations.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anderson, T. H. & Armbruster, B. B. (1982). Reader and text-studying strategies. In: W. Otto, & S. White (Eds.), *Reading Expository Material*. London: Academic Press.

Bramley, T. & Pollitt, A. (1996). Key Stage 3 English: Annotations Study. A report by the University of Cambridge Local Examinations Syndicate for the Qualifications and Curriculum Authority. London, QCA.

Bretzing, B. H. & Kulhavy, R. W. (1979) Notetaking and depth of processing. *Contemporary Educational Psychology,* **4**, 145-153.

Crisp, V. & Johnson, M. (2007). The Use of Annotations in Examination Marking: Opening a Window into Markers' Minds. *British Educational Research Journal*, **33**, 6, 943-961.

Elder, C., Barkhuizen, G., Knoch, U. & von Randow, J. (2007). Evaluating rater responses to an online training program for writing assessment. *Language Testing,* **24**, 1, 1-28.

Fischer, M. H. (1999). Memory for Word Locations in Reading. *Memory*, **7**, 1, 79-118.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Greatorex, J. (2004). *Moderated E-portfolio Project Evaluation*. Cambridge: University of Cambridge Local Examinations Syndicate.

Hamilton, J., Reddel, S. & Spratt, M. 2001: Teachers' perceptions of online rater training and monitoring . *System,* **29**, 505-520.

Hartley, J. (1983). Note-taking Research: Resetting the Scoreboard. *Bulletin of the British Psychological Society,* **36**, 13-14.

Hartley, J. & Davies, I. K. (1978). Note-taking: A Critical Review. *Innovation in Education and Teaching International*, **15**, 3, 207-224.

Hsieh, G., Wood, K. R. & Sellen, A. (2006). *Peripheral Display of Digital Handwritten Notes*. Proceedings of the Conference on Human Factors in Computing Systems, Montreal, Quebec, 2006

Hudson, T. (1991). A Content Comprehension Approach to Reading English for Science and Technology. *TESOL Quarterly*, **25**, 1, 77-104.

Johnson, M and Nádas, R. (2009). An investigation into marker reliability and some qualitative aspects of on-screen essay marking. *Research Matters*: A Cambridge Assessment Publication, Issue 8, July 2009.

Johnson, M & Shaw, S. (2009). *'The lack of annotation undermines my confidence…': Towards an understanding of the impact of annotations on returned examination scripts*. Internal Cambridge Assessment Report.

Johnson, M & Shaw, S. (2008). Annotating to comprehend: a marginalised activity? *Research Matters Issue 6*, June 2008, 19-24.

Just, M. A. & Carpenter, P. A. (1980). A Theory of Reading: From Eye Fixations to Comprehension. *Psychological review*, **87**, 4, 329-354.

Just, M. A. & Carpenter, P. A. (1987). *The Psychology of Reading and Language Comprehension.* Boston: Allyn and Bacon.

Kennedy, A. (1992). The Spatial Coding Hypothesis. In: K. Rayner (Ed.) *Eye Movements and Visual Cognition*. New York: Springer-Verlag.

Khan, F. (1994). *A Survey of Note-taking Practices*. HP Labs Technical Reports HPL-93-107.

Kintsch, W. & van Dijk, T. A. (1978). Toward a Model of Text Comprehension and Production. *Psychological Review*, **85**, 363-394.

Knoch, U. Read, J. & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing,* **12**, 1, 26-43.

Kobayashi, K. (2005) What limits the encoding effect of note-taking? A meta-analytic examination. *Contemporary Educational Psychology,* **30**, 242-262.

McMahon, M. & Dunbar, A. (2003). Mark-UP: Facilitating Reading Comprehension through On-Line Collaborative Annotation. In: N. Smythe (Ed.), *Digital Voyages.* Proceedings of the Apple University Consortium Conference, Adelaide, South Australia, September 28 – October 1, 2003.

Marshall, C. C. (1997). *Annotation: From Paper Books to the Digital Library*. Proceedings of the Second ACM International Conference on Digital Libraries; Philadelphia, Pennsylvania.

Marshall, C.C. (1998) Toward an ecology of hypertext annotation, Proceedings of HyperText '98 (Pittsburgh, PA, USA, June 1998), ACM Press, 40-48.
Marshall, C.C. (2001). *The Haunting Question of Intelligibility*. Paper presented at the ACM Conference on Hypertext and Hypermedia, Aarhus, August 14-18, 2001.
O'Hara, K. (1996). *Towards a Typology of Reading Goals.* Rank Xerox Research Centre Affordances of Paper Project Technical Report EPC-1996-107. Cambridge: Rank Xerox Research Centre.
O'Hara, K. & Sellen, A. (1997). A Comparison of Reading Paper and Online Documents. In: S. Pemberton (Ed.), *Proceedings of the Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery.
Piolat, A., Roussey, J-Y. & Thunin, O. (1997). Effects of Screen Presentation on Text Reading and Revising. *International Journal of Human-Computer Studies*, **47**, 565-89.
Price, B. & Petre, M. (1997). *Teaching Programming through Paperless Assignments: An Empirical Evaluation of Instructor Feedback.* Milton Keynes: Centre for Informatics Education Research, Open University.
QCA (2007). *GCSE, GCE, GNVQ and AEA Code of Practice*. London: QCA.
QCA (2005). *A Review of GCE and CSE Coursework Arrangements*. London: QCA.
Reinking, D., Labbo, L. D. & McKenna, M. C. (2000) From assimilation to accommodation: a developmental framework for integrating digital technologies into literacy research and instruction. *Journal of Research in Reading* 23 (2) pp 110-22
Sellen, A. & Harper, R. (2002) *The Myth of the Paperless Office.* Cambridge, MA: MIT Press
Shaw, S. (2005). *On-screen Marking: Investigating the Examiners' Experience through Verbal Protocol Analysis*. Internal ESOL Validation and Research Report.
Shaw, S. (2007). On-screen Marking of Extended Writing: towards an understanding of examiner on-line assessment behaviour. Internal Cambridge International Examinations Internal Research Report.
Shaw, S. (2008a). Marking Essays on screen: towards an understanding of examiner assessment behaviour. *Research Matters Issue 6*, June 2008, 9-15.
Shaw, S. (2008b). Essay Marking On-Screen: implications for assessment validity. *E–Learning*. 5(3), 256-274.
Urquhart, A. H. & Weir, C. J. (1998). *Reading in a Second Language: Process, Product and Practice*. London: Longman.
Weir, C. J. & Khalifa, H. (2008). A Cognition Processing Approach Towards Defining Reading Comprehension. *Research Notes 31*, March 2008, 4-16.
Williamson, P. (2003). Setting, Marking and Awarding: The Examination Process. In: K. Tattershall, J. Day, H. James, D. Gillan & A. Spencer (Eds.), *Setting the Standard.* Manchester: AQA.
Wolfe, J. L. (2000). Effects of annotation on student readers and writers. JCDL '00, San Antonio, Texas, 19-26.

Wolfe, J. L. & Neuwirth, C. M. (2001). From the Margins to the Center: The Future of Annotation. *Journal of Business and Technical Communication,* **15**, 3, 333-371.