

An examination of the design and implementation logistics of electronic reading assessment in PISA 2009 study: The Macao experiences

Pou-seong Sit & Kwok-cheung Cheung
University of Macau
Macao, China

Paper presented at the PISA Research Conference, Kiel, Germany, 14-16 September, 2009.

Abstract Macao, a special administrative region of People's Republic of China, participated in PISA 2009 reading literacy study. In this study, Electronic Reading Assessment (ERA) is regarded as an innovation in comparative assessment research (Cheung & Sit, 2008). Capitalizing on Macao's experiences in the field trial and main survey of PISA 2009 Study, this paper examines the design and implementation logistics of ERA. The paper consists of three main sections: (1) An explication of the design of ERA test delivery system; (2) An elucidation of task characteristics and item statistics of released ERA test units, and (3) A discussion of issues of reliability and validity warranted attention in ERA by Macau-PISA Centre. This paper ends with a summary of eight pertinent issues, the corresponding guidelines of which are concluded as having been fulfilled satisfactorily by the PISA Consortium.

Keywords: reading literacy, computer-based assessment, PISA

1. An explication of the design of ERA test delivery system

Reading literacy is “*understanding, using and reflecting on and engaging with written texts, in order to achieve one's goals, to develop one's knowledge and potential and to participate in society*” (OECD, 2007a, p.14). This formal definition is needed in the design of ERA test delivery system. Compared with print-based reading literacy, which is envisaged as a major assessment domain in PISA 2009, ERA is treated as a minor domain. Unlike print-based assessment of reading, mathematical and scientific literacy, ERA is conceived of and designed as a form of computer-based assessment. An online web-based reading environment is simulated within which both online reading literacy and information and communication technology (ICT) literacy are the foci of assessment.

Starting with the above formal definition of reading literacy and capitalizing on the advances of ICT in the development of CBA, four components for the delivery of online reading literacy assessment are delineated, namely: (1) Assessment generation; (2) Assessment delivery; (3) Assessment scoring and interpretation; and (4) Storage, retrieval and transmission (Professional Practice Board Steering Committee on Test Standards, 2002). Figure 1 below shows schematically what these four operative components entail.

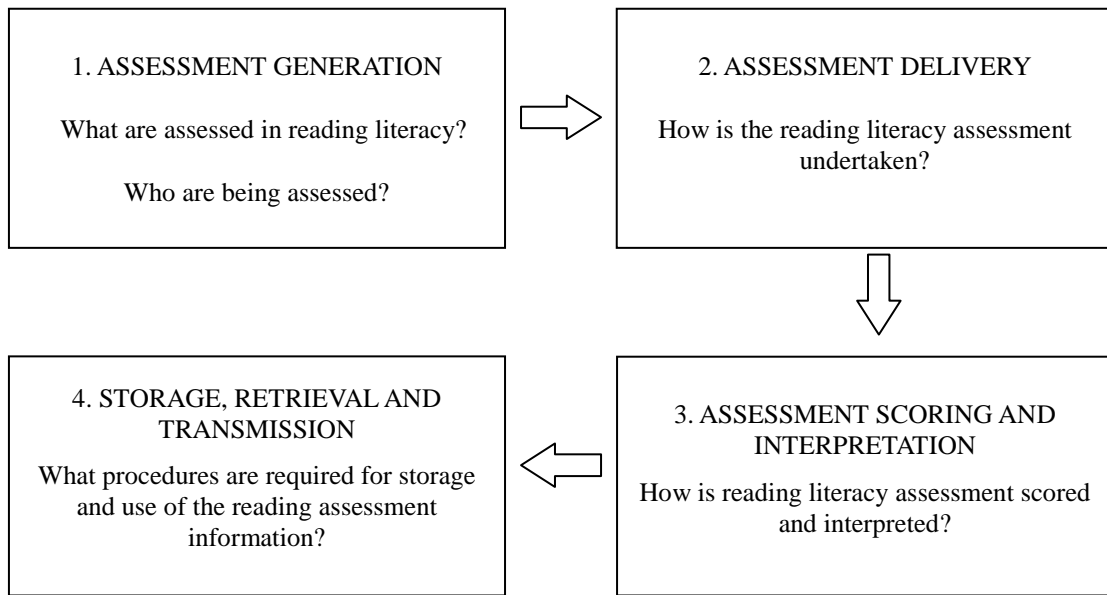


Figure 1 The four operative components for the delivery of online reading literacy assessment

(1) Assessment generation

The first operative component concerns who are being assessed and what are assessed in reading literacy. ERA assessment framework incorporates the reading assessment framework used by countries to assess their 15-year-old students' reading literacy not only in the print medium but also online electronically. In this framework, five aspects of reading literacy are delineated: (1) retrieve information; (2) form a broad understanding; (3) develop an interpretation; (4) reflect on and evaluate context of text; and (5) reflect on and evaluate form of text (OECD, 2007b).

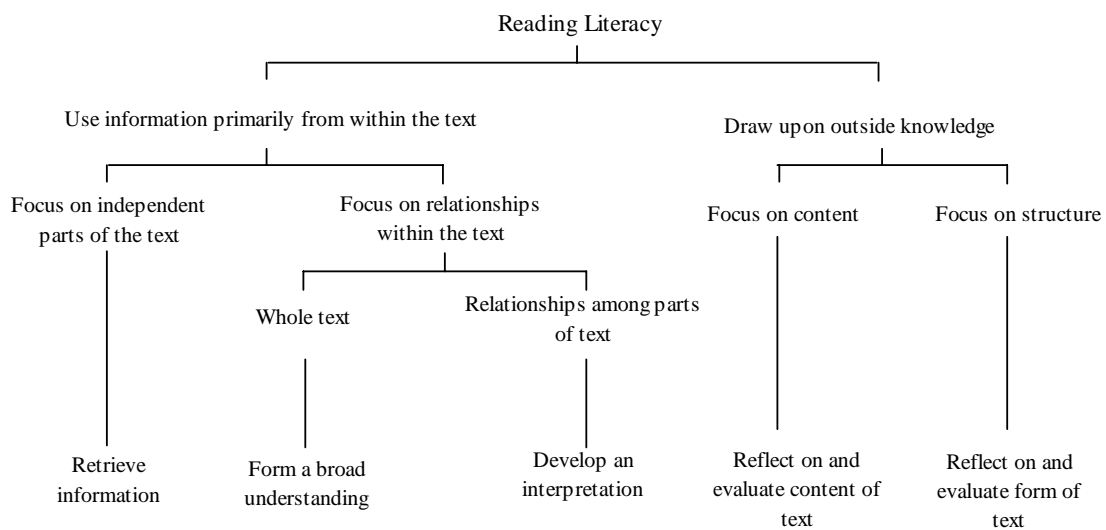


Figure 2 The five aspects of reading literacy (OECD, 2007b)

With regard to the relationships amongst tasks, texts and reading aspects in the print medium and electronic medium, there are important differences. Figure 3 and Figure 4 make a contrast between fixed texts in the print medium (e.g. books and magazines) with dynamic texts in the electronic medium (e.g. multiple texts in the form of linked webpage with navigation tools and features).

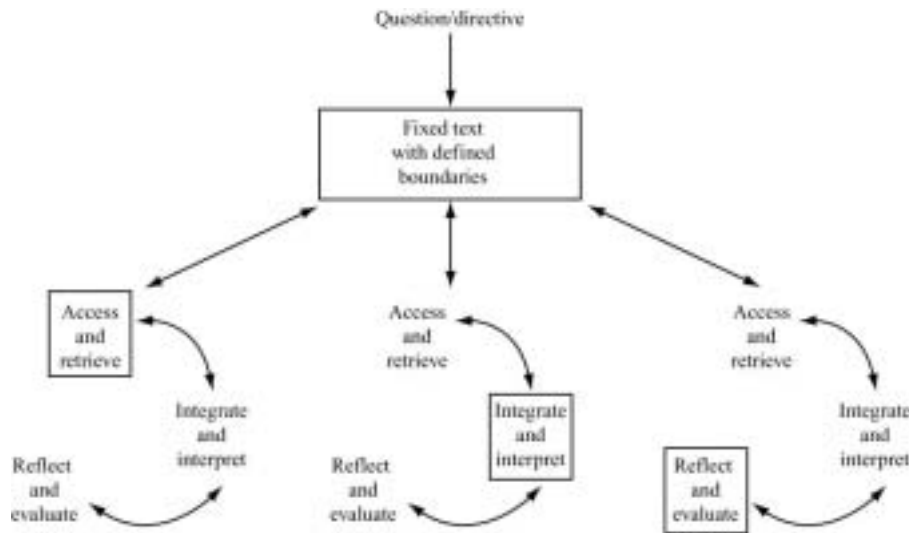


Figure 3 Relationships amongst tasks, texts and reading aspects in the print medium (OECD, 2007b & 2007c)

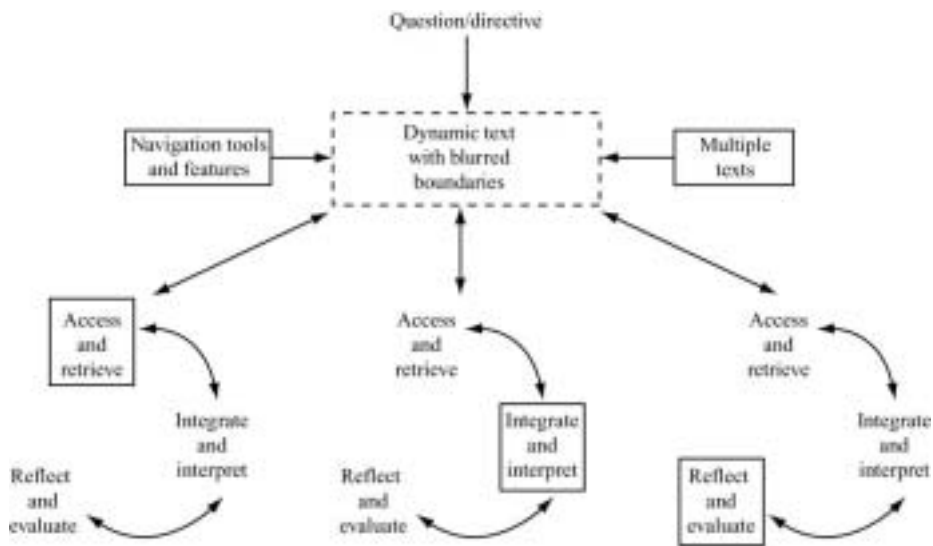


Figure 4 Relationships amongst tasks, texts and reading aspects in the electronic medium (OECD, 2007b & 2007c)

In addition, Table 1 summarizes the comparison results of the reading aspects between the print and electronic medium (see OECD, 2007c for details). As analyzed, there are important differences to justify ERA as an important add-on to the print reading literacy for 15-year-olds in the digital age – a time when ICT is increasingly envisaged as an important literacy that may be deployed in students’ online reading activities.

Table 1 Comparison of print and electronic reading in terms of the reading aspects of a test item

Reading Aspect	Print Reading	Electronic Reading
1. Access and Retrieve	<p>Orient and navigate in concrete information space, <i>e.g. Go to library, search in a catalogue, find a book</i></p> <p>→ Use navigation tools and structures, <i>e.g. Table of contents; page numbers; glossary</i></p> <p>→ Select and sequence information (low reader control; one sequence of linear reading)</p>	<p>Orient and navigate in abstract information space, <i>e.g. Enter URL; Google</i></p> <p>→ Use navigation tools and structures, <i>e.g. Menus; embedded hyperlinks</i></p> <p>→ Select and sequence information (high reader control; multiple sequences of linear reading)</p>
2. Integrate and Interpret	<p>Integrate at a lower level of demand: larger portions of text are simultaneously visible (one or two pages)</p> <p>→ Develop an interpretation</p> <p>→ Form a broad understanding</p>	<p>Integrate at a higher level of demand: limited parts of text are simultaneously visible (limited by screen size)</p> <p>→ Develop an interpretation</p> <p>→ Form a broad understanding</p>
3. Reflect and Evaluate	<p>Pre-evaluate information, <i>e.g. use table of contents; skim passages, checking for credibility and usefulness</i></p> <p>→ Evaluate credibility of source (<i>usually less important due to filtering and pre-selection in the publishing process</i>)</p> <p>→ Evaluate plausibility of content</p> <p>→ Evaluate coherence and consistency</p> <p>→ Hypothesize</p> <p>→ Reflect in relation to personal experience</p>	<p>Pre-evaluate information, <i>e.g. use menus; skim web pages, checking for credibility and usefulness</i></p> <p>→ Evaluate credibility of source (<i>usually more important due to lack of filtering and pre-selection in open environment</i>)</p> <p>→ Evaluate plausibility of content</p> <p>→ Evaluate coherence and consistency</p> <p>→ Hypothesize</p> <p>→ Reflect in relation to personal experience</p>
4. Complex	<p>The range of sources to be consulted is relatively undefined. The sequence of steps within the task is undirected, <i>e.g. finding, evaluating and integrating information from multiple printed texts</i></p>	<p>The range of sources to be consulted is relatively undefined. The sequence of steps within the task is undirected, <i>e.g. finding, evaluating and integrating information from multiple electronic texts</i></p>

Note: Adapted from OECD (2007c).

PISA assessment tasks (or items) focusing on same or different reading aspects are typically assembled and packaged in the form of test units (see Appendix 1-7 for examples). During test development, each test unit is categorized by four characteristics: (1) situation; (2) environment; (3) text type; and (4) item format. Table 2 compares print reading with electronic reading in terms of these four characteristics of test units (see also Cheung & Sit, 2008a, p.3-8, for a delineation of characteristics of test units in print assessment). Table 3 summarizes the percentage of different categories of ERA tasks selected in the PISA 2009 Main Survey. This distribution of assessment tasks has a bearing on the interpretation and use of the composed ERA constructs for cross-country comparisons.

Table 2 Comparison of print and electronic reading according to characteristics of the test units

Characteristics of the test units	Print Reading	Electronic Reading
1. Situation	Personal; Public; Occupational; Educational	Personal; Public; Occupational; Educational
2. Environment	None	Authored sites; Message based sites
3. Text Type	Argumentation; Description; Exposition; Narration; Instruction	Argumentation; Description; Exposition; Transaction
4. Item Format	Continuous; Non-continuous; Mixed; [Multiple]	[Continuous]; Non-continuous; Mixed; Multiple

Note: Texts in square brackets indicate that this feature is given relatively little emphasis in the PISA reading assessment framework

Table 3 Percentage of classification of ERA tasks in PISA 2009 Main Survey

Characteristics of assessment task	%	
1. Situation	Personal	32.1
	Public	44.6
	Occupational	10.7
	Educational	12.5
2. Environment	Authored sites	67.9
	Message-based sites	25.0
	Mixed	7.1
3. Text Type	Argumentation	22.6
	Description	31.5
	Exposition	35.7
	Transaction	10.1
5. Item Format	Multiple Choice	64.3
	Open Constructed Response	28.6
	Complex Multiple Choice	7.1
4. Reading Aspect	Access and Retrieve	25.0
	Integrate and Interpret	39.3
	Reflect and Evaluate	21.4
	Complex	14.3

(2) Assessment delivery

The second operative component concerns how reading literacy assessment is undertaken. The TAO (French acronym for *Testing Assisté par Ordinateur*) is used to deliver ERA. It is designed as a modular assessment platform for CBA delivery and management (Cheung & Sit, 2008b; Goldhammer, et al., 2008; Latour, et al., 2008). Typically, ERA is delivered on school computers via a USB flash drive or CD-ROM. Macao used CD-ROM in the field trial but shifted to use USB in the main survey in order to increase the processing speed of the assessment system. As shown in the ERA delivery system interface shown in Figure 5, there is a time-bar at the top of the screen to show how much time in the testing session the student has left. Also, the question numbers change colour as the student has completed them. There is a *Help* button with general instructions about features of the system. Students are able to copy and paste and to use a *Find* function within a web-style page.

The TAO system, once operational, can capture time taken for each response and the pathways and links followed by the student within each task. Both test units and cluster of tasks within test units are delivered in a fixed ‘lockstep’ fashion. The examinees are not able to return to a task once they have moved to the next one. Each time a student clicks the *Next* button the examinee is informed about to move on to the next task and it is not possible to return to the previous task. This approach enables test developers to specify the starting page for each task. In this way, all students begin in the same place within the stimulus. If they have navigated through a series of less relevant pages, they do not have to find their way back to begin the task. The stimulus materials are situated in the browser area and the task in the task area. There is a variety of Chinese-input methods (e.g. Cangjie, Simplified Cangjie, and Pinyin) for use by the students to respond to open constructed response tasks. The TAO system is able to register mouse clicks made at both browser and task areas.

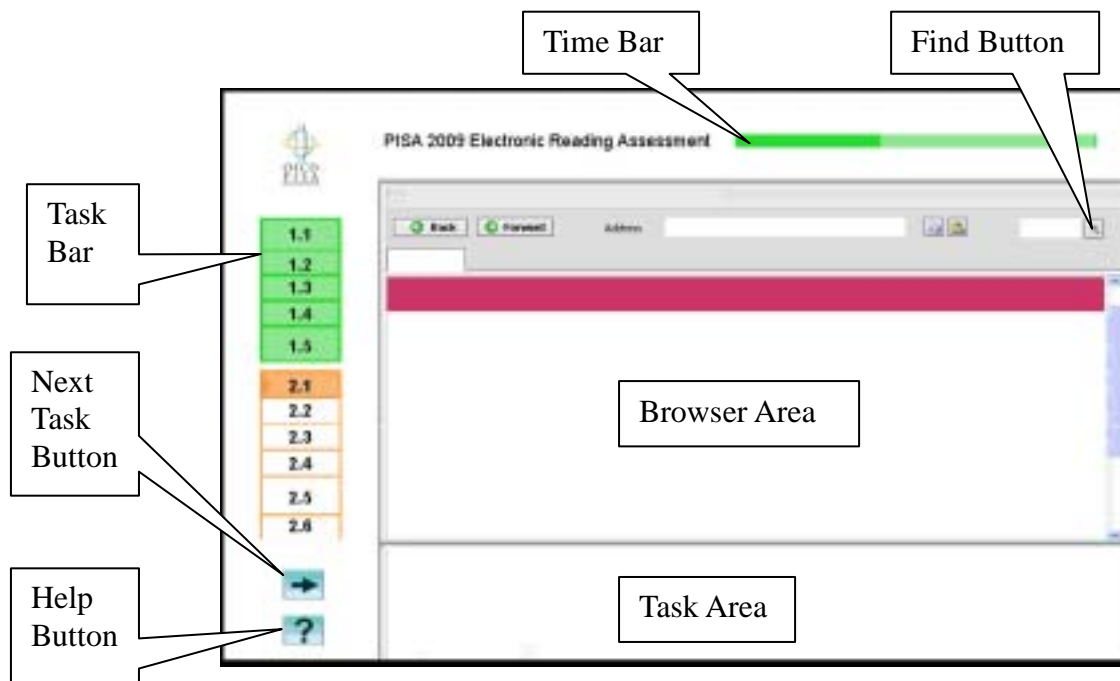


Figure 5 The ERA delivery system interface

(3) Assessment scoring and interpretation

The third operative component concerns how reading literacy assessment is scored and interpreted. The vigorous coding process is done online using Online Coding System developed by the PISA Consortium. Except for multiple choice and closed constructed response tasks the responses of which are automatically captured, all open constructed response tasks are coded in accordance with internationally agreed coding guides. To ensure consistency so as to achieve maximal reliability the tasks are coded one by one by a team of qualified coders. Appendix 1-7 details the coding guides of two released ERA test units, alongside delineation of question intent and access paths pertinent to answering the seven assessment tasks correctly (see <http://erasq.acer.edu.au> for the official website of the released ERA items). Using E022Q09 (LET'S SPEAK) as an example, students' responses are coded against the international coding guides and the associated workshop examples. Table 4 presents the coding guide of this task, illustrated with typical student responses in Macao's PISA 2009 Field Trial.

Task E022Q09 (LET'S SPEAK):

Look at Mischa's post for March 10. Click on "Write a Reply" and write a reply to Mischa. In your reply, answer her question about which writer, in your opinion, knows the most about this issue. Give a reason for your answer. [Note: use the Back button to refer to the Forum page.]

Click "Post Reply" to add your reply to the forum.

Table 4 Coding guide of E022Q09 (LET'S SPEAK), illustrated with typical student responses in Macao's PISA 2009 Field Trial

Code	International Coding Guides*	Typical student responses in Macao's PISA 2009 Field Trial (in Chinese)
1	Identifies <u>Doctor Nauckunaite and/or Psychologist O.L.</u> (explicitly or implicitly) AND refers to their <u>professional status</u> . May express skepticism about their professional status.	我覺得心理學家(Psychologist)知道得較多，如果要我相信的話，我令(寧)可相信一些有科學的事實。
	Identifies <u>any of the four writers</u> named by Mischa (Julie, Tobias, Psych OL or Dr. Nauckunaite) AND <u>gives a reason that is consistent with the text, related to the cogency, practicality or logic</u> of the text.	我認為嘉穎(Tobias)對公開演講知道最多。因為她說了一個重點，就是公開演講是需要經驗的，沒有人是一出生就懂得公開演講的技巧的。 由於人在開始演講時最緊張，克服恐懼的實際方法是背熟演辭開首部分。開始演講前，環顧觀眾。如果知道你對甚麼人說話，便會心安一點。那你就不會害怕了，駱君麗博士(Doctor Nauckunaite)知道的是最多。

0	Names <u>any of the writers without explanation.</u>	莉莉(Julie)知道得最多。
	Gives <u>insufficient or vague</u> answer.	<p>駱君麗博士(Doctor Nauckunaite), 因為她可以教我們怎去克服演講時的困難。</p> <p>我覺得駱君麗博士(Doctor Nauckunaite)比較清楚, 因為我比較認同她的看法。</p> <p>心理學家 O.L. (Psychologist O. L.), 他很長(詳)細說明演講的事。</p> <p>偉綸(倫)(Andrew), 因為他認為公開演講是應該要去克服, 自己演講前, 應該要預備好一切。</p> <p>我覺得公開演講的能力取決於個人的個性。</p> <p>美琪(Mischa), 沒有說誰好, 只要勇敢一些, 主動一點, 多給自己信心, 一定會成功。</p>

* Please refer to the coding guide in Appendix 7 for details. Words in brackets are added by the authors to clarify names and rectify incorrect use of words by the examinees.

One finding of the PISA 2009 Field Trial is that Macao's 15-year-olds are more used to answering multiple choice items by clicking answer boxes than typing in constructed responses using Chinese input methods. They are less inclined to earn full credits for open constructed responses that are demanding reflection and evaluation of materials read.

(4) Storage, retrieval and transmission

The fourth operative component is about what procedures are required for the secure storage and efficient use of the reading assessment information. OECD requires that any computer-delivered assessment in PISA 2009 be implemented using existing school infrastructure. To meet this requirement, ERA is based on a test delivery system that runs off its own operating system so that all students will use the same Linux operating system and the same browser loaded from the bootable CD/USB. This ensures test security by preventing copying of test materials elsewhere. Additionally, the use of a bootable CD/USB avoids problems associated with internet use, such as inadequate upload/download speeds/bandwidths, computer security settings, and local/system-wide firewalls. For the ERA Field Trial, five 15-minute clusters of test units (i.e. A, B, C, D, and E) are developed. These clusters are packaged as five electronic forms (i.e. AB, BC, CD, DE and EA), so that each cluster appears first in one form and second in another form. For the ERA Main Survey, three 15-minute clusters of test units (i.e. A, B, and C) are developed. These clusters are packaged as six electronic forms (i.e. AB, BA, BC, CB, CA and AC) so as to balance the positioning effects of each of the three clusters of test units.

In this way, assessment tasks are bundled together into test forms each consists of a number of test units on a CD/USB with a Linux-based operating system, a TAO assessment platform, and a Mozilla Firefox web browser for viewing the TAO interface. Flash media player is included to enable the viewing of the stimulus material and student responses are stored in RAM during the testing session and written to a USB stick at the conclusion of the test. For security reasons, each CD/USB, not altered during the testing session, must be recovered at the conclusion of an ERA session. It can be reused in a subsequent session. In Macao, test security is guaranteed because all ERA assessments are centralized in designated test centers, and online transmission of assessment results stored in the USB sticks to Australian Council for Educational Research (ACER) are done centrally at Macau-PISA Centre.

2. An elucidation of task characteristics and item statistics of the two released ERA test units

It is not allowed to reveal the contents of the assessment tasks and discussed item statistics generated in the assessment surveys. However, it is possible to illustrate the concepts and procedures entailed using the two released ERA test units (see Appendix 1-7 for details of the seven assessment tasks comprising the two test units). In accordance with the ERA assessment framework explicated in section one, the task characteristics of the two released ERA test units are summarized in Table 5 below.

Table 5 Task characteristics of two released ERA test units

ERA Test Unit	Task	Situation	Reading Aspect	Item Format	Response Format
E010: PHISHING	E010Q02	Public	Retrieve information	Multiple	MC
	E010Q01		Retrieve information		MC
	E010Q04		Retrieve information		MC
E022: LET'S SPEAK	E022Q01	Public	Develop an interpretation	Multiple	MC
	E022Q04		Develop an interpretation		MC
	E022Q08		Develop an interpretation		MC
	E022Q09		Reflect on and evaluate the content of a text		Open_ CR

Note: MC= multiple choice; Open_CR= Open Constructed Response

For purposes of illustration using the two released ERA test units, one can see that implicit in each assessment task is an access structure envisaged by the test developer as essential to tackle/answer the task accordingly (see Appendix 1-7 for the access structures and associated coding guides). It is informative to examine the pros and cons

of the two main item format, i.e. multiple choices (e.g. E010Q02) versus open constructed response (e.g. E022Q09), so as to understand issues pertaining to assessment generation of ERA student responses for 15-year-olds in Macao.

Specifically, E010Q02 (PHISHING) requires students to retrieve information by locating an important component of an explicitly stated definition. This task is of the multiple choices format and the preferable access structure is quite straight forward, i.e. based on the information shown on the webpage *Online Phishing Resource Site* to retrieve the correct answer. There is no need to click to another webpage. Table 6 shows that E010Q02, compared with Macao’s international counterparts, has unexpectedly high item facility. Particular mention is that the ERA PISA Consortium has kindly analyzed the field trial data, and asked the Macau-PISA centre to base on the item analysis reports to explain dodgy items spotted. Detailed analyses revealed that this task is a dodgy item identified as having positive country differential item functioning. It is heartening to learn that this task has been excluded in the Main Survey 2009.

Table 6 Item statistics of ERA test unit PHISHING (3 tasks)

Task ID	Country	Language	Maximum Score	Valid N	Score=0 (%)	Score=1 (%)	Facility
E010Q02#	Macao	Chinese	1	94	19	81	81
	International*	All	1	1702	35	65	65
E010Q01	Macao	Chinese	1	93	30	70	70
	International*	All	1	1668	27	73	73
E010Q04#	Macao	Chinese	1	92	25	75	75
	International*	All	1	1655	41	59	59

* Based on test data analyzed by the ERA PISA Consortium, countries/economies included are: Austria, Belgium, Canada, Colombia, Germany, Denmark, France, Hong Kong-China, Hungary, Iceland, Ireland, Japan, Korea, Macao-China, Norway, Poland, Spain, and Sweden.

Tasks identified as having positive country differential item functioning for Macao’s 15-year-old students in PISA 2009 Field Trial.

As mentioned in earlier section, E022Q09 (LET’S SPEAK) requires students to support an opinion about the authoritativeness of a text by combining prior knowledge with information from the text. This task is of the *open constructed response* format. Examinees need to reflect on and evaluate the content of a text, and the access structure is less straight forward. There are at least three steps in order to secure an answer. First, examinees need to click on *Write a Reply* button either at the top or at the bottom of the *Education Network Forum* webpage (i.e. Step 1). Second, they need to write by typing a reply to Mischa (i.e. Step 2). Third, they should click on *Post Reply* to post the answer (i.e. Step 3), and if necessary to edit the answer by clicking *Edit Reply* button either at the top or at the bottom of the webpage (i.e. Step 4).

Table 7 shows that E022Q09, compared with Macao’s international counterparts, has unexpectedly low item facility. Again, the ERA PISA Consortium has analyzed the field trial data, and asked the Macau-PISA centre to base on the item analysis reports to explain dodgy items spotted. Acknowledgment is due to PISA Consortium for its interaction with Macau-PISA centre to refine the test units for use in the main survey. Detailed analyses revealed that this task is a dodgy item identified as having negative country differential item functioning. It is heartening to learn that this task has been

excluded in the Main Survey 2009.

Table 7 Item statistics of ERA test unit LET’S SPEAK (4 tasks)

Task ID	Country	Language	Maximum score	Valid N	Score=0 (%)	Score=1 (%)	Score=2 (%)**	Facility
E022Q01	Macao	Chinese	1	118	59	41	0	41
	International*	All	1	1874	50	50	0	50
E022Q04	Macao	Chinese	1	113	44	56	0	56
	International*	All	1	1856	53	47	0	47
E022Q08#	Macao	Chinese	1	110	86	14	0	14
	International*	All	1	1834	73	27	0	27
E022Q09#	Macao	Chinese	2	105	75	19	6	15
	International*	All	2	1803	58	25	17	30

* Based on test data analyzed by the ERA PISA Consortium, countries/economies included are: Austria, Belgium, Canada, Colombia, Germany, Denmark, France, Hong Kong-China, Hungary, Iceland, Ireland, Japan, Korea, Macao-China, Norway, Poland, Spain, and Sweden.

** After scaling the ERA student response data using item response theory, the coding guide for E022Q09 has been revised by recoding all scores coded as 2 into 1.

Tasks identified as having country differential item functioning for Macao’s 15-year-old students in Field Trial 2009.

In sum, this section illustrates what an ERA task looks like, and how item statistics generated in the field trial helps to screen items of inferior quality from entering the item pools used in the main survey.

3. Issues of reliability and validity warranted attention in ERA by Macau-PISA Centre

National centers conducting ERA assessment are urged to exercise stringent quality controls monitoring the reliability of codes assigned to constructed responses. For multiple choice responses reliability is guaranteed because they are automatically coded by the ERA Online Coding System. The following mandates are put into effect to guarantee the reliability of the codes assigned to the constructed response items at stay a high level of 85% set by the Technical Committee of the PISA Consortium.

- (1) Organizing and managing the ERA coding operations by National Project Manager strictly in accordance with PISA technical standards;
- (2) Training coders by the coding supervisor in accordance with PISA guidelines and workshop examples;
- (3) Resolving by leading coder of any coded response that has been marked for review by a coder;
- (4) Spot-checking of 2.5% of student responses to the first coding done by a coder to an item each day by the leading coder;
- (5) Blind-coding by a second coder of a minimum of 25% of student responses to an item;
- (6) Discrepancy-checking by the leading coder of different codes assigned in the first and second coding (see Table 8);

- (7) Attaining for each coder 85% coding accuracy to an item;
- (8) Consulting ERA coder query service when codes assigned by a coder do not meet the specifications of the coding procedures set by the PISA Consortium.

Table 8 Discrepancy coding to attain 85% coding accuracy to an item

First Code * (C1)	Second Code (C2)	Discrepancy (C1=C2?)	Third Code (C3)	Final Result
<i>Single coding (up to 75% of student responses to an item)</i>				
Yes	Not required	Not required	Not required	C1
<i>Double coding (at least 25% of student responses to an item) **</i>				
Yes	Yes	Yes	Not required	C1 (=C2)
Yes	Yes	No	Required	C3

* After any change due to spot-checking (minimum of 2.5% of student responses recommended)

** No more than 15% of double-coded responses to an item require third discrepancy coding. If 15% is exceeded report to PISA Consortium for remedial action.

The coding guides and the technical standards, if followed exactly, guarantee that reliability of codes assigned to constructed responses is at least at the 85% consistency level. Up to now, it is not clear how the PISA Consortium will examine the construct validity of the ERA measures constructed in the ERA main survey. The ensuing discussion pertains to Macao’s proposal regarding the construct validation of ERA constructs for informed educational policy making in Macao.

According to Messick (1995), there are six aspects of construct validity the evidences of which are to be sought: (1) content; (2) substantive; (3) structural; (4) generalizability; (5) external; and (6) consequential. Macau-PISA Centre needs to collect evidences to establish the construct validity of the ERA constructs before any cross-country comparisons can be made. To be specific, the *content* aspect asks the important question of content relevance, representativeness, and technical quality of the assessment tasks designed to span the ERA literacy constructs. The *substantive* aspect refers to that the theoretical processes underlying test responses are exemplified with the empirical evidences, e.g. the task performance model underlying different reading aspects at different proficiency levels of the ERA literacy constructs are supported by empirical data. The *structural* aspect refers to the adequacy of the coding of the open/closed constructed and multiple choices responses so as to appraise the fidelity of the scoring structure to the make-up of the ERA literacy constructs. These three kinds of evidence are important to serve as the evidential basis of test interpretation (see Step 1 in Table 10).

Because PISA 2009 is a study of international assessment of reading literacy, it is important to examine the generalizability aspect of the ERA literacy constructs. Otherwise, trends cannot be gauged, nor literacy performance between groups can be compared with confidence and validity. This kind of evidence is important to serve as the evidential basis of test use (see Step 2 in Table 9). The external aspect includes converging and discriminating evidences between the print and ERA literacy constructs, as well as evidences of criterion relevance and applied utility (e.g. the use of the distribution of the ERA literacy constructs in the design of remedial educational programs for 15-year-olds in Macao). This kind of evidence is important to serve as the evidential basis for both test interpretation and test use (see Step 1 and 2 in Table 9).

Last, but not least, the consequential aspect appraises the value implications of score interpretation as a basis for action, as well as the actual and potential consequences of test use. For example, Macao educational practitioners should conceptualize low-performing schools in new lights—i.e. those schools with large proportion of 15-year-olds classified as “below level 1” should not be envisaged by the public as poor schools, but as schools with admirable missions catering for the special educational needs of the students they served. In these schools, low-performing students should be remedied with equitable policies so as to help students meet the challenges of life-long learning in the digital age (see Step 3 and 4 in Table 9).

Table 9 Adaptation of Messick’s (1989) validation procedures for informed educational policy making in Macao

	Test Interpretation	Test Use
Evidential Basis	<u>Step 1</u> : Construct Validity (CV)	<u>Step 2</u> : CV + (Relevance/Utility: R/U)
	<i>Scaling of both print and ERA coded responses using item response theory (IRT) models into proficiency levels that are explainable in terms of progressions of reading literacy.</i>	<i>Distributional statistics of the print and ERA literacy measures, as well as classification of 15-year-olds into graded proficiency levels allow Macao to compare with other countries/economies participating in PISA 2009.</i>
Consequential Basis	<u>Step 3</u> : CV + Value Implications (VI)	<u>Step 4</u> : CV + R/U + VI + Social Consequence (SC)
	<i>Students classified as “below level 1” in print and ERA literacy should be remedied with equitable policies by Macao schools so as to help students meet the challenges of life-long learning in the digital age.</i>	<i>Conceptualizing low-performing schools in new lights—i.e. those schools with large proportion of 15-year-olds classified as “below level 1” should not be envisaged by the public as poor schools, but with admirable missions catering for the special educational needs of the students they served.</i>

4. Conclusion

For purposes of summarizing computerized assessment and test administration issues examined in this paper, it is illuminating at this juncture to introduce to readers the eight guidelines for the development and use of CBA (Green, et al., 1995). This is because relevant to each of these eight guidelines there is a corresponding pertinent issue negligent treatment of which may threaten to a great extent the validity of the ERA constructs developed. To recapitulate, the eight guidelines needed attention and action for the development and use of CBA are: (1) The computerized testing system should be

appropriate for the purpose of test; (2) The psychometric model underlying the computerized test should be appropriate for the constructs being measured; (3) The item pool should be of appropriate quality to support the test purposes and to measure the examinee population; (4) Examinees should be well trained in the use of the computer testing system; (5) Software and hardware requirements of the computerized test delivery system should allow for the test to be administered in a fair and professional manner; (6) The item-selection algorithm and stopping rule should be appropriate for the purpose of the test; (7) The security should protect the integrity of the computerized test and the examinees' score records; and (8) The technical quality of test results should be sufficient for the purposes of score interpretation.

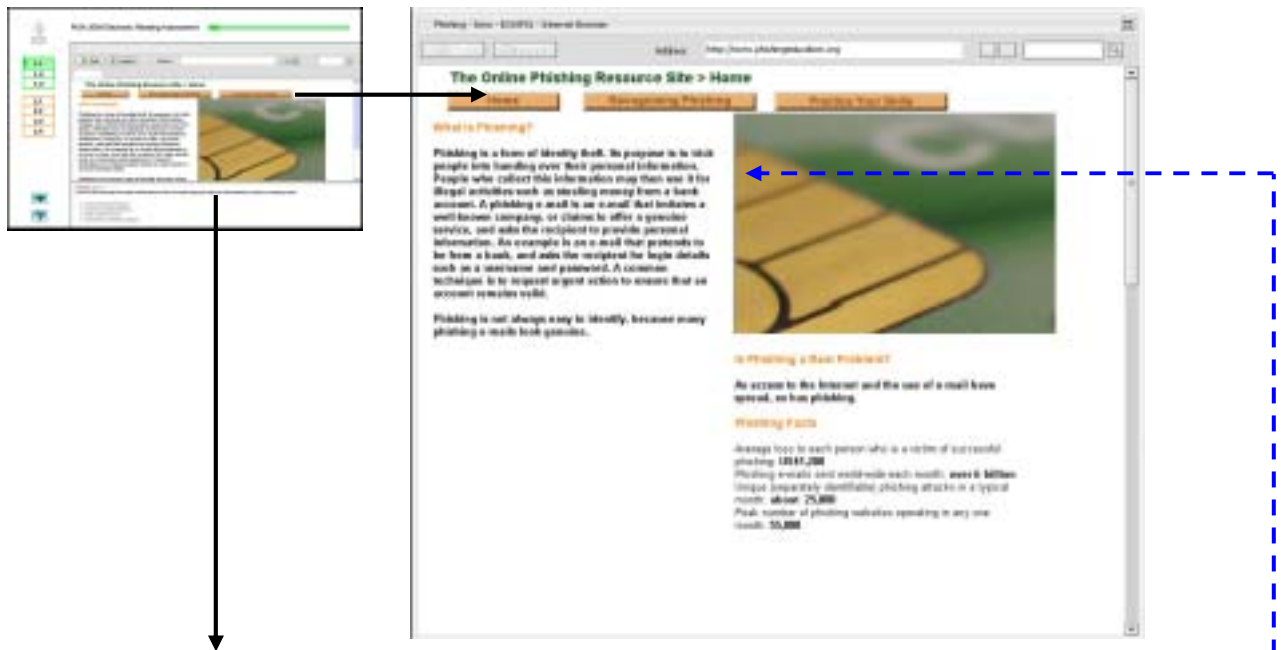
Corresponding ERA issues handled by the ERA PISA Consortium and examined in this paper are: (1) Elucidating target examinee population and definition of reading literacy; (2) Explicating ERA framework and the associated reading literacy constructs; (3) Designing item pools with pre-designed characteristics (e.g. situation, item format, reading aspect, text type) and item qualities (e.g. items of appropriate facility spanning the various levels of the reading literacy scale); (4) Familiarizing students with sample released items, as well as the operating environment of the computer-based assessment system; (5) Conducting testing in designated test centers following agreed test administration procedures; (6) Distributing randomly to students different test forms with a timed "lock-step" item design; (7) Facilitating coding of multiple choices and open constructed response data, alongside security of online transmission of assessment data to PISA Consortium; and (8) Scaling student response data using item response theory (IRT), and analyses of items having country differential item functioning.

To conclude: based on the Professional Practice Board Steering Committee on Test Standards, this paper introduced the four operative components of ERA test delivery system. After the field trial in 2008, two ERA test units were released by the PISA Consortium. This allowed the present authors to make use of them to elucidate the PISA approach of simulated online reading assessment. Based on Messick's (1989) unified construct validation theory and the ERA international coding procedures, this paper discusses issues of reliability and validity warranted attention by Macau-PISA Centre for cross-country comparison and informed policy making. This paper ends with a summary of eight pertinent issues, the corresponding guidelines of which can be concluded as having been fulfilled satisfactorily by the ERA PISA Consortium in its research and development of a CBA. More evidences are needed to be sought in due course in order to validate the ERA constructs against the objectives of the PISA 2009 Reading Literacy Study.

References

- Cheung, K. C., & Sit, P. S. (2008a). *Preparing for Macao PISA 2009: Reading assessment framework, released items and coding guides*. Macao: Educational Testing and Assessment Research Center, University of Macau.
- Cheung, K.C., & Sit, P.S. (2008b). Electronic reading assessment: The PISA approach for the international comparison of reading comprehension. *Journal of Educational Research and Development*, 4(4), 19-40.
- Green, B., Kingsbury, G., Loyd, B., Mills, C., Plake, B., Skaggs, G., Stevenson, J. & Zara, T. (1995). *Guidelines for computerized-adaptive test development and use in education*. Washington, D.C.: American Council on Education.
- Goldhammer, F., Martens, T., Naumann, J., Rolke, H., & Scharaf, A. (2008). *Developing items for electronic reading assessment: The hypertext builder*. Paper presented at the XXIX International Congress of Psychology, July 20-25, 2008, Berlin, Germany.
- Latour, T., Martin, R., Plichart, P., Jadoul, R., Busana, G., & Swietlik-Simon, J. (2008). *TAO: Paving the way to new assessment instruments using an open and versatile computer-based platform*. Paper presented at the XXIX International Congress of Psychology, July 20-25, 2008, Berlin, Germany.
- Messick, S. (1989). Validity. In: R.L. Lin, *Educational Measurement (3rd Edition)*, New York: Macmillan, 13-103.
- Messick, S. (1995). Validity of psychological measurement: Validation of inferences from person's responses and performances as scientific enquiry into score meaning. *American Psychologist*, 50(9), 741-749.
- OECD (2007a). *PISA 2006 science competencies for tomorrow's world. Volume 1: Analysis*. Paris: Author.
- OECD (2007b). *PISA 2009 Reading Framework – 23rd Meeting of the PISA Governing Board, Oslo, 12-14 March, 2007*. Paris: Author.
- OECD (2007c). *Reading literacy: A framework for PISA 2009 (20 July, 2007 version) – Meeting of the PISA National Project Manager, Dubronik, 2007*. Paris: Author.
- Professional Practice Board Steering Committee on Test Standards. (2002). *Guidelines for the development and use of computer-based assessments*. Leicester: The British Psychological Society.
- Sit, P.S., & Cheung, K.C. (2009, April). *International comparison of electronic reading literacy: Uncovering Macao's 15-year-olds' unexpected strengths and weaknesses using country differential item functioning*. Paper presented at 2009 Chinese American Educational Research and Development Association International Conference on "Educational Technology: Enhancing Teaching and Learning with Technology in the 21st Century". San Diego.

Appendix 1: Access structure pertinent to answering Task E010Q02 (PHISHING)



PHISHING: Task 1

E010Q02

You are at the home page of the Online Phishing Resource Site. According to the information on this page, which one of the following is a feature of a phishing e-mail?

- A It asks for personal information.
- B It contains unwanted advertising.
- C It offers a genuine service.
- D It comes from a well-known company.

Question intent:

Access and retrieve: *Retrieve information*

Locate an important component of an explicitly stated definition

Access paths:

Based on the information shown on *Online Phishing Resource Site* to retrieve the correct answer to the multiple choice question. (n.b. Blue dotted arrow shows where the answer is located; there is no need to click to another webpage)

Coding guide:

Full Credit

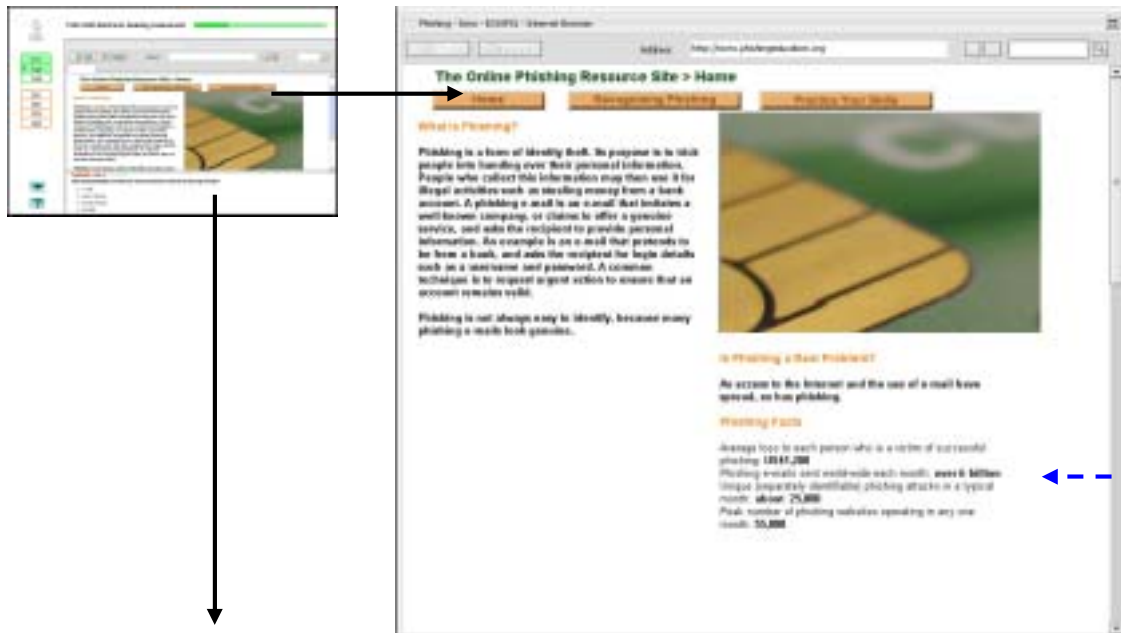
Code 1: A. It asks for personal information.

No Credit

Code 0: Other responses.

Code 9: Missing.

Appendix 2: Access structure pertinent to answering Task E010Q01 (PHISHING)



PHISHING: Task 2

E010Q01

How many phishing e-mails are sent around the world in an average month?

- A 1,200.
- B over 6 billion.
- C about 25,000.
- D 55,000.

Question intent:

Access and retrieve: *Retrieve information*

Identify the reference of a number in a list

Access paths:

Based on the information shown on *Online Phishing Resource Site* to retrieve the correct answer to the multiple choice question. (Blue dotted arrow shows where the answer is located; no need to click to another webpage)

Coding guide:

Full Credit

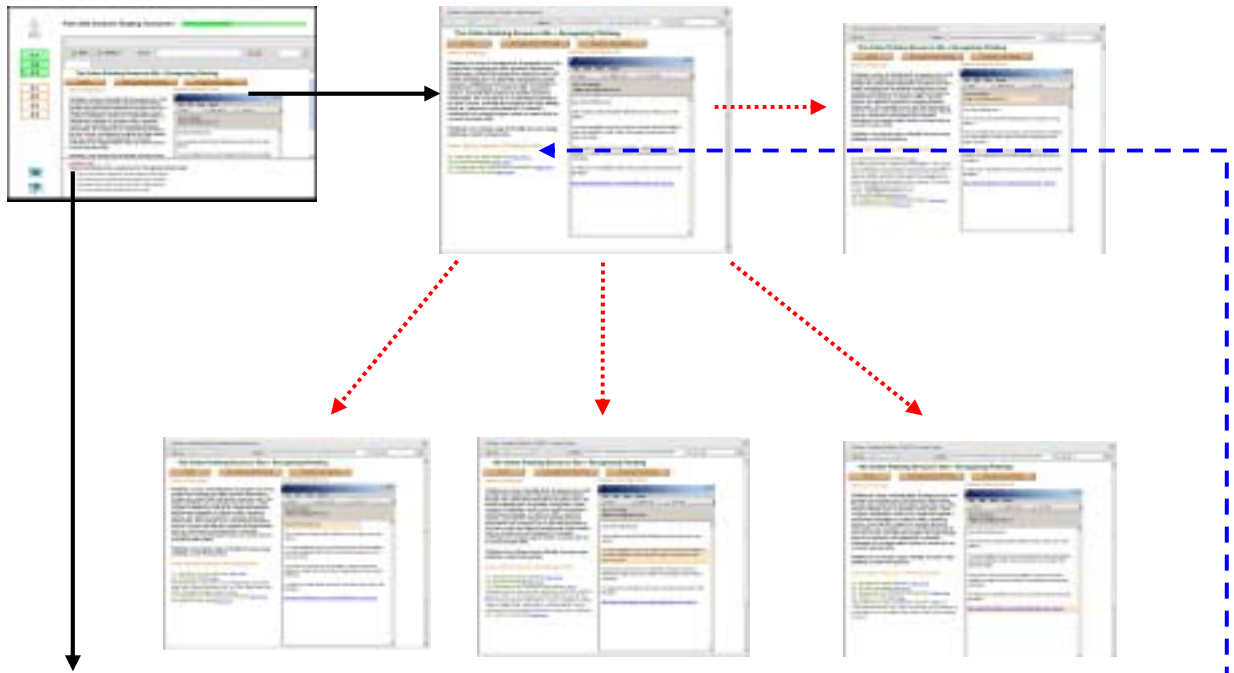
Code 1: B. over 6 billion.

No Credit

Code 0: Other responses.

Code 9: Missing.

Appendix 3: Access structure pertinent to answering Task E010Q04 (PHISHING)



PHISHING: Task 3

E010Q04

Which of the following tricks is explained on the “Recognising Phishing” page?

- A The e-mail asks the recipient to donate money to a fake charity.
- B The phishing e-mail installs spyware on the user’s computer.
- C The author of the e-mail inserts a fake link to a fake website.
- D The e-mail pretends the recipient has won a prize.

Question intent:

Access and retrieve: *Retrieve information*

Locate explicitly stated information

Access paths:

Click a series of hyperlinks in *Recognizing Phishing* to go to other webpage so as to retrieve the correct answer to the multiple choice question. (n.b. The four hyperlinks are located at where the blue dotted line points to, and to where they are linked are shown in red small dotted lines. Students may choose to click these four hyperlinks in any order they like to retrieve information.)

Coding guide:

Full Credit

Code 1: C. The author of the e-mail inserts a fake link to a fake website.

No Credit

Code 0: Other responses.

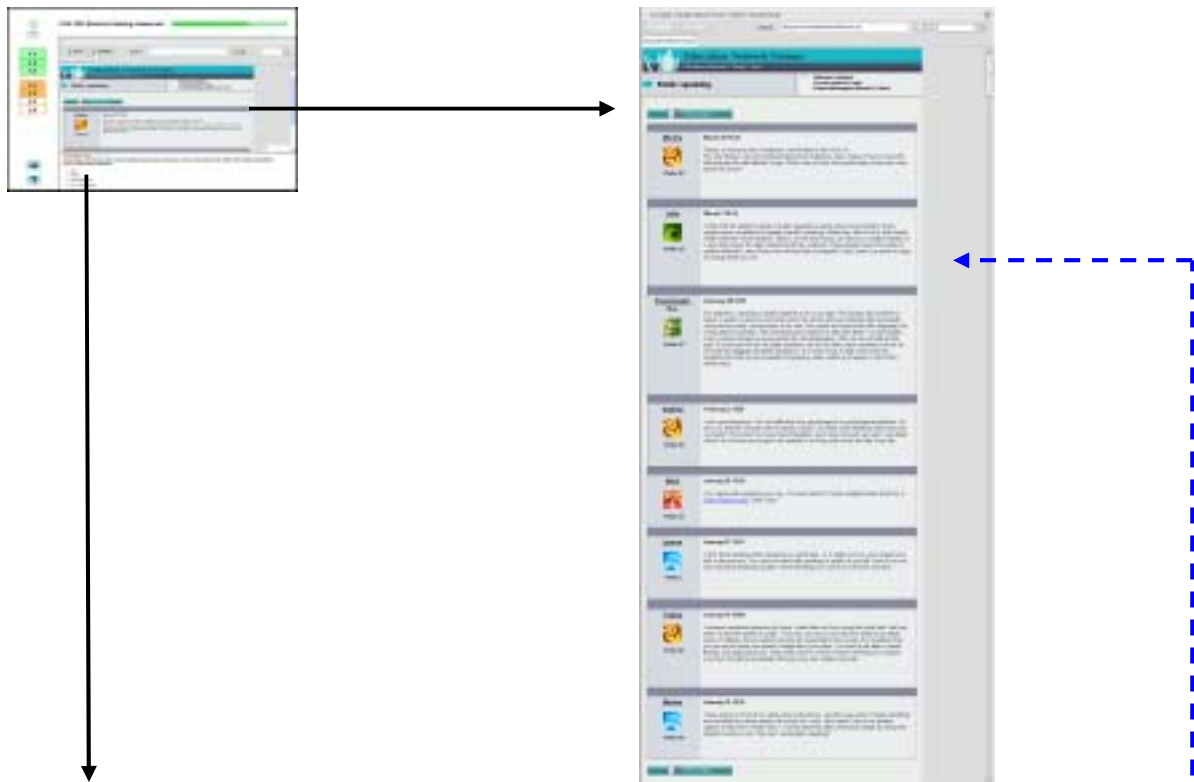
Code 9: Missing.

Appendix 4: Access structure pertinent to answering Task E022Q01 (LET’S SPEAK)



<p>LET’S SPEAK: Task 1 E022Q01</p> <p>Who wrote the first reply to Mischa in this Internet forum discussion?</p> <p>A Julie. B Mark. C Dr. Nauckunaite. D Tobias.</p>
<p>Question intent: Integrate and interpret: <i>Develop an interpretation</i> <i>Recognise the sequence of posts in a forum discussion</i></p> <p>Access paths: Based on the information shown on <i>Education Network Forums</i> to develop an interpretation of its layout format so as to answer the multiple choice question. (Blue dotted arrow shows where the answer is located; no need to click to another webpage)</p>
<p>Coding guide:</p> <p>Full Credit D. Tobias.</p> <p>No Credit Code 0: Other responses. Code 9: Missing.</p>

Appendix 5: Access structure pertinent to answering Task E022Q04 (LET’S SPEAK)



LET’S SPEAK: Task 2 E022Q04

Lauren writes, “Even if you are very scared of speaking in public, there are things you can do to overcome your fear.” Which writer would be most likely to **disagree** with Lauren’s statement?

- E Julie.
- F Tobias.
- G Psychologist O.L.
- H Dr. Z. Nauckunaite.

Question intent:
Integrate and interpret: *Develop an interpretation*
Compare two arguments to recognise a contrast

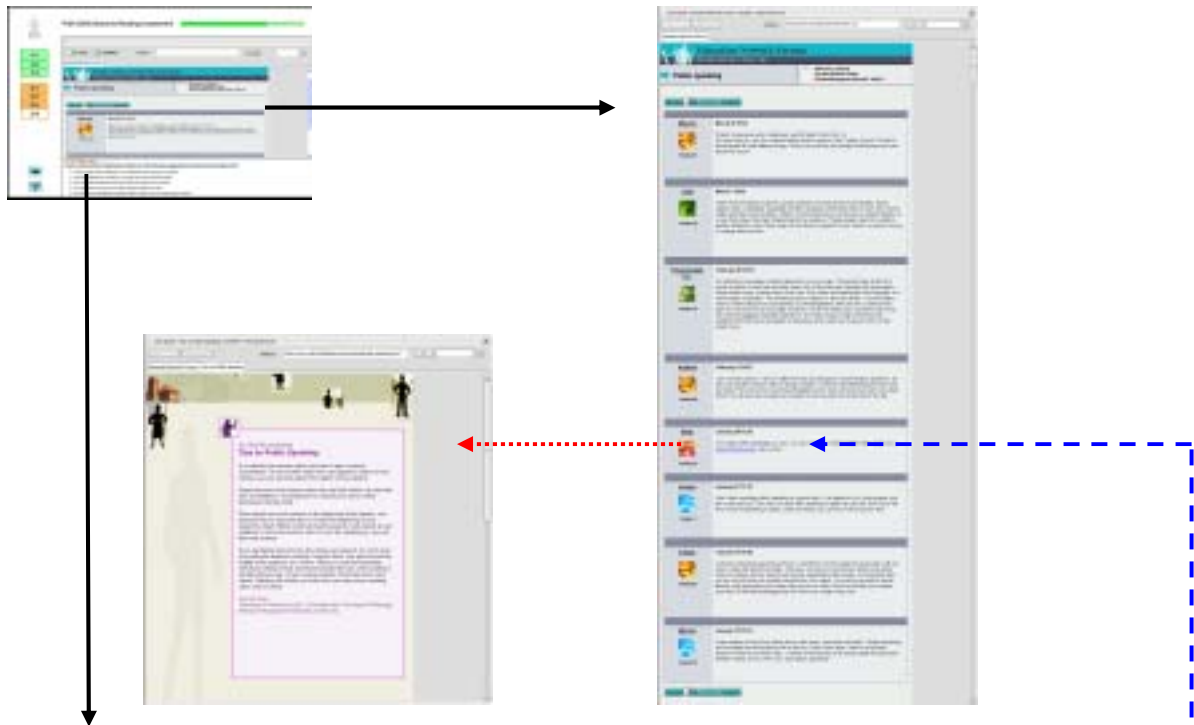
Access paths:
Based on the information shown on the *Education Network Forums* to develop an interpretation to answer the multiple choice question. (Blue dotted arrow shows where the answer is located; no need to click to another webpage)

Coding guide:

Full Credit
A. Julie.

No Credit
Code 0: Other responses.
Code 9: Missing.

Appendix 6: Access structure pertinent to answering Task E022Q08 (LET'S SPEAK)



LET'S SPEAK: Task 3 E022Q08

Find the article by Doctor Nauckunaite. Which one of the following suggestions does Doctor Nauckunaite make?

- A A casual and relaxed attitude is most effective when you give a speech.
- B If you think about your audience, you will worry less about yourself.
- C If you can hide the fact that you are afraid, you will feel less afraid.
- D It is safest to memorise your whole speech before you start.
- E It is best to look at different sections of the audience in turn during your speech.

Question intent:
Integrate and interpret: *Develop an interpretation*
Recognize which one of a set of suggestions is made in a text

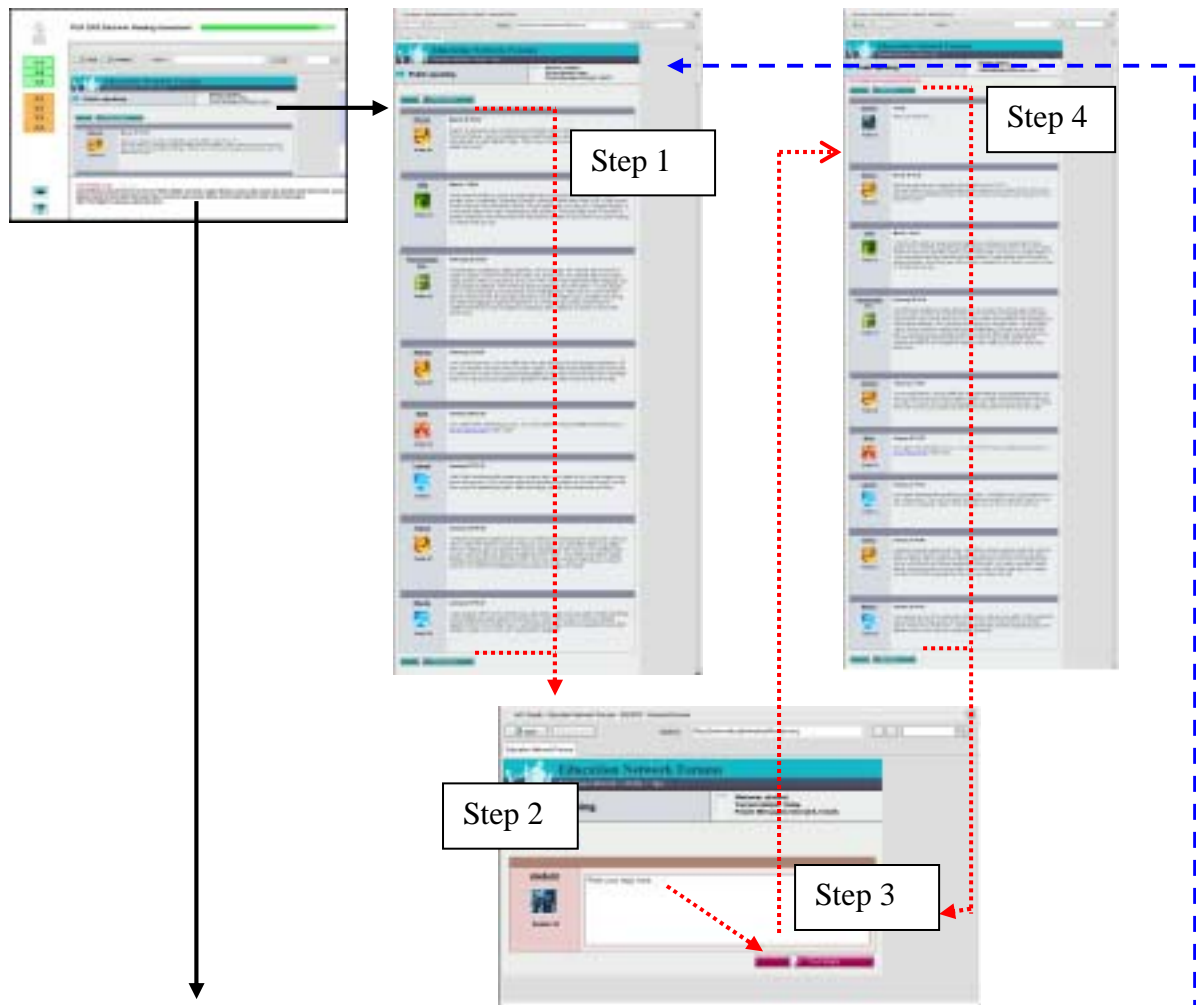
Access paths:
Click the hyperlink in *Education Network Forums* (blue dotted arrow shows where the hyperlink is located) to go to required webpage (shown by the red small dotted arrow). Based on the information displayed to integrate a set of suggestions made by Doctor Nauckunaite so as to answer the multiple choice question.

Coding guide:

Full Credit
C. If you can hide the fact that you are afraid, you will feel less afraid.

No Credit
Code 0: Other responses.
Code 9: Missing.

Appendix 7: Access structure pertinent to answering Task E022Q09 (LET’S SPEAK)



LET’S SPEAK: Task 4 E022Q09

Look at Mischa's post for March 10. Click on “Write a Reply” and write a reply to Mischa. In your reply, answer her question about which writer, in your opinion, knows the most about this issue. Give a reason for your answer. [Note: use the Back button to refer to the Forum page.]

Click “Post Reply” to add your reply to the forum.

Question intent:

Reflect and evaluate: *Reflect on and evaluate the content of a text*
 Support an opinion about the authoritativeness of a text by combining prior knowledge with information from the text

Access paths:

Click on “Write a Reply” button either at the top or at the bottom on the *Educaton Network Forum* webpage (i.e. step 1) and write by typing a reply to Mischa (i.e. step 2). Then click on “Post Reply” to post the answer (i.e. step 3). The answer can be edited by clicking “Edit Reply” button either at the top or at the bottom on the webpage (i.e step 4). (n.b. Blue dotted arrow shows where to start doing the tasks, and red dotted arrows in sequence indicate the planned access structure of the linked webpage)

Coding guide:

Full Credit

- Code 1: Identifies Doctor Nauckunaite and/or Psychologist O.L. (explicitly or implicitly) AND refers to their professional status. May express scepticism about their professional status.
- The two professionals are the ones knowing the most, but only Dr. N gives advice on how to work with the problem.
 - Psychologist O.L. or Dr.Nauckunaite because they are both trained in the area.
 - Doctor Nauckunaite. This is the only one that has the support of a university behind it.
 - Dr Nauckunaite, because she's from a university.
 - A university professor has the most practical experience in talking in public.
 - Mark has looked into it and found an article written by a person who knows what to do. This man is obviously a professional on the matter, so I think you should follow his advice.
 - I'd take most notice of the one who wrote the book, because she has published a book on this subject.
 - Psychologist O.L. sounds authoritative, but of course you can't really know that she is a psychologist.
 - You should follow Psychologist O.L's advice, as not only he is an experienced psychologist but answers all the questions concerning public speaking accurately, and is very believable.
 - Psychologist O.L. because he's a trained psychologist.
 - The person that probably knows most about this is the Doctor. He has had most experience (or at least more than Julie or Tobias) and I think he is therefore more trustworthy.

Identifies any of the four writers named by Mischa (Julie, Tobias, Psych OL or Dr. Nauckunaite) AND gives a reason that is consistent with the text, related to the cogency, practicality or logic of the text.

- Psychologist O.L. because what he says makes sense in terms of the way you see small children and teenagers behave.
- Tobias because he's actually done it.
- Doctor Nauckunaite because she has set out her ideas in a practical way.
- I think Tobias has the greatest idea of what he's talking about. He gives you concrete ways in which to improve your public speaking, and if you follow what he says I'm sure you'll do fine! :)
- Probably go with Mark's link, it has the most useful hint about how to overcome fear of public speaking. [*"Mark's link" implies Dr N.*]
- I think that Tobias is right. It does help to rehearse and know your topic well. I also agree with Julie to an extent, because some people are more outgoing than others. But with preparation and a good attitude you can make a good speech. Avoiding it altogether is not a solution!!
- Julie's ideas describe the way people differ, so she is the one I'd believe.

No Credit

- Code 0: Names any of the writers without explanation.
- Psychologist O.L.
 - Doctor Nauckunaite.

Gives insufficient or vague answer.

- Tobias because I agree with him.
- Doctor Nauckunaite because she is the best.
- Tobias because his ideas make sense.

Shows inaccurate comprehension of the material or gives an implausible or irrelevant answer.

- Psychologist O.L. because he's my favourite.
- Tobias because he tells you how to avoid public speaking. [*inaccurate comprehension*]
- I'd go for Mark. [*not one of the four writers named by Mischa*]

Code 9: Missing.