

# A Read to Succeed Project Primer

## Test purpose

### What is the purpose of the assessment suite and instructional system?

The **Prime** assessment suite and accompanying instructional system is being developed to provide a broad range of users with access to a flexible, multi-purpose, efficient and affordable toolkit that will serve all of their assessment needs and many of their instructional requirements. All the tools will be web-based and adaptive, so will rapidly give reliable and immediate results at low cost.

The assessment products will provide valid, reliable, comparable and interpretable estimates of skill in the following domains:

- Prose literacy
- Document literacy
- Numeracy
- Oral fluency

The instructional system will provide instructors with tools designed specifically to address the gaps identified by the assessments. The initial offering will be focussed on improving skills to move learners from Level 2 to Level 3. Future offerings will expand the scope to improve skills of learners at Level 1 and will also address skill improvement from Level 3 to Levels 4 and 5.

The **Prime** suite of products will include:

**TOWES Foundation (BASES):** a diagnostic assessment for IALSS prose literacy Level 1 and 2 readers. This tool classifies individuals into homogeneous groups of learners based upon patterns of strength and weakness in their reading decoding and comprehension skills, using the methods applied in the International Survey of Reading Skills (ISRS). These groups mirror those used in market segmentation analyses conducted by DataAngel Policy Research Inc.

**TOWES Focus (Focus):** a scalable measurement tool. Uses can range from providing instructors with a general understanding of a test taker's literacy skill level, to enable efficient instructional groupings, (with a score reported as a Level only) to a tool well suited to organizational or system accountability reporting.

Scores for system accountability will be reported on the IALSS 500 point scale and as a Level. As will be the case for all assessments in the suite, reporting scales and proficiency levels for **Focus** will be linked to the International Adult Literacy and Skills Survey (IALSS) and, by extension, to the Essential Skills framework and profiles. **Focus** will provide summative assessment by giving IALSS proficiency scores at program exit and learning gain information by giving IALSS scores from testing at program entry and exit.

**TOWES Scaffold (Construction):** is the online instructional system that will be developed using the same technology and linked to the same results parameters as the assessments. **Scaffold** can be used with any of the TOWES products in the suite.

**TOWES Sharp (PRÉCISION):** a measurement tool designed for use in high stakes situations such as job or training selection or occupational development. **Sharp** will be particularly useful for workplace and workforce development; scores will be reported on the 500 point scale and as a Level.

## A Read to Succeed Project Primer

|                                   |   |
|-----------------------------------|---|
|                                   | <p><b>What questions will it answer?</b></p> <p>The assessment system will allow users to select any combination of skill domains and any combination of uses. The system will adjust the number of items to yield the level of precision needed for a particular use. The instructional system will provide instructors and learners with material specifically aimed at addressing the gaps identified in the assessment results.</p>   |
| <p><b>Target Population</b></p>   | <p><b>For whom is the test intended?</b></p> <p>The assessments are intended for adults and youth aged 15 and over. The reading components assessment will work down to age 10 or younger. In the case of <i>Focus</i>, where it is used for system accountability measures, it will most commonly be used for adults participating in recognized literacy programming.</p> <p><b>Who will be eligible to take it?</b></p> <p>The tool may be used for all adults except the blind.</p>   |
| <p><b>Target Construct(s)</b></p> | <p><b>What traits or abilities will be measured?</b></p> <p>The tools assess proficiency as defined by the 2003 IALSS and 2005 ISRS assessment i.e. prose literacy, document literacy, numeracy and oral fluency.</p> <p><b>How are these abilities defined for testing purposes?</b></p> <p>The IALSS constructs are defined by the cognitive requirements associated with specific tasks. The cognitive characteristics of each task are the basis of a classification system that groups items according to their probability of being answered correctly by individuals with different levels of proficiency in the constructs. Each task, while embedded in context, represents a generalizable application of a single salient construct. Where accuracy permits, results may be reported for context-specific sub-domains or specific cognitive processes within the dominant constructs.</p>  |
| <p><b>Underlying Scale</b></p>    | <p><b>Along what scale is the test aligned?</b></p> <p>Results will be linked to the IALSS scale through the use of common items carried from IALSS. This scale has an international mean in the adult population aged 16-65 of approximately 250, and the range of scale scores is approximately 0 to 500. Proficiency will also be reported on the IALSS levels using the RP80 mastery standard; for each construct, each individual will be classified into a performance level that is defined by tasks he or she has at least an 80% chance of performing successfully. Reading components data gathered on tests takers who complete <i>TOWES Foundations</i> will be used to classify individuals into one of 6 learner groups.</p> <p><b>How is the test scored?</b></p> <p>The test will be scored automatically and instantly by the online system. All item responses are scored algorithmically using a combination of rule-based scoring and propensity matching. Item scores are then matched to item response theory parameters that have been linked to the IALSS scale. These parameters are used to estimate the final scale scores and standard errors by optimizing score likelihood functions.</p> |

## A Read to Succeed Project Primer

|                                   |  |
|-----------------------------------|--|
|                                   | <p><b>What levels or increments are of greatest interest?</b></p> <p>The IALSS levels are theoretically predicted and empirically confirmed with observational data from response statistics as well as neuroscience. The most significant distinction is between level 2 and level 3 on the IALLS scale. The numeric width of each level is 75 scale points. Based on previous secondary research, twenty-five points on the scale represents the average skill gain associated with an additional year of formal education.</p> <p><b>The IALS literature indicates that the lowest scale score (ranking) that can be assigned to a test item is 188. How does this ranking impact the test scores? Does this mean that the lowest score a test taker can achieve is also 188? If a test taker can score lower than 188 on IALS (using a TOWES result), how is their score determined if there are not items ranking below 188?</b></p> <p>The scale used to report the IALS/IALSS results is meant to describe how individuals perform relative to different types of tasks, rather than relative to an absolute scale. The value of 188 describing an item does not represent the lowest or highest possible score one can achieve, even with that single item. If a person has a very low chance of answering the item correctly, his or her estimated score may be much lower. Conversely, if the chance of answering the item is very high, the score may be much higher. The characterization of an item uses the score associated with a convenient or meaningful chance of answering correctly -- in the case of IALS, an 80% chance. The more an individual's chance of answering varies from the value used to characterize an item, the less useful the item is for estimating that person's score. In theory, there is no lower limit to the scores that may be estimated, if there are a sufficient number of low level items. Scores below 188 may be estimated accurately by giving large numbers of low level items; in IALS, the number of low level items was constrained because the test was designed for the general population. The ISRS includes a large battery of very low-level cognitive tasks that are essential for literate behaviour. However, the chance of having a score lower than zero are practically nothing, because very low level tasks become characterized by basic cognitive functions, such as pattern recognition and memory capacity. The ISRS improves on the cognitive interpretation of low literacy scores by taking a more diagnostic approach to the factors limiting a person's literacy; for example, a score of zero may be assigned if a person does not recognize a single letter of the alphabet.</p> |
| <p><b>Design and Format</b></p>   | <p><b>How is the test designed and formatted? Is it task-based, and what are the item types?</b></p> <p>All tasks mimic authentic individual interactions with stimuli encountered in non-testing settings. Stimuli are selected from a variety of contexts. In the current release version, all stimuli are static two-dimensional text forms or objects. Items are selected to provide full coverage of skill domain, range of difficulty, the underlying determinants of difficulty and the full range of contents and contexts defined in the IALSS assessment. See The Adult Literacy and Life Skills Survey: New Frameworks for Assessment <a href="http://www.statcan.ca">www.statcan.ca</a></p>  |
| <p><b>Development Process</b></p> | <p><b>What theoretical model underpins the test?</b></p> <p>The tests use the theory that was developed and validated for the IALSS and ISRS assessments. In general the theory specifies a set of variables that underlie or determine task difficulty. The test is designed to provide a representative sample of these underlying determinants. Test data is then summarized using advanced statistical techniques to provide empirical estimates of both task difficulty and individual proficiency. Provided these predicted and observed task</p>  |

# A Read to Succeed Project Primer

difficulties are in close agreement the test results can be safely generalized. The same methods provide an empirical method of detecting bias in item performance. The assessment is described as one that enables described proficiency.

## **What are the test development procedures?**

The test will use three large item pools – a selection of IALSS items with known psychometric properties, the TOWES item pool with known psychometric properties and a pool of newly developed items. Smaller numbers of proven test items will be added where possible.

## **Pilot process and sample:**

Test items will be administered to a minimum of 800 adults of varying skill levels in each of Canada’s official languages in order to confirm their psychometric properties. Once validated, the items will also be administered to an additional 400 individuals in each of Canada’s official languages as part of an intervention study. Both samples will contain sufficient Aboriginal adults to independently validate the tools.

## **Administration**

### **How will the test be administered?**

The test will be delivered over the web. The assessment will be administered by a certified test administrator. Results will be provided in real time.

### **Who will invigilate the test? Will there be trained experts at specified sites? Or will this be something that an instructor would learn to administer to her own students?**

There will be authorized testing centres that can be accessed when that makes sense for a group or individual, but an instructor will also be able to become certified to administer the test provided the location meets the minimum required standards for testing. These standards will be: access to the internet, computers in a reasonably ergonomic, secure environment with minimum screen size of 15 inches, screen resolution of 1024x768 and standard data entry devices, such as a mouse and keyboard (or equivalents).

### **What is the process of invigilator certification?**

Invigilator certification will be similar to the current online test administrator certification TOWES now uses which is a one hour online course followed by a brief exam.

### **How long will the test administration take?**

Test duration will depend upon the purpose selected and the consistency of the individual. Program triage will take the least time, certification the most. Adaptive algorithms will allow accurate scores to be generated using a much smaller number of items than the current paper-based TOWES assessment because the test taker’s responses to items are instantly recognized. This instant recognition enables the system to provide a follow up questions appropriate to the current estimate of a test-taker’s proficiency, depending on the pattern of responses demonstrated by the test-taker to that point.

The test will take 30-50 minutes to complete. Like TOWES, we would expect the test to be completed in one sitting; although, the system will be capable of storing a test taker’s progress and in the event of an interruption, returning to that point. Interruptions may include power or

# A Read to Succeed Project Primer

connectivity failures or voluntary breaks by the test takers, which may be allowed at different points in the tests. However, test takers need to be aware that, if their results are statistically anomalous and their session has an interruption, their results will be invalidated, which may require them to take a different version of the test. The maximum amount of time a test taker would be allowed to complete a test, once started, is 48 hours.

## **Will the test be fee-for-service, and if so, how much will it cost?**

The cost will depend on which tool is being used. A price list will be provided during or following the pilot phase. During the pilots, the Focus tool will be provided free. The only exception to the free tool is the Versant™ oral fluency test which is licensed from Pearson Technologies. For the duration of the pilot, we will have access to special pricing of \$8.00 per test.

## **What are the intervals for re-taking the test?**

The intervals for re-taking the test are determined by rules for item exposure. The adaptive test structure creates individualized testing experiences by varying content according to proficiency and randomly choosing between parallel testlets, which minimizes individual item exposure. However, the general principle guiding the administration of the tests is that no individual test-taker should be administered the same item within 30 days.

## **How will the test be operationalized?**

The tests will be sold by licensed distributors and administered by certified test administrators using a secure web application. Response data are stored temporarily on local clients and permanently on the web application servers. Test-taker interactions with the test items will be saved every 10 seconds.

## **When will it be available?**

**Focus** will be ready for use for system accountability by fall of 2011 in both English and French. Validation of the French versions will include Francophones from as many locations as possible including Ontario and New Brunswick as well as smaller populations across the country. College Lionel Groulx, our Francophone partner, will be conducting consultations with Francophones from a number of settings to gather information on their needs and to provide information on the project.

The tests will be available 23 hours of each day, with a maximum concurrent downtime of one 4-hour period per month for regular maintenance. When tests are offline, the web application will provide free resources related to the tests and learning.

## **Test results**

### **How will results be delivered to test takers?**

Test takers will receive a non-technical score report that summarizes their proficiency, identifies strengths and weaknesses and makes remedial recommendations. Test administrators will have access to a more technical score report. Immediately after completing the test, the score report is available for viewing onscreen in printable form and may be emailed to the test-taker or administrator email accounts, depending on the administration preferences. Both types of score reports will be available in archive form for print or download from the individual test-taker's account page or the account administration page, respectively.

# A Read to Succeed Project Primer

## **Will there be numerical scores and qualitative descriptors?**

Scores will be reported on the IALSS scales as scores on the 500-point scales and in the IALSS proficiency levels. Ultimately, the quantitative results will be accompanied by a qualitative description of strengths and weaknesses, a recommended remedial strategy, associated costs and expected benefits based upon the IALSS database.

## **How will results be interpreted?**

There are no pass-fail decisions based on the results of the test. The purpose of the tests is to locate the test-taker on a continuum of skills so that optimal decisions can be made about immediate learning goals and interventions. The results will include the probability that each test taker may be classified into each IALSS proficiency level. The results will be interpreted to the client by a certified test administrator who will have completed training in debriefing clients.

## **What will a result mean?**

Results may mean several things depending on the assessment used and the reason for which it was used. As a start, results from a test for a group or groups of clients will provide reliable information regarding the efficacy of the training that was delivered. Individual results will provide reliable information of a client's current skill level, so pre testing will describe the types of interventions that would be most useful to improve a client's skills. Pre and post-intervention test results will describe skill level changes and indicate where other interventions might be required.

## **What decisions can be reasonably informed by the test results?**

The tests are designed to support each of the uses. In a statistical sense, this means reducing the probability of Type I and Type II classification errors to a negligible level, depending on the relative importance of each error type for the specific test use. For example, false positives are less tolerable in learning-oriented uses, while false negatives are less tolerable in certification uses.

## **Technical Features**

### **What information is available on reliability?**

The system will generate complete statistical information automatically. Omnibus test statistics will include overall reliability for the complete sample of test-takers as well as for each IALSS proficiency level. In addition, expected conditional standard errors will be provided for each proficiency score for each construct measured by each test. Internal to the system, accuracy of each item will be monitored to identify and remove items that do not provide accurate information.

### **What validation research has been carried out?**

The IALSS framework and methods have been validated in a broad range of languages and cultures over the past 25 years through the conduct of IALS and IALSS/ALL. The frameworks and approach to measurement have also been applied in the OECD's PISA assessment of 15 year olds' reading, mathematics and science proficiency and for the design of the forthcoming PIAAC project.

# A Read to Succeed Project Primer

|                        |   |
|------------------------|---|
|                        | <p><b>What statistical methods were used to analyze the test quality?</b></p> <p>Item calibration and scoring makes use of conditional response probability models from item response theory (IRT). At an item level, inferences about item quality are based on the variability of response probability across different proficiency scores. At a test level inferences about test quality are made based on the variability of test scores across different proficiency scores (Fisher’s information) and the relative efficiency of the test in terms of the ratio of expected administration time to the test information. Full information factor analysis is used to verify the dimensional structure and appropriateness of IRT statistical methods.</p> <p><b>What further studies are planned?</b></p> <p>The project team is developing a web-based instructional system that is designed to offer custom remedial responses to the diagnostic profile generated by the assessment system. This system is the first in the world to embody the same theoretical framework that is used to drive assessment. Field testing of the instructional system will be undertaken in combination with pre and post assessment to accurately measure skill gains of learners.</p> <p>A regular program of data analysis will be undertaken using the proficiency data and background characteristics available on the database. All statistics and parameters will be updated on a quarterly basis using current item response data stores. These analyses will identify potentially problematic items and imbalances in content to inform ongoing instrument development.</p> <p>As noted above, the tools will be used in an OLES funded intervention pilot following validation. The tests will be also used in SRDC’s Accommodation and Food Service intervention RCT with some 1200 workers and in the ACCC’s National Essential Skills Intervention Study, with learners and workers.</p> |
| <b>French Versions</b> | <p><b>Will the francophone assessment versions be developed from scratch based on tasks and materials identified by a francophone needs analysis - or will they be a translation or adaptation of the Anglophone Scaffold?</b></p> <p>Both methods are being or will be used for creating items. Many of the French items will be translated from existing English versions in order to maintain statistical comparability between the English and French versions, but original French items will also be produced. Many of the original French items will be later translated into English, to establish a two-way linkage.</p> <p>The idea of a "needs analysis" is not applicable to the framework used by the assessments; the framework has been shown to have strong validity across many different languages and even character sets around the world. Fidelity to the framework is more important in ensuring validity of the results than cultural relevance. The tests are constructed to minimize the influence of culture-specific factors so that the results may be interpreted with respect to literacy development as opposed to cultural fluency. In addition, the inclusion of ALL items, which Statistics Canada has given us permission to use, provides a link to items developed in 26 countries and 16 languages including French and, through linkage to the IALS and PIAAC scales, empirically validated links to some 50 countries.</p> <p>Our interest is in individuals’ ability to deal with tasks that are unfamiliar to them. Thus the framework predicts the difficulty of the tasks; we sample the determinants of difficulty systematically and then summarize results to get a score we can interpret as a general indication of a respondent’s transferable skill level. Thus the measures are designed to transcend language and culture.</p>   |