Educ. Sci. **2013**, *3*, 172-192; doi:10.3390/educsci3020172

OPEN ACCESS

education sciences

ISSN 2227-7102

www.mdpi.com/journal/education

Article

Improving Science Assessments by Situating Them in a Virtual Environment

Diane Jass Ketelhut [1,*], Brian Nelson [2], Catherine Schifter [3] and Younsu Kim [4]

[1] Teaching and Learning, Policy and Leadership Department, College of Education, University of Maryland, Maryland 20742, USA

[2] Informatics Department, School of Computing, Informatics and Decision Systems Engineering, Arizona State University; Arizona 85281, USA; E-Mail: brian.nelson@asu.edu

[3] Psychological, Organizational, and Leadership Studies, College of Education, Temple University, Pennsylvania 02144, USA; E-Mail: ccs@temple.edu

[4] Curriculum and Instruction Department, Mary Lou Fulton Teachers College, Arizona State University, Arizona 85287, USA; E-Mail: yskim1@asu.edu

* Author to whom correspondence should be addressed; E-Mail: djk@umd.edu; Tel.: +1-301-405-3324; Fax: +1-301-314-9589.

Received: 22 January 2013; *in revised form: 2 April 2013 / Accepted: 22 May 2013 / Published: 30 May 2013*

Abstract: Current science assessments typically present a series of isolated fact-based questions, poorly representing the complexity of how real-world science is constructed. The National Research Council asserts that this needs to change to reflect a more authentic model of science practice. We strongly concur and suggest that good science assessments need to consist of several key factors: integration of science content with scientific inquiry, contextualization of questions, efficiency of grading and statistical validity and reliability. Through our Situated Assessment using Virtual Environments for Science Content and inquiry (SAVE Science) research project, we have developed an immersive virtual environment to assess middle school children's understanding of science content and processes that they have been taught through typical classroom instruction. In the virtual environment, participants complete a problem-based assessment by exploring a game world, interacting with computer-based characters and objects, collecting and analyzing possible clues to the assessment problem. Students can solve the problems situated in the virtual environment in multiple ways; many of these are equally correct while others uncover misconceptions regarding inference-making. In this paper, we discuss stage one in

the design and assessment of our project, focusing on our design strategies for integrating content and inquiry assessment and on early implementation results. We conclude that immersive virtual environments do offer the potential for creating effective science assessments based on our framework and that we need to consider engagement as part of the framework.

Keywords: assessment; scientific inquiry; immersive virtual environments; middle school

1. Introduction

The National Research Council [1] in the United States suggests that "Assessments that resonate with a standards-based reform agenda [need to] reflect the complexity of science as a discipline of interconnected ideas and as a way of thinking about the world" [p.12]. We concur and believe that current assessments of science fall short of this goal due to their format and emphasis on content only, context-free questions. Testing a series of isolated facts offers students a view of science at odds with the NRC recommendation and with real-world scientific practices, a view that foregrounds content over inquiry. Tests such as these cannot help guide instruction towards an overall goal as expressed in the new conceptual framework of science education that "*all* students have some appreciation of the beauty and wonder of science" [2]. Instead of separate questions on aspects of content and scientific inquiry, tests should include integrated and contextualized assessments of scientific concepts that focus on practices as well as content knowledge. We suggest that assessments embedded in immersive virtual environments offer one approach to fulfilling the NRC's recommendation for assessments to "reflect the complexity of science."

Our current project, *SAVE Science*, is an NSF-funded study developing an innovative system for contextualized, authentic assessment of learning in science. In *SAVE Science*, we are creating, implementing, and evaluating a series of computer-based modules for assessing both science content and inquiry in the middle grades. The modules, based in an immersive virtual environment (IVE), are designed to enable students to perform a series of assessment tasks that provide data about how well they have mastered and can apply content knowledge and inquiry skills taught via their regular classroom curricula. We hypothesize that through careful design of the virtual environment-based assessments, data can be collected and analyzed to produce meaningful and accurate inferences about student learning that provide additional insights about student understanding beyond what is possible from more traditional assessments.

Our first step in the *SAVE Science* project, reported in this paper, was to design such an assessment, integrating science content and inquiry. This paper reports on our initial usability study and our design strategies towards investigating the first two aspects of our integration, conceptualization, efficiency and statistical reliability/validity (ICES) framework, described later in this paper.

Our research questions, specific to this paper, are:

1. How can an IVE be designed for assessment with emphasis on integrating scientific inquiry with content?

2. Once designed, how is this IVE perceived by students and teachers in terms of engagement and usability?
3. What is the impact of the visual context on students' ability to demonstrate science learning?
4. What evidence is there in what students do and say that gives insight towards their understanding of scientific inquiry and content?

2. Theoretical and Research Context

2.1. Scientific Content and Inquiry

Prior to the twentieth century, science education primarily focused on the content of science, what Dewey called the outcomes of science, leaving the practice of inquiry to the scientists themselves [3–5]. An early scientific inquiry advocate, Dewey campaigned for an integration of content with process so that learning science better reflected science as practiced. By the middle of the twentieth century, the focus of science began to be more about relationships than 'uncovering truths', which led to a reinvigorated focus on teaching the processes of scientific inquiry in the K-12 classroom. The view today is that science cannot be understood, learned, or taught if content is separate from process [3]. The original National Science Education Standards (NSES) [6] suggested that students should understand inquiry as a multifaceted activity that involves actively making observations, formulating hypotheses, gathering and analyzing data, and forming conclusions from that data. Doing less than that minimizes the complexity that is science, and creates misconceptions around what constitutes science and how it develops [7]. For example, Klahr has researched the way science is conducted [8] and created a model - *The Scientific Discovery as Dual Search*-which states that scientific processes are conducted within a specific problem space and embedded in scientific content understanding.

As one example of how some are interpreting these recommendations, the Massachusetts State science standards call for inquiry and experimentation to be guiding principles that are integrated seamlessly into content strands [9]. As a historical aside, earlier drafts of the Massachusetts science standards actually had scientific inquiry as a specific set of standards, but it was argued that scientific inquiry was too important to be viewed as an item to be completed and checked off. Therefore, it was added as an overall principle whose impact should be considered on all that was accomplished in the science classroom.

While the research on the impact of teaching scientific inquiry is still under debate in part due to definitional differences of classroom-based scientific inquiry [10], we choose here to focus on studies that investigate the impact of integrating learning of content with scientific inquiry as defined above by the NSES and as the Massachusetts standards suggest be done. These studies indicate that integrating content and inquiry can positively impact both affective aspects as well as learning. For example, Gibson and Chase [11] found that a scientific inquiry-based summer science program had a long-term positive impact on attitudes toward science and interest in science careers for middle school children, even controlling for initial interest in a scientific career. One of us [12] found that after-school and summer scientific inquiry-based experiences also positively impacted the growth of scientific career interest among middle school students. Other large-scale research indicates that integrating scientific inquiry with content improves learning and retention [13,14] even as measured by standards-based

assessments [15,16]. Given that the research indicates that integrating scientific inquiry with content impacts learning and career interest, it is particularly important that this integration happens in the early grades since a student's decision to pursue a scientific career seems to be formed by the eighth grade [17].

2.2. Developing and Assessing Students' Scientific Inquiry and Content

Unfortunately, the goal of integrating inquiry with content in the science classroom is difficult. The new conceptual framework for science education draws attention to the fact, indicating that currently "[science education] emphasizes discrete facts with a focus on breadth over depth, and does not provide students with engaging opportunities to experience how science is actually done" [2; p1]. Indeed, in one study, 80% of K-8 science classrooms in the U.S. were found not to teach science content integrated with inquiry [18]. The obstacles to teaching content with inquiry are the subject of much research [19]. One of the identified problems is the impact that standardized testing has on classroom practice, the oft-mentioned teaching to the test. Numerous reports have indicated a need to match assessments to the curriculum (e.g., [20,21]). Thus, if students are supposed to learn with scientific inquiry, then their assessments should also be inquiry-based. In that case, teaching to the test is exactly what teachers should do as the tests reflect what should be taught. Further, if classroom instruction should be contextualized with real world connections as is typically recommended [21], then assessments should be similarly contextualized. This idea is strongly rooted in the tenets of situated theory, which states that learning happens best when conducted in the environment in which it is to be used [22,23].

Currently, the policy climate in the United States puts the burden of assessment on standardized tests. However, these tests often do not reflect what should be taught in the classroom or give a full picture of what a student knows or understands about the complexity of science [24]. Research indicates that students tend to take and pass science tests, but often are not able to understand larger concepts, which typically are not tested on multiple-choice tests [25]. In addition, the format of these tests makes it difficult to assess scientific inquiry as defined by the NSES since inquiry involves higher order skills that are not easily measured by multiple-choice tests [26,27]. Consequently, students are frequently assessed on whether they understand terms such as "hypothesis" or "control," while in-depth assessment of their abilities to formulate questions and hypotheses, and design and analyze experiments, is neglected [28]. In these cases, teaching to the test undermines the recommendations and directives of various state and federal policy doctrines. Concurring, the Carnegie report [20] suggests that the current testing system in the United States focuses heavily on assessing knowledge and interpretation to the detriment of more scientific inquiry-based topics.

Some state and federal assessments have tried to address this problem by including detailed open-ended questions. However, in order to set the context for the question, lengthy text-based scenarios are often included. These questions then rely on students' reading abilities as much as on their science knowledge. For example, one released 2009 NAEP scientific-inquiry based open-ended question asks the following:

> "Janet has four identical containers. In each container there are 200 grams of a different colored sand, as shown below. All the sand is at the same temperature and has the same grain

> size. Janet leaves the containers out in the full sun for three hours. Then she measures the temperature of the sand in each container. Her results are shown below. Explain why the temperature of the sand in each container is different."
>
> (http://nces.ed.gov/nationsreportcard/itmrlsx/search.aspx?subject=science)

The readability scores for this introduction based on the SMOG test is 10th grade, and yet this question is on the 4th grade test! In addition, the question includes a set of pictures and a data table to analyze. Clearly this question is testing reading as much as it is science, which disadvantages English language learners and poor readers. In Pennsylvania, a state where the population of English language learners has more than doubled in the last 10 years (in Philadelphia alone it tops 15% of the student body) and 33% of all students fail to reach reading proficiency (in Philadelphia 60% fail to reach reading proficiency), reliance on English reading skills calls into question how to use scores on tests like these. Does a student's score indicate their science knowledge or their reading ability? For some students, this is unclear.

This issue is not relegated to U.S. tests alone. It has also been raised regarding the large-scale "Trends in International Mathematics and Science Study" (TIMSS) science assessment, where in one study it was shown that students could correctly answer science questions in an interview that they had answered incorrectly on a TIMSS implementation because of poor reading and English language skills [29].

2.3. Better High Stakes Tests

Shavelson and Baxter [30], working from a premise that good assessment is directly linked to learning, compared five different forms of assessment of scientific inquiry—conducted to solve problems in specific content areas—to what they believed was the gold standard: direct observations of students solving problems. Their five assessment formats were: (1) lab notebooks, (2) computer simulations, (3) short answer paper and pencil questions, (4) multiple choice questions derived from practice, and finally (5) decontextualized standardized test questions of science achievement. Shavelson and Baxter found that the design of the non-test assessments was time-consuming but could result in high-quality assessments of learning. They also found that multiple-choice tests seem to tap into a different aspect of learning than other forms of assessment, as the correlation between student outcomes on these tests and the performance assessments was only moderate. They further suggest that each assessment format taps into a different facet of student understanding, and that to get a full understanding, multiple pathways should be used. This last piece of advice portends the recommendations in the Carnegie report, nearly 20 years later.

Clearly, the question then becomes: what would a better science test look like? We propose that there are four conditions for designing better science assessments. First, the test must assess scientific inquiry in the context of content. While understanding how scientists work is itself a content area, it is only one aspect of scientific inquiry [6]. Students should also be asked to use science practices in a natural way to show a deeper understanding of what scientists do and why they do it. It is impossible to assess this latter aspect of scientific inquiry without embedding its assessment in content. Second, the theory of situated cognition would argue that students will have a difficult time applying their classroom-based understanding of science inquiry and content to the decontextualized questions found

on a multiple choice test. On typical paper-and-pencil tests, the material is not only isolated from the context in which it was learned, but also from where it will be used [31]. Thus, our second condition is that assessments should contextualize the question. Third, the tests must be time and cost efficient for grading. The results need to be provided to schools, teachers, parents, and students in a timely fashion if they are to inform teaching and learning. Fourth, they must be able to show evidence of reliability and validity. Students' scores on tests should reflect their full understanding of science, not their reading abilities or any single facet of their science knowledge. These four conditions—integration, contextualization, efficiency, and statistical reliability/validity ("*ICES"*)—together can create a litmus test for good high stakes science assessment.

To address some of these conditions, alternative methods of assessment have been promoted [19]. However, in most cases, such methods are unable to address all four of our conditions. For example, the Carnegie report suggests that portfolios can be used to integrate assessment of scientific inquiry with content and clearly situate the assessment in the environment in which learning happened (the classroom). Thus, the use of portfolios does indeed address the first two of our conditions. However, portfolios require time-intensive hand grading by well-trained assessors, violating the third condition, and because they are hand-graded, portfolios suffer from a lack of reliability and, depending on how they are structured, issues of validity as well [1,32]. Therefore, while useful in many ways, portfolios fail as good high stakes science tests based on our four conditions.

In 2009, the NAEP outlined their framework for an approved science assessment. The new framework highlights a number of key features, among which are the inclusion of questions requiring inquiry skills and the piloting of interactive computer tasks [33]. Unfortunately, while this is an improvement, even the computer tasks fail at the least on providing contextual clues. One new option that is being considered is using immersive virtual environments as a platform for science assessments that might meet these four conditions.

2.4. Immersive Virtual Environments and Assessment

In an attempt to move beyond cookbook inquiry instruction toward a comprehensive approach for science learning, a growing number of researchers are turning to immersive virtual environments (IVEs). These game-like environments enable the situating of science inquiry practices and content in realistic contexts that have been shown to be engaging for students and beneficial for learning—particularly for students who do not do well with more traditional science instruction [19,34, 35].

IVEs are often created as simplified multimedia simulations of realistic situations and places. Players navigate IVEs, represented by an avatar, wandering through virtual countryside and cities, viewing landscape, entering buildings, driving vehicles or riding animals, and exploring. IVEs typically utilize a back-end database that records all interactions that take place within the environment, producing data from students that can be analyzed to infer evolving levels of competency around science inquiry and concepts [36].

IVE-based science curriculum has the advantage of placing problems in an authentic context for students to solve while providing meaningful information on patterns of learning over time to both students and teachers. Steele [37] reports that if students can learn to connect any concept they are learning to real-world situations, not only will it make the concept more meaningful but also help make

it easier to understand and remember. Part of the promise of IVEs is their capability to create immersive experiences with problems in contexts similar to the real world. In particular, research indicates that using IVEs for learning and assessment offers additional details about student understanding, giving information about students' strategies in solving the problem plus their solutions [38].

Research is beginning to emerge on the question of how well IVEs can be used to situate assessments. For example, Shute *et al.* [39] explore the idea of conjoining immersive games with embedded assessments to create what they label "stealth" formative assessments. Shute and her colleagues argue that player interactions in a game can be assessed in real-time using probability analysis techniques. The sum of these interactions over the course of a game adds up to meaningful evidentiary records of understanding of the content and processes taught in the game. Game players can be continuously and invisibly assessed as they work through series of challenging tasks situated seamlessly into game play and narrative [40].

In a study of the River City IVE curriculum, one of us [38] used assessment data gathered on students' actions in River City to investigate changes in their scientific inquiry processes over four visits to the virtual town, and to explore how their self-efficacy in science impacted those changes. The study found that on average students increased their data-gathering activities with each visit to the IVE, and that initially self-efficacy levels correlated with the number of data-gathering actions in which they engaged. In other words, high self-efficacy students engaged in more data gathering than students with low self-efficacy, as self-efficacy theory would predict. However, by the end of their time in the River City IVE, initial student self-efficacy no longer correlated with data-gathering behaviors. As another example of how embedded assessment techniques can be used to understand more about student inquiry learning in virtual environments, Ketelhut and Dede [41] discovered in further analysis that, on average, increases of only eight inquiry activities in the River City IVE were associated with an increase of 5% on science content scores ($p < 0.02$).

2.5. SAVE Science

Building off our work in River City and enacting similar ideas to those proposed by Shute, we are developing assessments of scientific inquiry embedded with assessments of content. The *SAVE Science* (*Situated assessment using virtual environments for science content and inquiry*) project utilizes our four conditions for good high stakes assessment to design and implement a series of IVE-based modules that assess both science content and inquiry taught in the middle grades science classroom. In the *SAVE Science* project, middle school students have an overall goal of uncovering the likely contributors to a series of problems facing a virtual world (e.g., sick farm animals, weather-related disasters, and urban planning). Students enter the *SAVE Science* IVE multiple times over the course of a school year, conducting a new inquiry quest on each visit, being assessed on understanding and application of content just studied through classroom-based instruction.

The *SAVE Science* project makes use of an IVE called Scientopolis that we have developed. Throughout the assessment modules, student interactions are recorded in a database allowing us to analyze both explicit answers to questions posed by the lead computer-based character that the students meet in each module as well as students' actions in coming to those answers. Automatically recorded in-world interactions are used to understand students' problem-solving behaviors, and are

recorded in the database with a location and time-stamp. This allows us to assess students' processes in solving the posed problems in addition to their solutions.

While automatically recorded process data helps us assess students' understandings of material and concepts in the virtual world, it is important to assess students' ability to articulate and apply what they have learned. Therefore, each assessment module ends with a series of embedded questions about what students conclude and the evidence for that conclusion. These are integrated into the storyline as help for the main computer-based character.

Further, we are looking to see if our assessments provide new information about student understanding of the content and skills being assessed beyond that which the schools can learn from their in-class tests. The participating school district has several different standardized assessments of their curriculum that students are required to take. Using questions from these district assessments, we have designed the SAVE Science assessment modules to clearly show, either through student behaviors, articulation or both, whether they can answer those questions.

2.6. SAVE Science and the Four Assessment Conditions

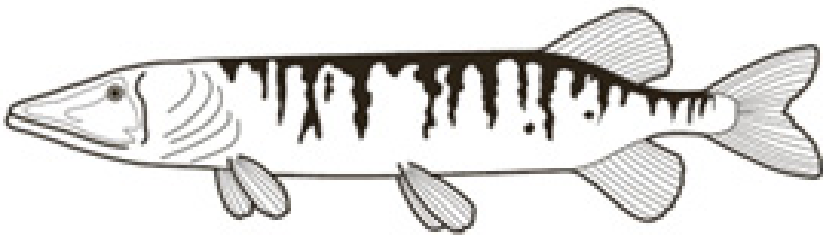
The design of the assessments takes into consideration our four high stakes science assessment conditions, *ICES*, described above. Here we will describe how these four conditions guide our design work. Later in the paper, we will detail how these are effectuated in one module, "Sheep Trouble." The first condition requires that we *integrate assessment of content with scientific inquiry*. By creating problem-based assessments, students must use their scientific inquiry skills to solve the content-based problem, while the tools they choose and how they interpret data gathered through use of those tools relies on their understanding of the content. For example, if you are given a barometer and a thermometer and asked to predict if a storm is coming, you would choose the barometer to see if the air pressure is dropping indicating an arriving storm. Choosing the thermometer would imply a lack of understanding of air pressure changes associated with weather fronts.

In SAVE Science, we carefully outline specific content and scientific inquiry objectives for each module and design the scenario with gaps that students must fill with their own knowledge. Their success in doing so, as indicated by both their processes in the world and their articulation at the end, reflects how well they understand those specific objectives.

The second condition, *contextualization*, is difficult to achieve on a multiple-choice text-based test, but is actually the main strength of virtual environments. For example, on a 2010 Pennsylvania System of School Assessment (PSSA) standardized test [42], the following question on adaptation is asked baldly with little context:

> "The picture (*see Figure 1*) below shows a type of fish that is adapted to live in the weedy areas of freshwater lakes. How is this fish adapted to live in the weedy areas in freshwater lakes?

Figure 1. The fish.



a) The upper fin looks like another fish.
b) The lower fins look like the legs of a turtle.
c) The stripes of the fish look like plants in the water.
d) The mouth of the fish looks like the bottom of a lake."

To interpret this question, students would need to imagine a freshwater lake with weeds, thus testing their ability to imagine the lake as well as their understanding of adaptive coloration (the correct answer is "C"). For urban-based students without experiences with freshwater lakes, this becomes a much more difficult question than for their suburban or more well-travelled peers. However, if this same question was asked embedded in a virtual environment, the IVE could include a freshwater lake with weeds that has the appropriate fish swimming around. Students could observe the fish swimming among the pond weeds before answering the question. Their difficulty in finding the fish (since the stripes camouflage it) would give them the contextualized clues needed to apply their understanding of adaptive coloration to answer the question.

Efficiency is the hallmark of the third condition. Our purpose in *SAVE Science* is to create an assessment that integrates science content with scientific inquiry, and has the ultimate goal of data mining the recorded student actions and communications in order to produce an assessment report. Since all students' actions and answers are recorded in a database, we envision ultimately having the virtual environment 'score' students' progress through the assessment module and deliver both a quantitative score on various measures and key phrases from the students' answers for teachers' formative use. Progress toward this goal is ongoing. In this paper, we report on our first interpretation of how successful our design is in creating tasks that could be automatically scored.

The most challenging conditions to ascertain with IVE-based assessment are reliability and validity. For if we accept that current tests are fraught with difficulties, then we may not want to look for correlations between the results of our IVE-based assessment and results from current tests—the traditional way of assuring validity—as we are suggesting that there are validity issues with current tests. Therefore, in the SAVE Science modules we are building in triangulation data to allow us to ascertain if computer generated scores match up with other measures. These other measures include: classroom observations (as discussed above this was the gold standard according to Shavelson and Baxter), interviews and focus groups (the measure used in the previously measured TIMSS study), and open-ended questions where students must explain their answers in their own words. Using content experts scoring these last, we can look for content validity in our assessment.

3. Methodology

3.1. Research Questions

1. How can an IVE be designed for assessment with emphasis on integrating scientific inquiry with content?
2. Once designed, how is this IVE perceived by students and teachers in terms of engagement and usability?
3. What is the impact of providing visual *versus* textual context for assessment questions on students' ability to demonstrate learning?
4. What evidence is there in what students do and say that gives insight towards their understanding of scientific inquiry and content?

3.2. Site and Sample

Sheep Trouble was the first assessment module to be designed for the SAVE Science project. It was first implemented in a mid-Atlantic school district, classified as a near-urban district. Of the students in this school, 19% are on free or reduced lunch—a proxy for poverty— 18% are racial minorities. The seventh grade science teacher in the sole middle school volunteered one of his classes to participate. The mid-morning class consisted of 23 students, of which 20 were present on the day of the pilot, evenly split between girls and boys with demographics typical for this school. The content of the relationship between biological structure and its function which the Sheep Trouble module assesses had been covered earlier in the semester.

3.3. Procedure

Students met in the computer laboratory for a single class period. Several research staff, including the authors, were in attendance and introduced the project to students by telling them that they were about to take a new kind of science test, that this test would assess their understanding of adaptations, and that they would have the entire period to complete it. Further, they were told that since this was a test, silence would be expected and enforced, but if they had issues with the software they should let us know. After this introduction, students logged in to the module and began trying to solve the problem. As students completed the module, we engaged them as junior researchers to write an evaluation of the test along with their recommendations for improvement. Circuit breakers tripped repeatedly on six computers preventing those students from completing the assessment. The 14 other students completed the assessment in under 20 min; a ten minute class discussion (led by the first two authors on this paper) on what caused the problem with the new sheep ensued. Three students were chosen by the teacher to be interviewed by a regional official, not directly affiliated with the project or the school.

3.4. Design of Sheep Trouble Assessment

"Sheep Trouble" is the initial module in SAVE Science and assesses student understanding of concepts of adaptation and structure/function that underlie beginning speciation. The module is designed so that by interacting with farmers and sheep on a virtual farm, students gain the contextual

understanding they need to apply their classroom learning to more meaningful assessment tasks. We designed the assessment curriculum and the Sheep Trouble module to address our four conditions for a high quality assessment that we highlight below. All interactions within the module are written at a 7th grade reading level using the Flesch-Kincaid Grade Level test [43].

In Sheep Trouble, students enter a medieval-like world (see Figure 2 below) where they are met by a farmer character. The farmer asks students to help him find a scientific explanation for why his recently imported flock of sheep is in poor health. Students discover that many of the townsfolk think that the new sheep are sick due to "bad magic" and that the new sheep must be destroyed quickly before the bad magic spreads to the local sheep. The farmer asks the student-scientists to apply their skills to suggest possible science-based contributors to the new sheep's poor condition. Students are given virtual tools for data collection and can also observe visual clues about the environment itself (for example, most of the newly imported sheep are gathered at the bottom of a hill on the farm, where there is little grass; while the 'original' sheep are mostly at the top of a steep hill in lush grass). Students use a question and answer system to communicate with a farmer and his brother (see Figure 3). They can also interact with a large number of the new and local sheep scattered around the farmyard. The local "original" sheep (sheep from stock that have been raised locally for centuries) are all healthy. Students can use virtual rulers to measure the sheep's legs, body length, and ears, and can access information on sheep weight loss or gain, age and gender. Once students feel they have gathered enough information to form a hypothesis, they return to the farmer and explain their hypothesis for why the new sheep are failing to thrive.

Figure 2. Scientopolis: the SAVE Science world.

Figure 3. The question and answer communication process.

Students are provided with access to information related to the problems through posters, computer agents, and the design of the virtual world. For example, a poster on the farm explains that the new sheep come from a very different geographic locale (a flat, snowy island) from the current farm (hilly, rocky, and dry). By applying their knowledge of scientific inquiry and of adaptation through exploration of the virtual world, students can discover that to get to the best grass, the new sheep need to climb a hill. Students gather evidence and can reach the conclusion that the newly imported sheep are not adapted for this new environment, which requires them to climb steep hills for food. As a result, they are failing to thrive.

In this module, students are being assessed on specific state science standards (from an East Coast state) related to *both content and scientific inquiry* (our first assessment condition). Using the associated state standards, we identify how they are embodied in Sheep Trouble:

1. "Changes in environmental conditions can affect the survival of populations and entire species"—students need to recognize that the new flock of sheep has been transplanted to a very different environment from the one to which they are adapted and using their classroom learning about adaptations, infer that this might be impacting their survivability.
2. "Describe the structures of living things that help them function effectively in specific ways (e.g., adaptations, characteristics)"—students describe the structural differences between the two flocks as indications of adaptations to different environments.
3. "Explain how different adaptations in individuals of the same species may affect survivability or reproduction success"—students explain how the different adaptations impact survivability of the two flocks.
4. "Apply appropriate measurement systems (e.g., time, mass, distance, volume, temperature) to record and interpret observations under varying conditions"—students should choose the appropriate tools (ruler, scale, graphing tool) for gathering, recording and interpreting data.
5. "Interpret data/observations"—students collect data but before they report to the Farmer at the end on what they have found, they must make sense of this data.

6. "Use evidence, such as observations or experimental results, to support inferences about a relationship"—students need to identify the evidence that supports their hypothesized relationship.
7. "Use evidence from investigations to clearly communicate and support conclusions"—students must communicate their findings and conclusions to the Farmer at the end of the test, using evidence they have collected to support their conclusions.

Our second condition, *contextualization*, is derived from allowing students to manipulate an avatar who represents them as a student-scientist and explore the farm. Rather than list the various clues to solve the problem as is done on tests (for example, the previously mentioned fish example lists that it lives in a freshwater pond with weeds), students can see the clues in context (grass at the top of the hill, different shaped sheep, *etc.*) and then apply their understanding of adaptations to gather the appropriate data and interpret the clues.

Efficiency is the third condition and is found in how Sheep Trouble will be graded. Ultimately, we plan for the assessment anchors embedded in the module to be computer-graded and so have built into this various aspects that allow for this. As an example, assessment anchor #5 above states that students must interpret data. In Sheep Trouble, students are provided with a graphing tool. However, to use that tool, students must make choices that show their understanding of how to interpret data. The first choice that students face is whether to include both sheep types on their graph or only one. Choosing both types to graph indicates student knowledge of the role graphing plays in helping them understand what if any differences exist between the two flocks. Their second choice is on what data to graph, and preferably would be only one type of data so that the graph clearly illustrates whether that characteristic differs from new to old sheep. Since these are clickable choices, they are recorded in the database, and a grade report showing whether students can use graphs to interpret their data can be immediately available to teachers.

In Sheep Trouble to gather validity data for this measurement, we ask students through the conversation with the Farmer at the end of the assessment module to explain why they think the problems with the sheep do not have magical causes in several ways: open-ended response, listing three sources of evidence used to reach their conclusions and then ranking that evidence in order of importance. This data, along with the computer generated grade report, will allow us to gather reliability and validity evidence for the computer-generated report. Further, once students finish the module, they are asked to 'debrief' their experience by explaining what they found in writing on a hardcopy report. This report is then submitted to the research team to triangulate with what students did in the environment. Once the entire class is finished, the teachers conduct a class discussion on the module which is recorded and used similarly by the research team.

3.5. Analysis

Qualitative content analysis was used to analyze the data collected for the pilot study which included: students' actions in the module, observers' notes of all comments students made as they explored and completed the module, the students' written evaluation, and the observers' notes of the ten minute class discussion. This data was content analyzed specifically looking for terms and phrases alluding to engagement e.g., "fun", "looking for", "figure out", and other action verbs indicating

engagement in the game module [44]. Transcriptions of each source of student input, both written and verbal, were used to identify evidence of engagement and usability, understanding of scientific inquiry, and aspects of improvement of the module. The student answers to the two questions embedded in the module were used to assess students' knowledge of adaptation content. Two researchers independently reviewed coded the transcriptions, and then compared analysis. Initially, there was a 77% agreement level. Discrepancies were discussed using individual researcher analyses and consensus achieved, leading to a 100% agreement. This process was used for each type of qualitative data used for this analysis to insure internal validity of coding and analysis. In addition, the students' performance in the module, as recorded by the underlying database, allowed the researchers to compare module performance with answers to the two end-of-module questions.

4. Results and Discussion

4.1. Engagement and Usability

In order to evaluate whether students were engaged and had few issues in using the environment, we looked at whether students were active in the module, and interacting with the various objects and computer-based characters. In addition, student comments about their impressions of Sheep Trouble were considered. Overall, we found students were active, interacting with both of the computer-based characters in the IVE, attempting to gather information from the sheep, and interpreting their findings in response to the farmer's questions. In addition, student comments indicated a high level of interest. Table 1 has illustrative student comments on engagement, indicating their enjoyment of the challenge. For example, one typical comment was "*it was a brain puzzle but still lots of fun*." Other comments, column 2 of Table 1, indicate that students were challenged and interested by the problem: "*it was intriguing*." Overall, student comments indicated that students thought of the Sheep Trouble module more as a game than as an assessment. As one student stated, "*It didn't feel like we were just taking a test on a blank screen*."

Table 1. Example student comments (n = 20).

Engagement	Assessment aspects	Overall
it was fun with the evidence that you had to find on the two kinds of sheep	It seemed like a real-life question	I liked how you could interact with the different people
the game was very intriguing. It was a brain puzzle but still lots of fun	it was sort of a challenge	Give a second hint about what the problem is
Its really fun	I think the barriers of the game was too small	I need more stuff to interact with
You get to figure out what's wrong	It was fairly easy	It was really fun because you got to go around and explore why the new sheep were sick

Table 1. *Cont.*

Engagement	Assessment aspects	Overall
the most interesting part was trying to find out what was wrong with sheep.	It was fun and realistic. It didn't feel like we were just taking a test on a blank screen.	Make a bit harder and longer
its real enough looking that I can really get into it.	I think this a great way for students to test their skills	But the story will get old and it would probably be better if there were different challenges

All students were engrossed in the task throughout the class period. While the students were encouraged to be quiet much as during a "real" test, from time to time when they discovered something interesting, they burst out with a comment of "*cool*" or "*did you see that*?" Several students finished the task quickly and then asked if they could continue to explore the IVE. Looking at the data from the database, we found that 11 of the 20 students went back into the virtual world and measured more sheep or talked with a farmer again after completing the embedded questions from the farmer at the end of the quest, supporting our observations in the classroom of their sustained interest. Two of the participants actually completed the embedded final conversation with the farmer twice, but in both cases the second attempt was accompanied with odd answers that in one case, the student identified was just '*for fun*'.

Because this was our first implementation, we were very interested to evaluate how easily students would interact with the assessment module and how well it would run on typical (old) school computers. We found that students had no difficulties moving their avatar around or in interacting with the various objects in the IVE. Indeed, a couple of students figured out how to do different things while in the world beyond the typical avatar movement via arrows on the computer keyboard, including learning how to make their avatar run. Issues of running the software on the school computers were non-existent, although the school's electrical circuits were found to be unable to run a full class of computers simultaneously, resulting in circuit breakers tripping periodically. However, apparently, this was a known and expected problem for teachers in this school. Unfortunately, this meant that only 14 of the 20 students were able to complete the module.

4.2. Evidence for Scientific Inquiry Understanding

One of our participating school districts indicated an interest in knowing if students understood the importance of forming conclusions based on data. This is not currently well-tested in their district-wide assessments from their perspective but it is one of the state standards ("Use evidence, such as observations or experimental results, to support inferences about a relationship") that Sheep Trouble is assessing as described above. The data from this first Sheep Trouble implementation does indicate that we can gather this information. Students made use of the science inquiry methods they learned through regular classroom instruction prior to the module to gather information by measuring the sheep, exploring and talking with the farmers. They then used their gathered data to make an inference about the problem to the farmer at the end of the assessment. Automatically collected data indicated that all students gathered data before reaching conclusions. Some students asked questions of the two non-player characters multiple times, and all observed sheep characteristics.

Further, students used their observations to support their responses to questions from the farmer, indicating their understanding of evidence to support ideas in their answers. For example, when asked if they knew why the sheep were dying, several students said: "*grass is dead in a lot of areas*", "*not the type of land they (new sheep) are used to–mostly flat where they came from while hills here*". Some students did show a confusion of the concept of data as evidence. When asked for their data to support their conclusion, they answered with an inference instead, "*(they) need different grass," "(they are) not eating enough*".

4.3. Knowledge of Adaptation Content

We included two questions from the district science assessment test in an embedded interview given by "Farmer Brown" at the end of the module to see if the visual context provided through the immersive virtual environment to these questions improved students' ability to answer them as compared to the success rate for the district test where the questions are presented via text as story problems. To do this, we slightly modified the questions so that they were about sheep on a farm instead of their original context. For example, one question asks about beginning speciation very generically, "How would different structures most likely benefit two subpopulations going through speciation?" We took this question and situated in on our farm and in our problem: "If these two sets of sheep are undergoing speciation, what is the most likely benefit to each sub-population of the different leg lengths?" No other changes were made to this question. The second question is about adaptation and shows the front half of a head of a tropical-appearing bird, asking: "Based on its physical appearance, what is the most likely function of this bird's beak?" We situated this question in the sheep on the farm and asked, "It looks like the sheep's legs are different lengths. What is the most likely function of the old sheep's shorter legs?" in this situation we also needed to adjust the responses but were careful to match the intent of the original question's choices: "(a) Kicking rocks over. (b) Walking through snowy ground. c. Climbing steep hills. d. Running fast on flat ground".

Fourteen students answered both of these questions. Two students answered both incorrectly, eight students answered one or the other correctly, and four answered both correctly. Our results for the second question on adaptation mirrors those found in the larger school district; however, the percent of our students who answered the speciation question correctly was nearly double that of the school district. In investigating this further, we found that for the adaptation question there were no differences between students who answered correctly or incorrectly in how many sheep they looked at, however, students who answered the speciation question correctly had measured 69% of the sheep they looked at while those who answered incorrectly had only measured 54% of the sheep they looked at ($p = 0.10$). This provides a tentative indication that contextualizing questions does improve students' ability to answer, if students take an active part in the contextualized assessment tasks, however, since our sample was very small and we did not find the same results for the speciation question, we are investigating this further.

4.4. Aspects for Improvement

As can be seen in the third column of Table 1, many students asked for more complexity or difficulty in future modules. Nearly all of the students suggested changes that involved more interactivity and complexity to the story. For example, several students wanted to have the sheep and farmers move naturally and this is a design feature we have since built in to subsequent versions of the module. They also wanted to be able to explore the hypothetical town that the farmer lived near. Subsequent modules allow students to explore different parts of the larger virtual world of Scientopolis.

Several students indicated a desire for more complexity. One student wanted us to "*make the game more challenging*". The intriguing part of that comment was that while we told them that this was a test, the students still overwhelmingly referred to it as a game, despite clearly telling them that it was a test. Indeed, one student in an interview afterwards said, "*[the game] seemed like a real life question. Tests just seem like a made up question on a piece of paper. Even though the game is kind of a made up question*, *it still seems more like a real question*". Clearly, students responded to this assessment more as a fun challenge than as a test.

5. Conclusion and Future Research

The purpose of this study was to investigate the viability of designing IVE-based science tests that integrate content and scientific inquiry assessment. We identified four characteristics, *ICES*, that we think identify high quality science assessments:

1. Integrated content with scientific inquiry as opposed to separate questions
2. Contextualized questions to help student apply their learning
3. Efficient means for grading
4. Statistically reliable and valid assessments

We designed the Sheep Trouble assessment module around these four characteristics as indicated above. Study data indicates that we have been successful at integrating assessment of content with scientific inquiry and in contextualizing the problem. Further, we have evidence that the contextualization was helpful for students in showing us what they knew.

The strength of students' responses around engagement, however, has caused us to rethink our four characteristics. Typically, engagement is not an aspect as seriously considered in test design as it is in curriculum design. Given that we were designing a test inside an environment that evoked gaming, we planned for visual and interactive engagement of students. For example, the sheep are more cartoon-like than real, the farm has interesting places to explore, and there are several different avatar designs. However, game designers often state that students see embedded assessments in computer games as the tests they are. In an unpublished study of a summer camp built around the River City IVE, one of us found students avoiding interactions with a computer-based newspaper reporter (the embedded assessor) because they said they did not want to stop exploring the IVE to take a test. In other words, the students recognized the embedded character as a thinly disguised test. However, keeping students' attention on a test embedded in a virtual environment means that they are more likely to engage in the difficult cognitive activity required by well-designed test questions. This

overlooked characteristic of well-designed science assessments, *engagement*, therefore becomes our fifth aspect in the newly named ICESE framework.

The implementation described in this paper improved our understanding of how to create efficient assessments of scientific inquiry and content. In Sheep Trouble, we found it relatively easy to identify students who understand what constitutes evidence and how to interpret that data. What was much more difficult was discerning gradations of student understanding related to the embedded assessment. In this study, we had evidence that higher levels of engagement with the module represented by increased levels of sheep investigations predicted for better scores on traditional multiple choice questions, but that connection was not fine-grained enough to suggest that these type of tests can assess student knowledge of concepts along with scientific inquiry. We have used insights from this first pilot study to redesign Sheep Trouble and other subsequent assessment modules to include embedded prompts that will give us clearer evidence of student understanding of the integrated content and inquiry assessed in this module.

Based on findings from this initial study, and on subsequent Scientopolis implementations, our future research in this area will focus on ICESE framework development and validation. In the study reported here, we focused on the first two aspects of this framework: designing modules that ***i***ntegrate assessment of science content with science inquiry and investigating the impact of ***c***ontextualizing assessment. Future studies will investigate the other three components of the framework: ***e***fficiency and ***s***tatistical validity of our modules, and the impact of creating ***e***ngaging assessments on students' self-efficacy and interest in science. To explore these components, we plan to (1) investigate our system's ability to automate evaluation of learners' evolving state of understanding as indicated by their actions in the IVE assessment, and provide clear formative and summative evaluation data back to teachers and students; (2) redesign our existing IVE-based assessment modules based on sophisticated data mining findings to bolster assessment validity, while investigating validity of the redesigned assessments across diverse traditional and non-traditional settings; and (3) explore factors affecting student engagement, teacher uptake of the technology, and sustainability of IVE-based assessments by implementing the redesigned modules across delivery platforms (mobile and desktop).

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 0822308.

Conflict of Interest

The authors declare no conflict of interest.

References

1. National Research Council. *Classroom Assessment and the National Science Education Standards*; The National Academies Press: Washington, DC, USA, 2001.
2. National Research Council. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*; The National Academies Press: Washington, DC, USA, 2012.

3. National Research Council. *Inquiry and the National Science Education Standards: A Guide for Teaching and Learning*; The National Academies Press: Washington, DC, USA, 2000.
4. Bybee, R. Teaching Science as Inquiry. In *Inquiring into Inquiry Learning and Teaching in Science*; Minstrel, J., Van Zee, E.H., Eds.; American Association for the Advancement of Science (AAAS): Washington, DC, USA, 2000; Chapter 3, pp. 20–46.
5. Dewey, J. *Democracy and Education (First free press paperback 1966 ed.)*; Macmillan Company: New York, USA, 1944.
6. National Research Council. *National Science Education Standards: Observe*, *Interact*, *Change*, *Learn*; The National Academies Press: Washington, DC, USA, 1996.
7. Rutherford, F.J. Vital Connections: Children, Books, and Science. In *Vital Connections: Children*, *Science*, *and Books*; Saul, W., Jagusch, S.A., Eds.; Library of Congress: Washington, DC, USA, 1991; pp. 21–30.
8. Li, J.; Klahr, D. The Psychology of Scientific Thinking: Implications for Science Teaching and Learning. In Teaching Science in the 21st Century, 1st, ed.; Rhoton, J., Shane, P., Eds.; National Science Teachers Association: Arlington, VA, USA, 2006; pp. 307–327.
9. Massachusetts Department of Education. Massachusetts Science and Technology/Engineering Curriculum Framework [Electronic Version], 2006. Available online: http://www.doe.mass.edu/frameworks/scitech/1006.pdf (accessed on 15 January 2008).
10. Anderson, R. Reforming science teaching: What research says about inquiry. *J. Sci. Teach. Educ.* **2002**, *13*, 1–12.
11. Gibson, H.; Chase, C. Longitudinal impact of an inquiry-based science program on middle school students' attitudes toward science. *Sci. Educ.* **2002**, *86*, 693–705.
12. Savage, L.; Ketelhut, D.J.; Varnum, S.; Stull, J. *Raising Interest in Science Careers through Informal After-School Experiences*; Paper presented at the National Association for Research in Science Teaching: Philadelphia, PA, USA, 2010.
13. Leonard, W.H.; Speziale, B.J.; Penick, J.E. Performance assessment of a standards-based high school biology curriculum. *Am. Biol. Teach.* **2001**, *63*, 310–316.
14. Alberts, B. Some Thoughts of a Scientist on Inquiry. In *Inquiring into Inquiry Learning and Teaching in Science*; Minstrel, J., Van Zee, E.H., Eds.; American Association for the Advancement of Science: Washington, DC, USA, 2000; Chapter 1, pp. 3–13.
15. Blanchard, M.; Sutherland, S.; Osborne, J.; Sampson, V.; Annetta, L.; Granger, E. Is inquiry possible in light of accountability? A quantitative comparison of the relative effectiveness of guided inquiry and verification laboratory instruction. *Sci. Educ.* **2010**, *94*, 577–616.
16. Marx, R.; Blumenfeld, P.; Krajcik, J.; Fishman, B.; Soloway, E.; Geier, R.; Tal, R. Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *J. Res. Sci. Teach.* **2004**, *41*, 1063–1080.
17. Tai, R.; Liu, C.; Maltese, A.; Fan, X. CAREER CHOICE: Enhanced: Planning early for careers in science. *Science* **2006**, *312*, 1143–1144.
18. Jorgenson, O.; Vanosdall, R. The death of science: What we risk in our rush towards standardized testing and the three R's. *Phi Delta Kappan* **2002**, *83*, 601–605.
19. Nelson, B.; Ketelhut, D.J. Designing for real-world inquiry in virtual environments. *Educ. Psychol. Rev.* **2007**, *19*, 265–283.

20. Carnegie Corporation. *The Opportunity Equation: Transforming Mathematics and Science Education for Citizenship and the Global Economy*; Carnegie Corporation of New York: New York, NY, USA, 2009.
21. Krajcik, J.S.; McNeil, K.L.; Reiser, B.J. Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Sci. Educ.* **2007**, *92*, 1–32.
22. Lave, J.; Wenger, E. *Situated Learning: Legitimate Peripheral Participation*; Cambridge University Press: New York, NY, USA, 1991.
23. Brown, J.S.; Collins, A.; Duguid, P. Situated cognition and the culture of learning. *Educ. Res.* **1989**, *18*, 32–42.
24. Songer, N.; Wenk, A. *Measuring the Development of Complex Reasoning in Science*; Paper presented at the American Education Research Association (AERA) Annual Meeting: Chicago, April 25, 2003.
25. Michael, J. Conceptual assessment in the biological sciences: A National Science Foundation sponsored workshop. *Adv. Physiol. Educ.* **2007**, *31*, 389–391.
26. Resnick, L.B.; Resnick, D.P. Assessing the Thinking Curriculum: New Tools for Educational Reform. In *Changing Assessments: Alternative Views of Aptitude*, *Achievement*, *and Instruction*; Gifford, B., O'Connor, M., Eds.; Kluwer Academic Publishers: Norwell, MA, USA, 1992; pp. 37–75.
27. Southerland, S.A.; Smith, L.K.; Sowell, S.P.; Kittleson, J.M. Resisting unlearning: Understanding science education's response to the United States' national accountability movement. *Rev. Res. Educ.* **2007**, *31*, 45–77.
28. National Research Council. *America's Lab Report: Investigations in High School Science*; National Academies Press: Washington, DC, USA, 2005.
29. Harlow, A.; Jones, A. Why students answer TIMSS science test items the way they do. *Res. Sci. Ed.* **2004**, *34*, 221–238.
30. Shavelson, R.J.; Baxter, G.P. What we've learned about assessing hands-on science. *Educ. Leadersh.* **1992**, *49*, 20–25.
31. Behrens, J.T.; Frezzo, D.; Mislevy, R.; Kroopnick, M.; Wise, D. Structural, Functional, and Semiotic Symmetries in Simulation-Based Games and Assessments. In *Assessment of Problem Solving Using Simulations*; Baker, E., Dickieson, J., Wulfeck, W., O'Neil, H., Eds.; Lawrence Erlbaum Associates: New York, NY, USA, 2007.
32. Stecher, B.M.; Klein, S.P. The cost of science performance assessments in largescale testing programs. *Educational Evaluation and Policy Analysis*, **1997**, *19*, 1–14.
33. National Assessment Governing Board (NAGB). *Science Framework for the 2009 National Assessment of Educational Progress*; NAGB, U.S. Department of Education: Washington, DC, USA, 2008.
34. Barab, S.; Arici, A.; Jackson, C. Eat your vegetables and do your homework: A design based investigation of enjoyment and meaning in learning. *Educ. Technol.* **2005**, *45*, 15–20.
35. Nelson, B. Exploring the use of individualized, reflective guidance in an educational multi-user virtual environment. *J. Sci. Educ. Technol.* **2007**, *16,* 83–97.

36. Nelson, B.; Erlandson, B.; Denham, A. Global channels for learning and assessment in complex game environments. *Br. J. Educ. Technol.* **2011**, *42*, 88–100.
37. Steele, M. Teaching science to middle school students with learning problems. *Sci. Scope* **2005**, *29*, 50–51.
38. Ketelhut, D.J. The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in river city, a multi-user virtual environment. *J. Sci. Educ. Technol.* **2007**, *16*, 99–111.
39. Shute, V.J.; Ventura, M.; Bauer, M.I.; Zapata-Rivera, D. Melding the Power of Serious Games and Embedded Assessment to Monitor and Foster Learning: Flow and Grow. In *The Social Science of Serious Games: Theories and Applications*; Ritterfeld, U., Cody, M.J., Vorderer, P., Eds.; Routledge/LEA: Philadephia, PA, USA, 2009; Chapter 18, pp. 295–321.
40. Clark, D.; Nelson, B.; Sengupta, P.; D'Angelo, C. Rethinking Science Learning through Digital Games and Simulations: Genres, Examples, and Evidence. An NAS Commissioned Paper. Available online: http://www7.nationalacademies.org/bose/Clark_Gaming_CommissionedPaper.pdf (accessed on 7 October 2009).
41. Ketelhut, D.J.; Dede, C. *Alternative Assessments of Students' Understanding of Scientific Inquiry via a Multi-User Virtual Environment*, Invited Paper Presented at the Distributed Learning and Collaboration (DLAC-II) Symposium, Singapore, Singapore, 11 June 2007.
42. The Commonwealth of Pennsylvania. Pennsylvania System of State Assessment, 2011. Available online: http://www.portal.state.pa.us/portal/server.pt/community/pennsylvania_system_of_school_assessment_(pssa)/8757/resource_materials/507610 (accessed on 30 December 2011).
43. Du Bay, W. 2004, the Principles of Readability. Available online: http://www.nald.ca/library/research/readab/readab.pdf (accessed on 21 March 2013).
44. Creswell, J.W. *Qualitative Inquiry and Research Design*; Sage: Thousand Oaks, CA, USA, 1998.