Independent Living Resources, Inc.

411 Andrews Road, Suite 230

Durham NC 27705

Final Research Report:

SBIR Phase I

Adolescent Real World Simulation

(ARWS)

SBIR Grant:

1R43MD005173-01

October 2009 – July 2010

Real World Simulation Phase I Evaluation Report – October 2009-July 2010

Summary – The feasibility of the Phase 1 approach is strongly supported, as demonstrated by:

- robust, statistically significant, positive knowledge-gain scores,
- the absence of time effects,
- the absence of any independent effect of the pretest on posttest scores,
- the absence of any a-priori between-group demographic differences, and
- positive affective feedback from study participants.

These findings, and the results of each of the statistical analyses, are presented in detail throughout the body of this report.

Evaluation Logic Model - The evaluation of Phase 1 focused on the users' perceptions of the training materials and whether they were effective in imparting new information to training participants. Accordingly, the Phase I feasibility plan focused on fundamental issues of satisfaction and knowledge.

- Satisfaction – pertains to satisfaction with the materials, the look and feel of the website, and ease of navigation. Satisfaction, relevance of material, and user's motivation was assessed with a combination of open-ended questions and 7-point Likert scales constructed to rate affective dimensions.

- Knowledge– knowledge gain was measured objectively and subjectively. Knowledge was measured objectively by a short direct assessment focused on the competencies/learning objectives. Evaluation staff used module content to develop objective test items. Knowledge gain was assessed using pretest/posttest differences, and subjectively by using open-ended questions asking participants about the relevance of the materials and the extent to which they learned from the material comprising the training module.

Design - A pretest/posttest with comparison group design was used in the study. The comparison group was actually a delayed-treatment group, used in the design to test for time effects and the effect of taking the pretest. This group received the pretest at the same time as the treatment group, and received the posttest (on knowledge items) at the same time as the treatment group, to account for time and testing. The comparison group was then given access to the website, and experienced the training. They were then posttested again on knowledge, and also queried on satisfaction and motivation to coordinate a real world simulation. Thus, for both groups, user satisfaction data was collected at the end of the learning segment, and at the end of the Phase I introductory training program, but the posttesting of the delayed treatment group lagged the treatment group by approximately three weeks. The following table illustrates the study timeline and the "controlling" features of the delayed treatment group.

Table 1. Timeline of study test conditions for Treatment and Comparison groups.

Time period →	Month 5, Day 1	Month 5, Day 21	Month 6 Day 14
Treatment Group	Pretest followed by web training	Posttest on knowledge and satisfaction	
Comparison (Delayed Treatment) Group	Pretest followed by 3-week waiting period	Posttest on knowledge, only, followed by web training	Posttest on knowledge and satisfaction

Both open-ended and close-ended questions were employed to learn about users' reaction to the material, quality and presentation of the material. Descriptive analysis focuses on identifying course material and design elements that work well together, and understanding how other aspects of the training can be improved.

Knowledge assessment data is based directly on course material and consists of items measuring knowledge acquisition and knowledge comprehension. A pretest-posttest design with treatment and comparison (delayed treatment) groups provided maximum internal validity (e.g., selection effect, which is the most common threat to the casual effect of educational interventions), minimized competing explanations for the training effect, and reduced extraneous sources of variability. Purposive random assignment was employed in order to balance the groups as much as possible with respect to the variables of age, race, gender, education, geography, and experience.

Measures - Training satisfaction survey
Pretest/Posttest – knowledge and user motivation assessments

Research Questions - The study addressed four a-priori evaluation questions:

1. Is there a significant gain in knowledge for users?
2. Does the Phase-1 material appear to motivate users to coordinate and plan a real world event?
3. Is the pretest instructive independently of the training?
4. Are the materials well constructed and embraced by the user community?

Analyses – Analysis of Variance (ANOVA) was used to analyze all interval-level data from scaled questions. A within-subjects, one-way ANOVA was used for testing within-group knowledge-gain for both groups. To test between-group differences on dependent measures, data were analyzed using a one-way, between-subjects ANOVA.

User satisfaction was addressed in the Likert scale responses to satisfaction questions. These data were analyzed using descriptive analytic techniques (e.g., mean ratings, self-assessment of knowledge increase, etc.). Strength of

knowledge gain data, differential effectiveness findings, and qualitative information/feedback will be used to refine the intervention for Phase II.

In order to have confidence that the training increased knowledge, meaningful changes in objective measures must occur that cannot be explained by effects of time, a-priori between-group differences, or the influence of the pretest on the posttest that occurs independently from the training curriculum. The design employed allows the testing of each of these requirements.

The treatment and comparison groups were compared on traditional demographic variables to assure comparability of the groups prior to the study. Variables included participants age (collapsed into 10-year increments: 21-30 years, 31-40 years, etc); race (Black, White and American Indian); Gender; Education (high school, some college, college graduate, post college); rural versus urban program setting; any prior experience with Real World Simulation materials or programs. All analyses were performed using Chi Square analysis of categorical data (group assignment x demographic variable).

Results -There were no differences between the groups on any of the demographic variables. For example, the treatment and comparison groups were 31% and 27% male, respectively; and 62% and 55% urban, respectively. Similarly, the treatment and comparison groups were 38% and 35% Black, 59% and 62% White, and 3% and 3% American Indian, respectively. Although variables with multiple response categories (e.g., age, education) were not necessarily equally distributed along the ordinal categories (e.g., there tended to be more younger respondents than older respondents) there were no between-group differences. The results of the analyses are summarized in Table 2, below. The very low values of Chi Square and very high p-values reflect the equivalence of the groups, a-priori.

Table 2. Summary of between group demographic comparisons using Chi Square.

Variable	Chi-Square Value	Degrees of Freedom	p value
Age	1.725	4	.786
Race	0.06	2	.963
Gender	0.083	1	.773
Education	0.847	3	.838
Rural/Urban Program Setting	0.284	1	.594
Previous RW Simulation Experience	0.090	1	.764

To test for group differences on a-priori knowledge relating to the contents of the RWS curriculum, the pre-test scores of the two groups were compared using ANOVA. Recall that the two groups were pretested at the same time (Time 1). Differences on each individual knowledge question were tested, as was the Composite Index of Knowledge which is the sum of the answers to all questions. There were no differences observed on any question. The mean rankings of the two groups, differences between group means, and F-ratio tests of significance are presented in Table 3.

The between-group differences on each question range from 0.03 to 0.42, using a 7-point scale. The F-ratios and p-values clearly indicate that these differences are not only very small, but entirely random. The between-group difference on the composite index is also very small (1.11 out of a possible 55-point difference), random and insignificant. Furthermore, recall that the scale anchors were "strongly agree" to "strongly disagree" passing through a neutral

Table 3. Analysis of pre-test scores of knowledge test items for both groups.

Test Item	Tx Group Pretest Mean	Comp Group Pretest Mean	Between-Group Difference	F-ratio*	P-value**
Question 1	4.21	3.79	0.42	.883	.352
Question 2	4.62	4.83	-0.21	.152	.698
Question 3	3.41	3.38	0.03	.006	.937
Question 4	3.93	3.59	0.34	.716	.401
Question 5	3.36	3.48	-0.12	.098	.755
Question 6	3.69	3.28	0.41	.966	.330
Question 7	3.52	3.17	0.34	.428	.515
Question 8	5.72	5.76	-0.04	.006	.939
Composite Index	32.39	31.28	1.11	.378	.541

*In each case, df = 1/56 **Alpha set at p < .05

or "4" (conceptually "unsure," or "unable to decide"). Thus, group means hovering about the "4" rating indicate that each group was essentially naïve with respect to the contents of the curriculum prior to training. Only Question 8 had a mean score departing from "4," but the groups were essentially equivalent in their departures.

Having determined that the two groups were equivalent, and naïve, at the beginning of the study (Time-1) the main effect of the treatment (web-based instruction on conducting a Real World Simulation) was tested by comparing the

pretest and posttest scores of the treatment group. The posttest was administered to all participants in both groups at Time-2, which immediately following completion of the training by the Treatment Group. Results of the ANOVA on individual questions and the Composite Index of Knowledge are presented below in Table 4.

Table 4. Pretest (Time-1) and Posttest (Time-2) scores for the Treatment Group.

Test Item	Treatment Group Time-1 Mean	Treatment Group Time-2 Mean	Between-Time Difference	F-ratio*	P-value**
Question 1	4.21	5.31	-1.1	4.87	<.05
Question 2	4.62	6.17	-1.55	12.15	.001
Question 3	3.41	5.14	-1.73	17.69	<.001
Question 4	3.93	5.59	-2.20	19.36	<.001
Question 5	3.36	3.76	-0.40	1.02	.317
Question 6	3.69	4.45	-0.76	2.90	.094
Question 7	3.52	5.03	-1.51	8.98	<.01
Question 8	5.72	5.97	-0.25	.372	.57
Composite Index	32.39	41.41	-9.02	23.79	<.001

*In each case, df = 1/56 **Alpha set a p < .05

The analysis in Table 4 indicates that there was a 27.8% improvement in the Composite Index of Knowledge (mean difference = 9.02). Results of individual question analyses show increased knowledge scores (shown by negative numbers in the Between-Time Differences column) on all questions, with those differences being robust and significant for questions 1, 2, 3, 4 and 7. Question 6 approached significant (p = .09). Only questions 5 and 8 showed weak and insignificant gains, although they did contribute to the Composite Index of Knowledge.

In order to be sure that the differences observed for the main effect of training on the treatment group were not due to random effects of time or history, the Time-2 and Time-3 scores for the comparison group were analyzed in juxtaposition to the timeline for both groups. That is, the Time-1 Comparison Group scores were compared to the Time-2 Comparison Group scores (obtained at the same time as the Treatment Groups Time-1 and Time-2 scores, but without having received training), and the Time-2 Comparison Group scores were compared to the Time-3 Comparison Group scores (after that group had received training). The Time-3 testing was unique to the Comparison Group, and represents their knowledge scores following their exposure to the treatment. In

effect, the testing sequence for the Comparison Group at Times-1, -2 and -3 represent pretest-1, pretest-2, and posttest.

The first of these comparisons (Time-1 and Time-2, Comparison Group) tests for any instructive effect of the pretest, alone, on the posttest. It also tests for possible effects of time or history on the posttest. Thus, if there are Time-1/Time-2 differences for the Comparison Group, the main effect of training on the treatment group scores might not be a pure treatment effect, but might be influenced by intervening or random variables. If there are no significant time or random variable effects between Time 1 and Time 2 for the Comparison Group, the Time-2/Time-3 comparisons for the Comparison Group test for the main effect of training on the Comparison Group, at a time later than that tested for the Treatment Group. All of these comparisons were conducted using a one-way ANOVA across the three time conditions. The results are presented in Table 5, below.

The results of the ANOVA reveal large differences between the Time-3 group mean scores (those following exposure to the training curriculum after two iterations of the pretest) and the Time-1 and Time-2 means, and relatively small differences between the Time-1 and Time-2 means. The F-ratios and p-values show highly significant overall effects for all questions except Questions 5 and 8, and for the Composite Index of Knowledge. However, post-hoc Scheffe tests are necessary to determine which differences across time are significant.

Table 5 Time-1 & Time-2 Pretest scores and Time-3 Posttest scores for the Comparison Group.

Test Item	Comp Group Time-1 Mean	Comp Group Time-2 Mean	Comp Group Time-3 Mean	F-ratio*	P-value**
Question 1	3.79	4.41	6.14	18.42	<.001
Question 2	4.83	5.10	6.28	7.33	<.001
Question 3	3.38	3.24	4.83	8.17	<.001
Question 4	3.59	3.62	5.28	11.73	<.001
Question 5	3.48	3.41	4.31	2.94	.06
Question 6	3.28	3.03	4.31	4.97	<.01
Question 7	3.17	3.38	5.38	10.59	<.001
Question 8	5.76	5.86	6.38	1.44	.24
Composite Index	31.28	32.07	42.90	23.86	<.001

*In each case, df = 2/84 **Alpha set at p < .05

The results of the post-hoc tests are presented in Table 6, below. To conserve space, the Time-1/Time-2 comparisons and the Time -2/Time-3 comparisons are presented, as these are the comparisons of most interest for determining the independence of the main effect of training. To demonstrate maximum independence of the main effect of training on both groups, and to minimize competing explanations, Time-1/Time-2 differences for the Comparison Group should be small and insignificant, and Time 2-Time-3 differences should be robust and significant.

The mean score differences between Time-1 and Time-2 in Table 5 reveal that virtually no differences occurred in pretest scores whether due to time, history, or any instructive value of the pretest, itself. This supports strongly the independence of the main effect of training for the Treatment Group. The two right-most columns in Table 6 present the effects of the training curriculum on the Comparison Group, following their second exposure to the pretest. Their knowledge increase pattern on the individual posttest questions is very similar to that of the Treatment Group, including the non-significance of the differences recorded for Questions 5 and 8.

Table 6. Results of Scheffe post-hoc analyses of Time-1/Time-2 (Pretest 1 and Pretest 2) and Time-2/Time-3 (Pretest 2 and Posttest) scores for the Comparison Group.

Test Item	Comp Group Time-1/Time-2 Differences	p-value of F-test Time-1/Time-2	Comp Group Time-2 /Time-3 Differences	p-value of F-test Time-2/Tiume-3
Question 1	-0.62	.306	-1.72	<.001
Question 2	-0.28	.791	-1.17	<.05
Question 3	0.14	.951	-1.59	<.01
Question 4	-0.03	.996	-1.66	<.001
Question 5	0.07	.986	-0.98	.10
Question 6	0.24	.854	-1.28	<.05
Question 7	-0.21	.927	-2.0	<.001
Question 8	-0.10	.966	-.517	.423
Composite Index	-0.79	.915	-10.83	<.001

To see if there were any differences in the way the Comparison Group respond to the training in comparison to the Treatment Group, the Treatment Group Posttest scores obtained at Time-2 were compared to the Comparison Group

Posttest scores obtained at Time-3, and no differences were found either on an individual question basis or with respect to the Composite Index of Knowledge. The Comparison Group Composite Index of Knowledge increase by 33.8 % between Time-2 and Time-3 ($p < .001$) which is very similar to and not significantly different from the Treatment Group increase of 27.8%. Thus, the training curriculum was essentially equally effective for increasing the knowledge of both the Comparison Group and the Treatment Group.

It is not enough to know that the training curriculum is effective, although the preceding analyses clearly demonstrate that it is. Developers also need to know if the participants found the website easy to navigate, found the training vignettes believable, and (independently of our knowledge test questions) did the users self-assess as having gained in knowledge. These qualitative questions were asked using 7-point Likert rating scales with anchors tailored to each question's content. Because the two groups were essentially equivalent after training, their qualitative data are combined in the following presentations.

When asked if the respondents found the curriculum to be informative, the mean group response rating was 6.24 (N=58), with 1 = Not at All and 7 = Very Informative.

When asked how knowledgeable the respondents considered themselves to be with regard to conducting a Real World Simulation prior to viewing the curriculum, the mean group response rating was 2.44 (N=58) with 1 = Not At Al Knowledgeable and 7 = Very Knowledgeable, representing a fairly low self-assessment of knowledge prior to training. After completing the curriculum, the mean group rating was 5.12 (N=58, same anchors), which represents a 38.3% increase relative to the 7-point scale. This increase suggests a substantial increase in self-assessed knowledge following training.

The last 5 qualitative questions inquired about the features and "feel" of the web-based presentation of the curriculum. All 5 questions used anchors of 1 = Not at All to 7 = Very; thus high ratings are desirable. The questions and mean group ratings are presented below; in each case, N = 58.

- Did you find the Real World Simulation website to be visually appealing? Mean group rating = 5.95.
- Did you find the scripted scenes to be realistic? Mean group rating = 5.85.
- Did you find the acted scenes to be convincing? Mean group rating = 5.69.
- Did you find the scenes to be informative? Mean group rating = 6.11.
- Did you find the website to be easy to navigate during the training? Mean group rating = 6.52.

Overall these are very positive responses to the qualitative inquiries. Combined with the results of the analysis of the knowledge test scores and determination of the efficacy of the curriculum as a pure treatment effect, the feasibility of this Phase-1 demonstration is strongly supported.

In addition to the objective data, subjects were provided the opportunity to respond to an open-ended question, and 33 of 58 did so. The large majority of comments were laudatory, suggesting that respondents had a very positive experience testing the website and being exposed to the introductory training materials. A sample of these comments follows:

- I thought it was great
- I did not know this kind of program existed. I think it is a great thing for all youths.
- I found this website visually attractive and easy to navigate. I am very interested in learning more about the material presented.
- This type of event is greatly needed in our communities. My husband and I teach a life skills program but it would make greater sense to use experts, within the community, to teach their specialty and have everything together during one day to see how it all relates instead of piece mealing each topic.
- I really found the information to be very useful and REAL.
- I feel that this website would be very useful in planning a Real World event.
- The website is very easy to use. It might be nice to have links to areas in the country where these events are taking place for those who would like to volunteer but not run the entire event.
- I feel using the real life people to share their own personal experiences was very beneficial.
- I really liked the design of the website. The colors are soft and appealing; it was easy to navigate, and it wasn't 'cluttered'. The videos were very professional and entertaining as well as informative. I liked that the program gave explanations of each question, too.

A few of the comments offered constructive criticisms of various features of the website. For example:

- I think there should be a little more detail about how the Real World Program works in the first module.
- Disconcerting that one answer was deemed 'incorrect' because I spelled 'toolkit' as two words rather than one.
- I did not like typing in the answer without options to choose from--my mind went blank but if I had a list of words to choose from I would have gotten the correct answer.

Taken as a whole, the comments offered by participants were very supportive, and many of the participants reported being inspired to pursue hosting a Real World Simulation event for youths with whom they work. They expressed eagerness to view additional curriculum segments. Both the complementary and critical comments will inform future design and development efforts.