Adolescent Real World Simulation (ARWS) Final Report

ARWS Phase II Research Evaluation Report

The feasibility of the Real World Simulation model was strongly supported by the findings of the Phase I evaluation. Phase II expanded the curriculum, produced the training materials, and conducted an evaluation of the final product using essentially the same design that was used in Phase I.

Summary – The success of the Phase II project is manifestly evident by the evaluation findings that included:

- 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
- the 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 II focused on the users' perceptions of the expanded curriculum and whether the curriculum was effective in imparting new information to training participants. Accordingly, the Phase II evaluation plan focused on user satisfaction and knowledge gain.

- 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 post-tested 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 2 training program, but the post-testing 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 evaluation conditions for Treatment and Comparison groups.

Time period →	Month 1, Day 1	Month 1, Day 21	Month 2 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 curriculum and training experience. Knowledge assessment data was based directly on course material and consisted 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 demographic 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 curriculum appear to motivate users to coordinate and plan a real world event?
3. Is the pretest instructive independent of the training?
4. Are the materials well-constructed and embraced by the user community?

Participants – A total of 59 participants contributed data to the evaluation: 29 in the Treatment Group and 30 in the Comparison/Delayed Treatment Group.

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.).

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; race (Black, White and American Indian); Ethnicity (Hispanic/Latino, non-Hispanic/Latino), Gender (male, female); Education (high school, some college, college graduate, post college); rural versus urban program, type of program setting (e.g., public or private child welfare, school, juvenile services),and whether the participant had ever received prior training on youth transitioning to independent living.

Being a ratio-level measure, age was analyzed using a group means test (ANOVA). All other comparisons on demographic variables 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. Although the mean age of the treatment group was slightly lower than the comparison group (41 years and 45 years, respectively), the difference was not significant (F=3.304, df = 1, 57, p = .074). There were not even trends among the categorical measures. For example, the treatment and comparison groups were 17% and 20% male, respectively; and 31% and 37% rural, respectively. Similarly, the treatment and comparison groups were 48% and 53% Black, 48% and 43% White, and 1% and 1% American Indian, respectively. 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**
Race	0.153	2	.926
Ethnicity	0.01	1	.972
Gender	0.074	1	.786
Education	1.054	2	..590
Type of Program Setting	1.868	5	.867
Rural/Urban Program Setting	1.144	2	..564
Participated in Youth Transition Training	0.203	1	.652

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 in absolute value from 0.16 to 0.57, 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 (1.69 out of a possible 60-point difference), random and insignificant. Furthermore, recall that the scale anchors were "strongly agree" to "strongly disagree" passing through a neutral rating 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.

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	3.79	4.03	-0.24	.231	.632
Question 2	3.90	4.47	-0.57	1.252	.268

Question 3	3.34	3.60	-0.26	.397	.531
Question 4	3.31	2.97	0.34	..757	.388
Question 5	3.00	3.50	-0.50	1.528	.221
Question 6	3.38	3.57	-0.19	..122	.728
Question 7	4.14	4.30	-0.16	..129	.721
Question 8	4.55	4.17	0.38	..558	.458
Question 9	4.14	4.37	-0.23	.249	.620
Question 10	4.86	5.13	-0.27	.458	.501
Composite Index	38.41	40.10	-1.69	.903	.364

*In each case, df = 1/57 **Alpha set at p < .05

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

The analysis in Table 4 indicates that there was a 32.9% improvement in the Composite Index of Knowledge (mean difference = 12.65). Results of individual question analyses show increased knowledge scores (shown by negative numbers in the Between-Time Differences column) on 8 of 10 questions. Only questions 8 and 9 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.

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	3.79	6.24	-2.45	40.54	<.001
Question 2	3.90	5.52	--1.62	11.99	<.001

Question 3	3.34	4.31	-0.97	5.38	<.05
Question 4	3.31	5.31	-2.00	27.90	<.001
Question 5	3.00	4.45	-1.45	11.27	<.001
Question 6	3.38	5.38	-2.00	19.37	<.001
Question 7	4.14	5.72	-1.58	20.18	<.001
Question 8	4.55	3.83	0.72	1.81	.184
Question 9	4.14	4.03	0.11	0.36	.851
Question 10	4.86	6.21	-1.35	14.09	<.001
Composite Index	38.41	51.00	-12.59	57.501	<.001

*In each case, df = 1/59 **Alpha set a p < .05

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.

The results of the post-hoc Scheffe 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 6 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 8 and 9. For this group, Question 3 was also insignificant.

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	4.03	3.77	5.83	11.43	<.001
Question 2	4.47	4.30	5.93	7.71	<.01
Question 3	3.60	3.67	4.00	.555	.576
Question 4	2.97	3.4	5.6	23.94	<.001
Question 5	3.50	3.2	4.47	5.50	.01
Question 6	3.57	3.63	5.47	8.78	<.001
Question 7	4.30	4.43	5.67	5.25	<.01
Question 8	4.17	3.90	3.50	.798	.454
Question 9	4.37	4.80	5.07	1.03	.363
Question 10	5.13	5.30	6.30	7.42	<.01
Composite Index	40.11	40.40	51.84	24.54	<.001

*In each case, df = 2, 87 **Alpha set at p < .05

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/Time-3
Question 1	.267	.852	-2.067	<.001
Question 2	.167	.936	-1.633	<.01
Question 3	-.067	.987	-.333	.716
Question 4	-.433	.571	-2.200	<.001
Question 5	.300	.754	-1.267	<.01
Question 6	-.067	.992	-1.833	<.01
Question 7	-.133	.960	-1.233	<.05
Question 8	.267	.882	.400	.754

Question 9	-.433	.681	-.267	.864
Question 10	-.167	.879	-1.000	<.05
Composite Index	-.300	.988	-11.433	<.001

*In each case, df = 1, 58 **Alpha set at p < .05

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 28.3 % between Time-2 and Time-3 (p < .001) which is very similar to and not significantly different from the Treatment Group increase of 32.9%. Thus, the training curriculum was essentially equally effective for increasing the knowledge of both the Treatment Group and the Comparison/Delayed-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 6.71 (N=59), 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 2.75 (N=59) 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 5.25 (N=59, same anchors), which represents a 91%% 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 = 59.

- Did you find the Real World Sim website to be visually appealing? Mean group rating = 6.14.
- Did you find the scripted scenes to be realistic? Mean group rating = 6.07.
- Did you find the acted scenes to be convincing? Mean group rating = 6.14.
- Did you find the scenes to be informative? Mean group rating = 6.36.
- Did you find the website to be easy to navigate during the training? Mean group rating = 6.29.

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 treatment effect, it is clear that the curriculum is not only effective, but the mechanisms for presentation were very well received, and the web interface was very easy to use by members of the intended audience.

In addition to the objective data, subjects were provided the opportunity to respond to an open-ended question, and 53 of 58 did so. The large majority of comments were very complimentary to the product and the experience. May comments were brief: “Great job;” “I loved the experience;” “This should be required in every high school in the country;” etc. Others expressed appreciation for being invited to participate in the test of the curriculum, and others indicated being motivated to carry out the curriculum and convene a Real World Simulation event.

There were a few expressions of frustration regarding streaming of videos, but these comments were associated with participants using Windows Explorer and older operating systems (Pre Windows 7). People using Google Chrome or Firefox did not report any streaming problems. Overall, these problems were infrequent, and not disabling, merely annoying.

Overall, the evaluation findings are uniformly positive, and robust. Users enjoyed the experience, reported learning a great deal, and also “tested” as having learned a great deal. Satisfaction measures were very high and consistent across both groups. The knowledge gain can be attributed solely to the curriculum and training experience due to the pre-trial equivalence of the Treatment and Comparison Groups on all demographic variable, absence of any time effects (controlled for in the design and statistically verified), and absence of any pre-test instructional effects (also controlled for in the design and statistically verified).).

Preparing for Phase III

Since completing the ARWS research, developers have upgraded the curriculum in preparation for a public launch. The Content Management System (CMS) has been upgraded and features have been added to accommodate research participant suggestions. The developers have evaluated the specific written feedback to generate the needed adjustments and create additions to the curriculum before promoting the site in Phase III.

Some additions to the ARWS curriculum include:

1. The ability to swipe the Interactive Timeline “clean” with one click to begin planning another Real World event.
2. A grant template that can be used to help communities request funding to support a community-wide Real World Simulation.
3. A curriculum summary allowing those users who have finished the curriculum to locate specific sections to relearn or reference information.
4. The addition of “words of encouragement” to accompany Interactive Timeline emails. This encouragement is meant to keep planners motivated during the planning process.
5. The ability for community business partners to see the video’s introducing them to the Real World concept without having to log into the website.

ILR has continued promoting the Real World program and testing the online features. During the past six months ILR used the online curriculum as a substitute for traditional “stand-up” training and community mentoring. ILR monitored the process to identify glitches in the system or gaps in learning. The training participants had no previous knowledge of the Real World program and successfully coordinated a community-wide event held on June 26, 2014. Over 160 people attended the event making a huge impact on the business and social service community. The collaboration of three agencies motivated their work and facilitated lasting professional partnerships.

Enthusiasm has been building among previous research participants, interested professionals and educators waiting for the curriculum to be available. ILR has been connecting with media outlets and using social media to help broadcast the effectiveness of the program and stimulate interest in advance of the school year beginning. An international publication wrote extensively on the benefits of the program and project partners have continued their interest in seeing the ARWS product become available commercially. It seems once someone learns about this product they share it with another. ILR anticipates a successful commercialization of the ARWS curriculum.