Construction Certification Test Can Now Be VR

ANSI has accredited ITI's virtual reality test for construction hazard education.

February of this year marked the first time that ANSI accredited a certification whose test was done through virtual reality.

That certification test was Industrial Training International's (ITI's) Construction Hazards Identification (CHID) exam.

At its core, the CHID exam is a VR shooter game. Users are turned loose on virtual construction sites that have various hazards. Players earn points for each hazard they see and mark, and they need to earn a specified score, or more, to pass.

The types of hazards on the test come from OSHA's list of the top 10 hazards it sees most often during site inspections.

As ITI and Authentic Testing Corp. developed the Construction Hazards Recognition test, they addressed many issues in order to assure the test is effective, accurate, and fair.

It had to be all three for ANSI to accredit it for use in certifying workers.

Here's how ITI and Authentic Testing solved each issue to perfect the VR test.

Cheating

Problem: The virtual hazards are memorable and workers talk to each other both before and after the test.

Solution: Vary the hazards. We created more than 120 hazards that could be shown in the 20+ item test. We just selected different subsets of the hazards for each testing session.

Loading Time

Problem: Loading took too long. Each hazard took five seconds to load if we chose them randomly during the test.

Solution: Preassemble different playlists of hazards. Using a playlist, there was no discernible loading time as the candidate moved around a virtual site or moved from one site to another.

Static Playlist

Problem: With a static playlist, the problem of cheating remained. If we had just one playlist, it would soon be discussed by all the workers on a site.

Solution: Put the hazards into multiple preassembled playlists, then select one playlist for any given testing session. With 20 or more playlists, there is little likelihood that candidates will encounter identical, or even similar, sets of hazards.

Unfair Difficulty

Problem: Some playlists could be more difficult, making the testing unfair.

Solution: We made sure the playlists were all equally difficult.

Selecting Equivalent Playlists

Problem: More hazards were available than were put on one playlist. How could we select those for each one?

Solution: We rated each hazard by its difficulty and risk, then equalized the overall difficulty of all the playlists. Difficulty is how hard the hazard is to recognize. Risk is the scope of the problem if the hazard isn't recognized.

Each hazard was assigned points based on the difficulty of seeing it and the consequences if it were not identified.

Quantifying Difficulty

Problem: How do you quantify difficulty?

Solution: The difficulty classifications are shown below. Visibility of Hazard = Difficulty.

Rating	Visibility Examples
1	**Immediately Visible** *Ex: No gloves, No hard hat*
2	**Add'l Equipment Needed** *Ex: No ditch frame*
3	**Regulation Infraction** *Ex: Oxygen near flame*
3	**Measurement Needed** *Ex: Unloading crane by hand*
4	**Situational Recognition** *Ex: Bad lanyard attachment point, Ladder on slick floor*
5	**Pair Needed for Inference** *Ex: Scaffolding - up, below*

The table above shows how we constructed initial estimates of hazard difficulty.

Once sufficient data is available to get results from more than 50 candidates for each item, the actual probability of success in identifying that hazard will be computed, rounded, and substituted into the rating scale.

Quantifying Risk

Problem: How do you quantify risk?

Solution: Quantify the danger to individual, team, and work site.

Although none of us was an actuary, we evaluated risk as shown below.

Factor is the factor by which the difficulty is multiplied if the candidate recognizes the hazard.

Level is the verbal rating of risk.

Consequence is the consequence if the hazard is not mitigated.

Wallace Judd, Ph.D., is president of Authentic Testing Corp., Leesburg, Virginia. He can be reached at wjudd@authentictesting.com or 703.777.7321.

Example illustrates a hazard of the specified level.

The table at right clarified the differences in risk between hazards, so it became more clear to the authors and developers specifically how to assess risk for each hazard.

Integrating Risk and Difficulty

Problem: How do you integrate risk and difficulty?

Solution: Define Hazard Points as risk *times* difficulty.

We could have added difficulty and risk, but we felt multiplying the two would create a greater spread between hazards. Also, we had no reason to assume that one factor should be weighted more than the other, so we didn't create weighting factors for risk and difficulty.

Equivalent Playlist Content

Problem: Playlists don't meet blueprint specifications.

Solution: Create stratified random template hazard domains. See table 3 - *Hazard Frequencies* at right.

Equal Playlist Item Totals

Problem: How can all playlists have identical numbers of items?

Solution: Randomly select from pairs that would equal a constant number.

The playlist content had to relate to the 10 OSHA citation categories. However, the OSHA categories were not evaluated as to frequency of occurrence.

At the same time, we didn't feel that all categories would be encountered with equal likelihood on the job. Consequently, we created a selection table that reflected our opinion of the likelihood of encountering a hazard.

By randomly selecting one pair from each of the item groups, then using the other pair, we could randomize the number of items in each domain pair and still assure that there were 23 items in each playlist.

See *Table 3 - Hazard Frequencies,* above right, to see the domain pairs that had equivalent sums.

Creating Equivalent Playlists

Problem: More hazards available than put on playlist. How to select?

Solution: Score hazards by difficulty, risk.

Here's how we created playlists that all have equivalent difficulty and risk:

Generate 1,000 random playlists that fit the blueprint.

Sum the total points in each template.

Find the mean and standard deviation of

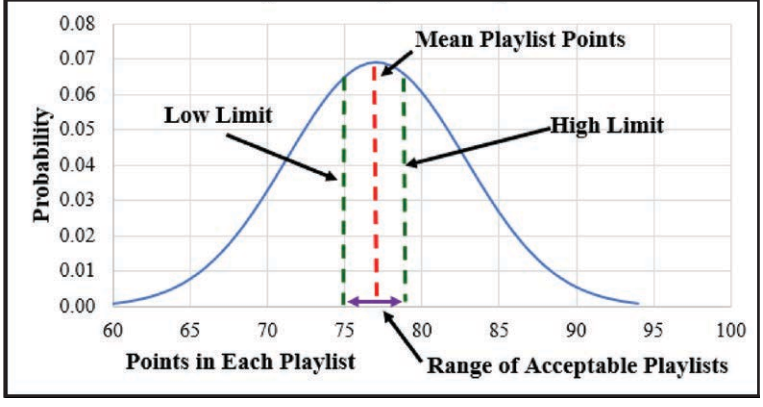
Factor Level Consequence Example - Risk of Non-Recognition

Factor	Level	Consequence	Example
1	Low	Inconvenience Trip 10 min. fix	Extension cord No marking Gas can alone
3	Mid	Fall Broken bones 1 hr. fix	Gas can by wood
5	High	Fatalities Multiple injuries Shut down project	Gas can - sparks No ditch frame Front loader backup

Table 3. Hazard Frequencies

CHID Blueprint	
Domain	**Items**
Confined Space and Hot Work	1–2
Environmental Hazard	1–2
Lifting and Rigging	1–2
Industrial Hygiene	1–2
Electrical Safety and LOTO	2–3
MEPI and Excavation/Trenching	2–3
Dropped Object Prevention/Protection	2–3
Scaffolding	2–3
Hand Tools	3–4
Fall Prevention	3–4

Figure 1. Range of Acceptable Playlists



Editor's note: Industrial Training International's virtual reality video test that is used to certify construction workers as having been educated to recognize construction hazards has been accredited by ANSI. It's the first VR test that ANSI has accredited. That means OSHA accepts it. This test is for general construction hazards, but it opens the door for the possibility that crane operator and rigger certification tests could someday also be done virtually.

template points.

Set selection bounds of mean ± 0.3 Std. deviations.

Generate playlists and calculate points.

Use only playlists whose total points are within bounds.

Scoring and Penalties

Problem: The optimal strategy would be to mark everything. That way a candidate could be sure to get all the hazards.

Solution: Penalty points for marking non hazards.

There are two types of penalties. The first is for failing to recognize a hazard. For doing this, a candidate simply fails to accrue the number of points the hazard is worth.

The second type of penalty is for pointing out an object, person, or location that is not a hazard. For the first two errors of this type in each area, no points were subtracted.

After that, 4 points were subtracted for each of the next two errors, 6 points subtracted for each of the next two errors, and so on.

The rationale for this was forgiveness — for 2 hazards. After that, the points rate made up for the forgiveness on the first 2 hazards. See *Table 4*, below.

Total Points

Problem: How to integrate penalties and score.

Solution: Score equals hazard ID points minus penalty points.

Total Points equals the points for each hazard identified, minus the penalty points calculated in *Table 4. Penalty Table.*

Table 4. Penalty Table

Incorrect Markers	Penalty
0	0
1	0
2	0
3	4
4	4
5	6
6	6
7	8
8	8
9	10
10	10

Actual Score

Problem: Cutscore (pass/fail score) can't be total points.

Solution: The cutscore is a percentage of total points in the playlist.

Because different playlists contain different total available points, the cutscore can't be based on total points a candidate earned.

The cutscore has to be the percentage of total available points in the playlist that the candidate achieved.

So the total score equals the points for each hazard identified correctly, minus the penalty points calculated in *Table 4* (the penalty table), divided by the total number of hazard points available in the playlist.

Candidate score equals (points for identified hazards minus penalty points) divided by total points in the playlist.

Time Limit

Problem: Candidates taking too long to complete the test.

Solution: Time limit = Beta test time + 2 Standard Deviations = 97%.

No time limits for the total test were enforced on the Beta version so that unlimited time for completion could be estimated.

The section time limits below shows the test time distributions for candidates, as well as the recommended time limit for the test when it is administered in the field.

Each of the six areas had a time estimate that was not enforced, but which was recommended by displaying blue "ghost" footprints in front of the candidate showing him how to find the area exit.

The table to the right shows a summary of the times for all candidates taking the Beta test. The Finished column shows data for those who finished all six areas of the test.

CHID Beta Test Times

Parm	Finished
N	46
Mean	16.46
Std.Dev.	6.09
Min.	11.78
Max.	33.27
μ+σ	22.55

The recommended time limit for administration of the exam, exclusive of the tutorial, was 23 minutes.

That would allow 84% of candidates to complete the exam without being hurried.

It is anticipated that as candidates get better training and preparation, overall times for completion will be reduced and fewer than 16% of candidates would time out.

Low Reliability

Problem: Low Alpha reliability.

Solution: Teach marker penalties.

When we calculated test statistics, the overall reliability was low. We began to wonder how the instability of scores could have occurred.

When we looked at the penalty points, we saw that some candidates received as many as 70 penalty points. Their scores were just barely positive.

We realized that these candidates had not been taught that marking objects that were not hazards would penalize them and lower their scores.

Looking at the results, those penalty points had made the total scores unreliable.

The solution was to give introductory instructions that explained how to erase markers over objects that were not hazards, and that also explained about the penalty points for marking non hazards.

Summary

To create a VR test that meets EEOC guidelines for employment, promotion, or retention, most of these issues will be relevant.

At Authentic Testing, we see VR as the next frontier in testing and certification.

Companies like ITI have taken up the challenge, and we hope that in the near future many more will do likewise.

References

This ANAB blog explains the requirements for 17024 Accreditation specific to VR: https://blog.ansi.org/anab/virtual-reality-assessment-iso-iec-17024/ ■

ANAB
ANSI National Accreditation Board
ACCREDITED
ISO/IEC 17024
PERSONNEL CERTIFICATION BODY