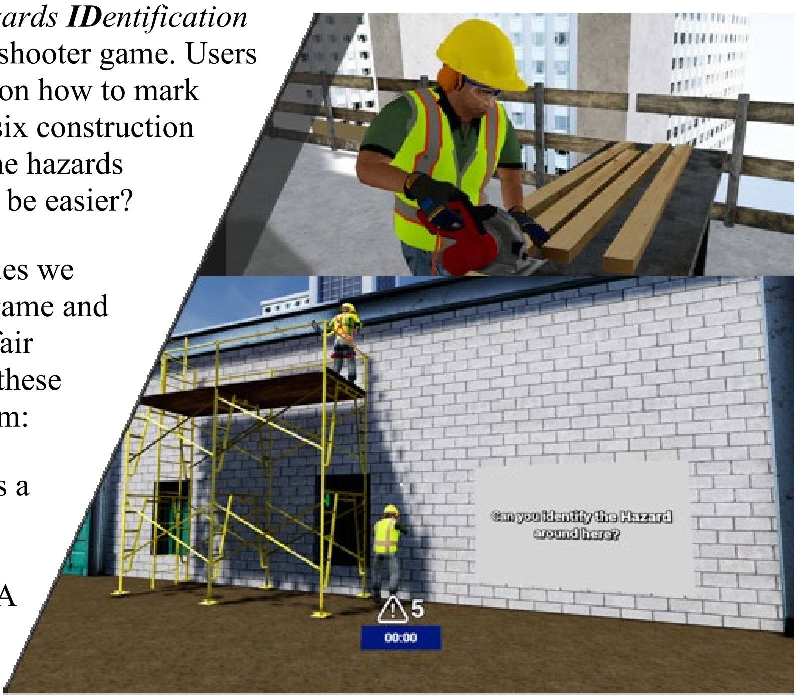
Authentic Testing®
The Gold Standard for Testing

AuthenticTesting.com
673 Potomac Station Rd NE No. 810
Leesburg, VA 20176
(703) 777-7321

Certification in the Metaverse

By Wallace Judd, Ph.D., President
Authentic Testing Corp.
WJudd@AuthenticTesting.com

In February 2022 ANSI (American National Standards Institute) awarded the very first accreditation to a VR certification, the Construction Hazards Identification exam by ITI (Industrial Training International).



In the beginning, ***C**onstruction **H**azards **ID**entification* (CHID) was just a VR first-person shooter game. Users got instruction in teleportation and on how to mark hazards, and were turned loose on six construction sites with various hazards. Count the hazards marked for a score. . . . What could be easier?

As we'll see, there were lots of issues we encountered as we beta-tested the game and turned it into a reliable, valid, and fair certification. So we had to remedy these issues to develop an accredited exam:

First a little background. OSHA has a requirement that construction employers train employees about construction hazards. Indeed, OSHA has a list of the top 10 hazard types cited in site inspections. These top 10 hazard classifications became the foundation for all the hazards shown in the exam.

I list the issues as we encountered them.

Cheating

1. Problem: Cheating—people talk. Memorable hazards
Solution: Variable hazards. We have plenty

As one might expect, the first issue we anticipated was cheating. Many of the construction sites had trailers in which classes on hazards were taught. And of course, workers both before and after the test would be talking with each other.

And VR testing presents one problem that multiple-choice tests don't have: they're easy to remember. The hazards are visible and easily communicated to those who haven't yet taken the test.

The way we solved this was to create over 120 hazards that could be shown in the course of a 20+ item test. We would just select a subset of the hazards and show them.

Load Time

2. Problem: Load time too long. Required 5 seconds to load hazard
Solution: Define a ***Playlist*** with hazards on it

The problem with this was that the load time would be too long if we randomly selected the hazards and assembled them in real time. So we could create a list of preassembled hazards we called a *playlist.* With a playlist there was no discernable load time as the candidate moved around the construction sites or moved from one site to another.

Static Playlist

3. Problem: Playlist is static
Solution: Multiple variable playlists

The problem of cheating still remained with a playlist. If we had just one playlist, it would soon be discussed by all the workers on a site.

The solution was to put the items in multiple playlists that were preassembled, then select a preassembled playlist to present to candidates. With 20 or more playlists, there is little likelihood that candidates would encounter identical or even similar sets of hazards.

Multiple Playlists

4. Problem: Variable playlists unfair—variable difficulty
Solution: Make playlists vary from candidate to candidate

The problem with multiple playlists is that some of the playlists would be more difficult than others. We had to make sure the playlists were of equal difficulty to be fair to all candidates.

Selecting Equivalent Playlists

5. Problem: More hazards available than put on playlist. How to select?
Solution: Score hazards by Difficulty, Risk

Our solution theoretically was to rate hazards by difficulty and risk. Then we could equalize. Difficulty is how difficult it would be to recognize a hazard. Risk is the scope of the problem if the hazard were not recognized. We didn't concern ourselves with reporting or resolving the hazard, since that process would differ at variable construction sites.

The problem remained as to how could we "objectively" quantify difficulty and risk.

Objectify Difficulty

6. Problem: How objectify Difficulty?
Quantify: Visibility: on the face, collectively, invisible

Items points were awarded based on the difficulty of seeing the hazard and the risk—essentially the consequences if the hazard were not identified. The difficulty classifications are shown below.

Visibility of Hazard = Difficulty		
Rating	**Visibility**	**Example**
1	Immediately Visible	No gloves No hard hat
2	Additional Equipment needed	No ditch frame
3	Regulation infraction	Oxygen near flame
	Measurement needed	Unloading crane by hand
4	Situational recognition	Bad lanyard attach point Ladder on slick floor
5	Pair needed for inference	Scaffolding - up, below

Table 1. Item Difficulty Categories

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.

Item Risk

7. Problem: How objectify risk?
Quantify: Danger to individual; team; worksite

While none on our team was an actuarial, we evaluated risk as well, as shown in the table below. In the table below:

Factor is the factor the difficulty is multiplied by if the candidate recognizes the hazard.
Level is the verbal rating of risk.
Consequence is the consequence if the hazard is not mitigated.
Example illustrates a hazard of the specified level.

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

Table 2. Item Risk Categories

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

Integrating Risk & Difficulty

8. Problem: Integrate difficulty with risk
Solution: Define hazard Points = risk * difficulty

We calculated the point value of a hazard as:

Points = Difficulty * Risk

We could have added difficulty and risk, but felt that their product 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

9. Problem: Playlists don’t meet blueprint specifications.
Solution: Create stratified random template hazard domains

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

Table 3. Hazard Frequencies

Equal Playlist Item Totals

10. 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 ten 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 which 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 23 items occurred in each playlist.

See Table 3. Hazard Frequencies to see the domain pairs that had equivalent sums.

Equivalent Playlists

11. Problem: More hazards available than put on playlist. How to select?
Solution: Score hazards by Difficulty, Risk

We can now return to the question of how to create equivalent playlists—ones of equivalent difficulty and risk. The solution was as follows:

Generate 1,000 random playlists that fit the blueprint
Sum total points in each template
Find mean and standard deviation of template points
Set selection bounds of mean ± 0.3 Std. deviations
Generate playlists & calculate points
Only use playlists with total points within bounds

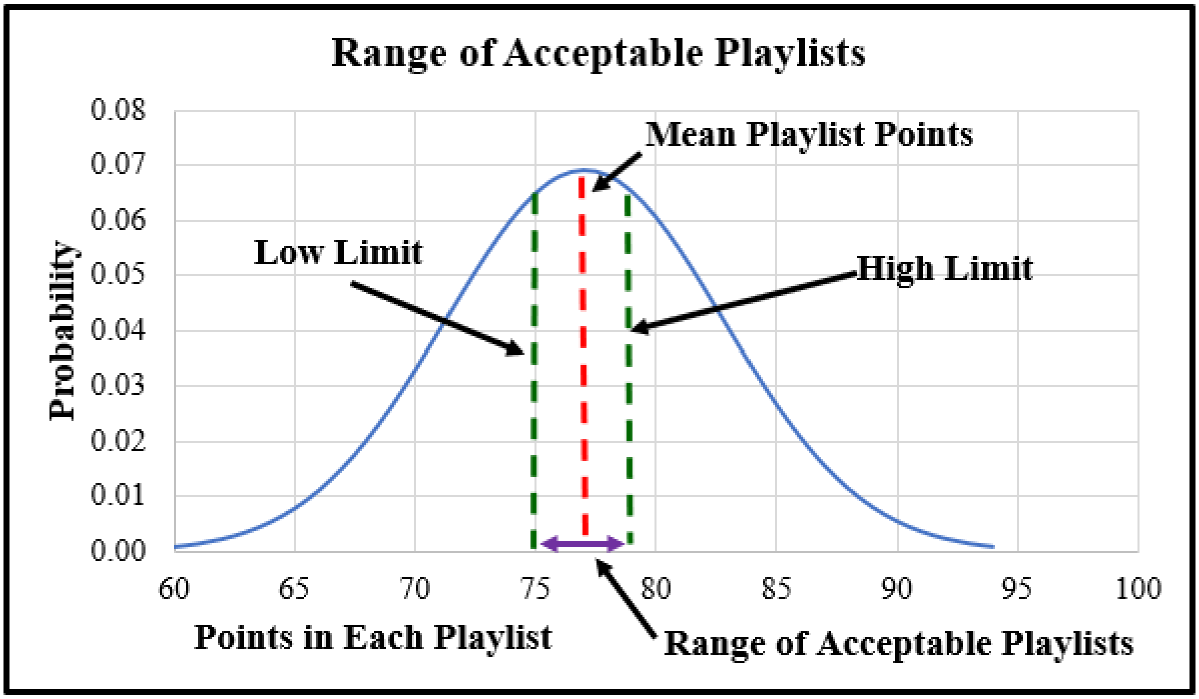


Figure 1 Range of Acceptable Playlists

Scoring – Penalties

12. Problem:	Optimal strategy is to mark everything
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 of this type of error 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.

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

Table 4. Penalty Table

Total Points

13. Problem:	Integrating penalties with score
Solution:	Score = hazard ID points – penalty points

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

Actual Score

14. Problem:	Cutscore can't be total points
Solution:	Cutscore is % of total points in playlist

Because different playlists contain different total available points, the pass/fail score (or cutscore) can't be based on total points a candidate achieved. 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 correctly identified, minus the penalty points calculated in the table above, divided by the total number of hazard points available in the playlist.

Candidate score = (Hazard points identified – Penalty points) / Total points in playlist

Time Limits

15. Problem: Candidates taking forever
Solution: Limit = Beta test time + 2 Std. Dev. = 97%

No time limits were enforced on the Beta for the total test so that unlimited time to completion could be estimated. The section Time Limits below shows the test time distributions for candidates and the recommended time limit for the test when it is administered in the field. Each of the six Areas had a time estimate which 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 below 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**
$\mu+\sigma$	**22.55**

Table 5. Candidate Test Times

The recommended time limit for administration of the exam, exclusive of the tutorial, was 23 minutes. This time limit 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 not even 16% of candidates would time out.

Low Reliability

16. 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. What we realized was 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, it was these penalty points which 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 which told of the penalty points awarded when non-hazards were marked.

<u>Summary</u>

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. We hope you will in the near future.

ITI received ANSI Accreditation under ISO/IEC 17024 in February 2022.

ANAB

ANSI National Accreditation Board

ACCREDITED

ISO/IEC 17024

PERSONNEL CERTIFICATION BODY

References

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