INSTITUTE FOR DEFENSE ANALYSES

DATAWorks 2022: Measuring Training Efficacy: Structural Validation of the Operational Assessment of Training Scale (OATS)

Dr. Vincent A. Lillard, Project Leader

Dr. Brian D. Vickers, Dr. Daniel J. Porter, Mrs. Rachel A. Haga,
Dr. Heather M. Wojton

March 2022

Public release approved. Distribution is unlimited.

IDA Document NS D-32972

Log: H 2022-000049

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task BD-09-229990, "Test Science Applications," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule and consisted of Elizabeth Green from the Operational Evaluation Division, and Emily Fedele from the Science and Technology Division.

For more information:
Dr. Vincent A. Lillard, Project Leader
vlillard@ida.org • (703) 845-2230

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-32972

DATAWorks 2022: Measuring Training Efficacy: Structural Validation of the Operational Assessment of Training Scale (OATS)

Dr. Vincent A. Lillard, Project Leader

Dr. Brian D. Vickers, Dr. Daniel J. Porter, Mrs. Rachel A. Haga,
Dr. Heather M. Wojton

Executive Summary

Effective training of the broad set of users/operators of systems has downstream impacts on usability, workload, and ultimate system performance that are related to mission success. In order to measure training effectiveness, we designed a survey called the Operational Assessment of Training Scale (OATS) in partnership with the Army Test and Evaluation Center (ATEC).

Two subscales were designed to assess the degrees to which training covered relevant content for real operations (Relevance subscale) and enabled self-rated ability to interact with systems effectively after training (Efficacy subscale). The list of 15 items was provided to over 700 users/operators across a range of military systems and test events (comprising both developmental and operational testing phases). Systems included vehicles, aircraft, C3 systems, and dismounted squad equipment, among other types.

We evaluated reliability of the factor structure across these military samples using confirmatory factor analysis. We confirmed that OATS exhibited a two-factor structure for training relevance and training efficacy. Additionally, a shortened, six-item measure of the OATS with three items per subscale continues to fit observed data well, allowing for quicker assessments of training. We discuss various ways that the OATS can be applied to one-off, multi-day, multi-event, and other types of training events.

Additional OATS details and information about other scales for test and evaluation are available at https://testscience.org/validated-scales-repository/.

Measuring Training Efficacy: Structural Validation of the Operational Assessment of Training Scale (OATS)

Brian Vickers, Ph.D.

April 2022

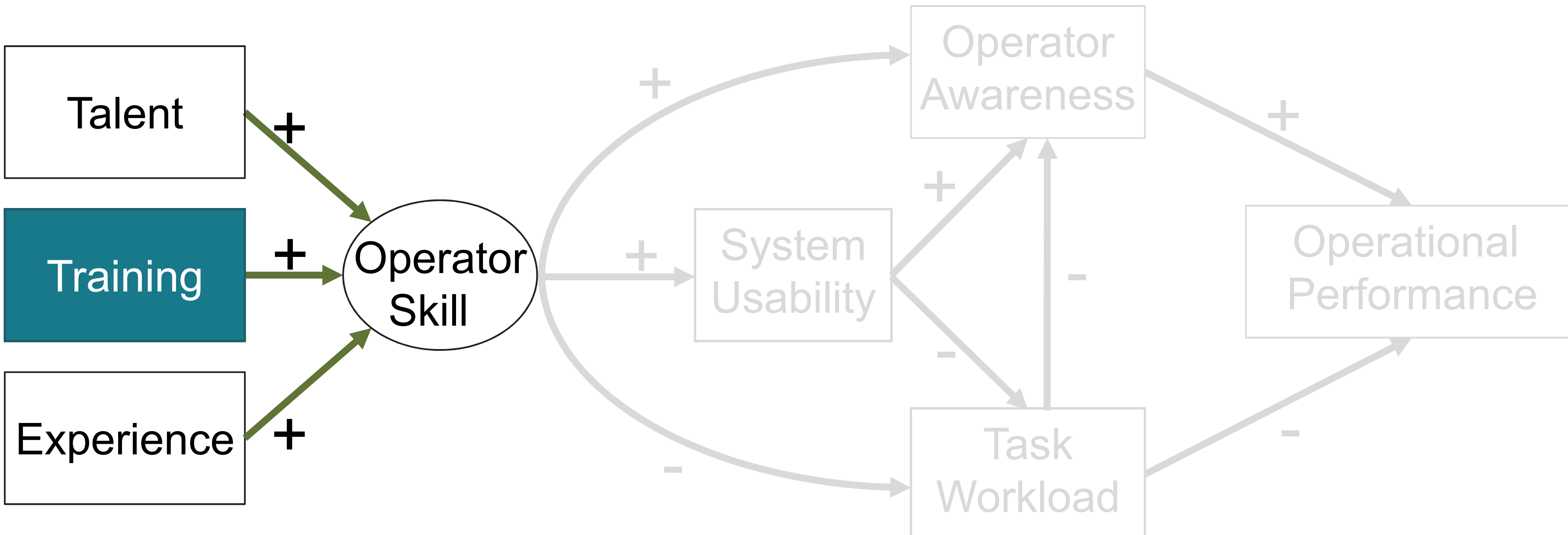
Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

Training is an important precursor for downstream human-system integration measures.

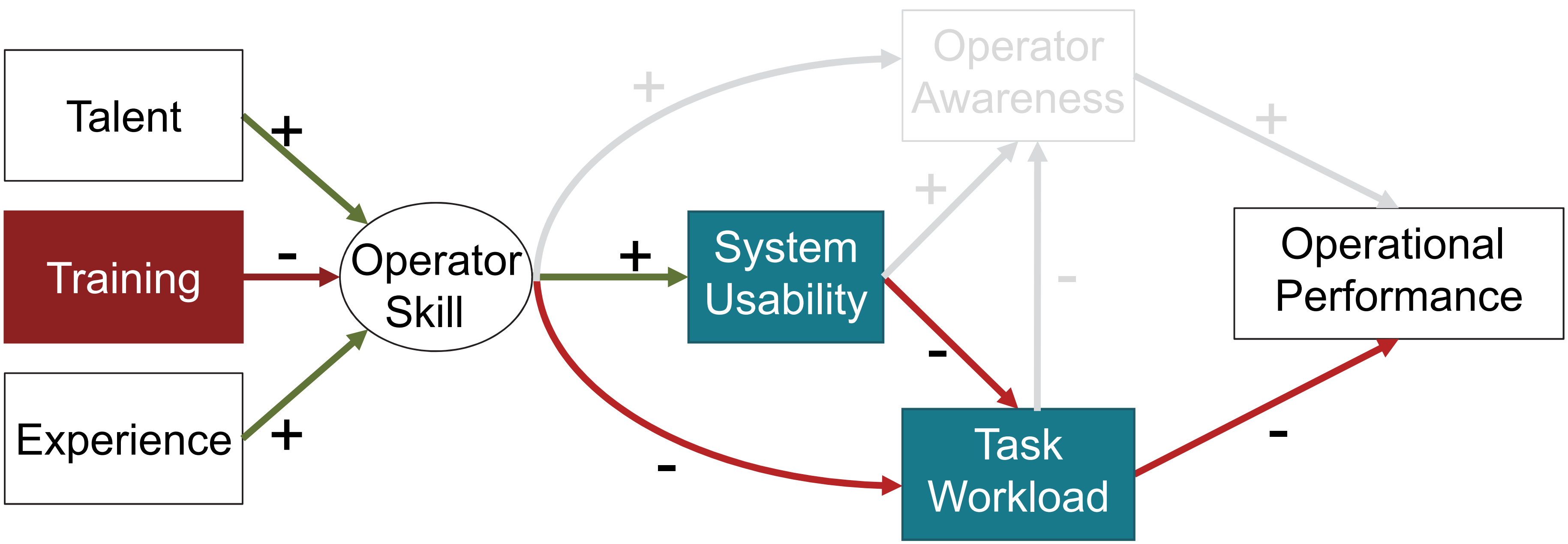
Training on DOD systems ranges from virtually nothing to custom-built learning management systems.

An easily administered, validated survey measure for the relevance and effectiveness of training will enable T&E of military systems.

Training is one of the earliest pieces of the human-system interaction (HSI) process



Training has downstream impacts on HSI processes, eventual system performance



When observing problems in usability, workload, or performance, it is important to consider the impact of training.

Training can increase job performance broadly

Training on tasks causally increases people's beliefs in their ability to perform tasks (Karl, O'Leary-Kelly, & Martocchio, 1993).

Belief in ability to perform a task (self-efficacy) impacts effort, persistence, interest, and success at difficult tasks (Gist, 1987).

Task-specific self-efficacy/confidence explains up to 28% of performance improvements in on-the-job performance (meta-analysis; Stajkovic & Luthans, 1998).

Goal: Develop the Operational Assessment of Training Scale (OATS) to assess training via survey

To assess training, we want to know (Bandura, 1977, Bandura & Adams, 1977)[1]:

1. To what degree does training impact self-efficacy? (Efficacy)
2. How pertinent was training for mission tasks and real operations? (Relevance)

- Wanted the survey applicable across systems, operators.
  - NOT a test of knowledge or learning (would need to be tailored to systems knowledge and operator tasking)

[1] Magnitude, or the idea that self-efficacy is associated with taking on more difficult tasks, was not assessed because operators typically do not have the ability to choose which tasks they are assigned.

How does survey development and validation work?

- It can be a long, iterative process
- We'll skip around to focus on relevant aspects of OATS at this time

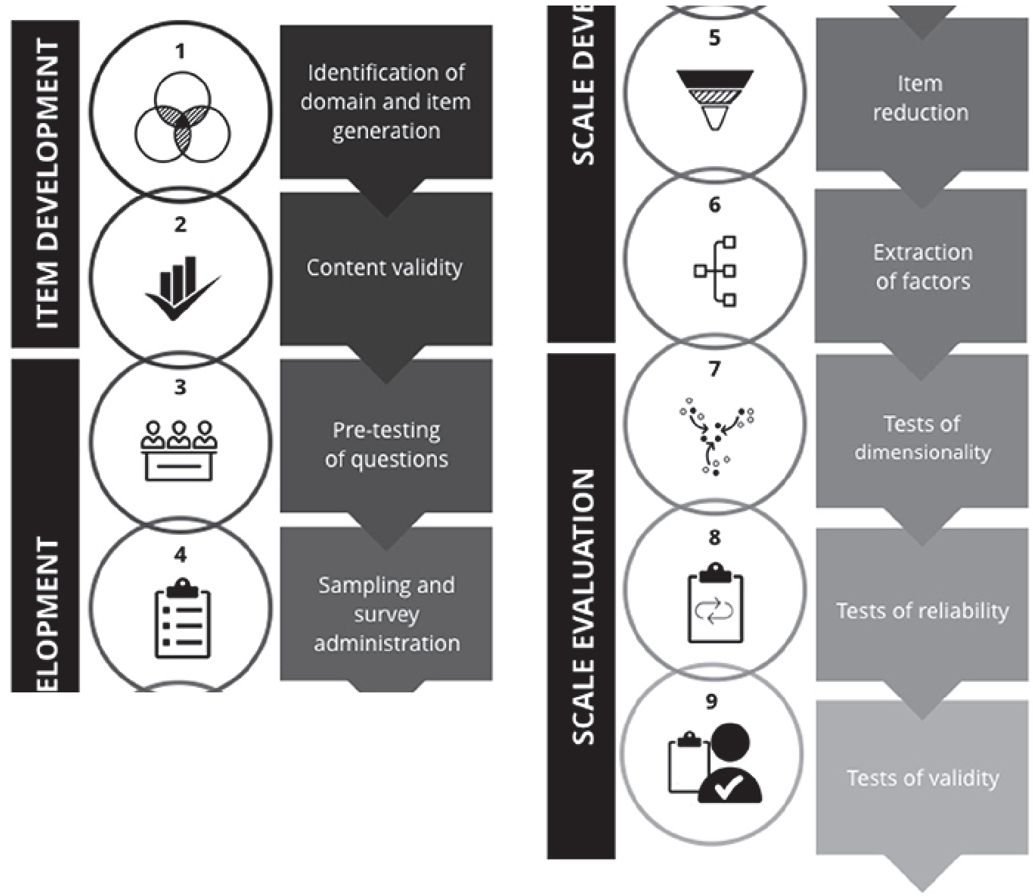


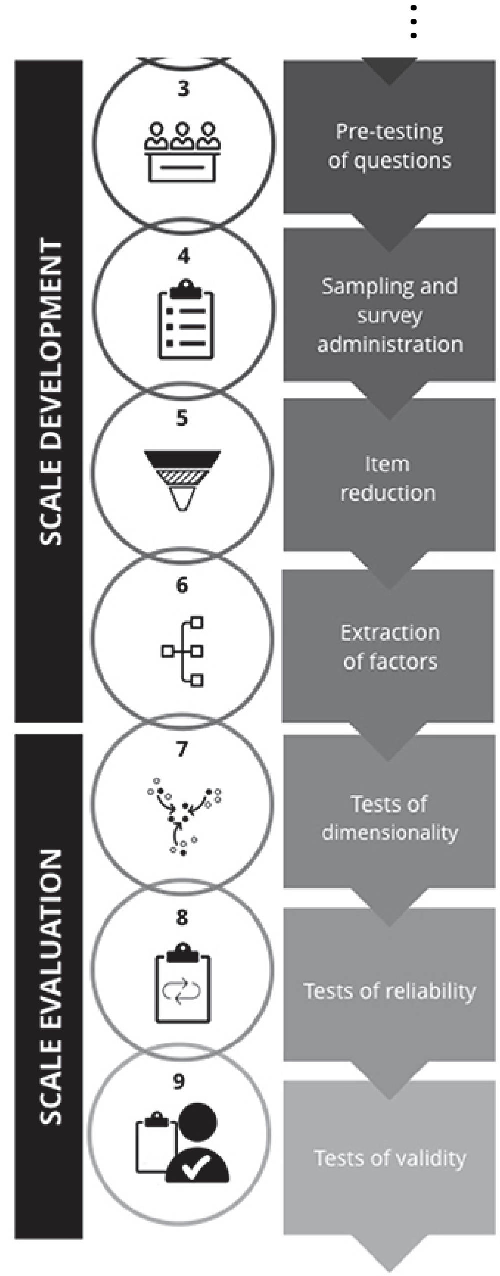
Figure adapted from Boateng et al., 2018

OATS Item Subscales

Created a set of items to see if we could capture two training constructs.

1. Efficacy: To what degree does training impact self-efficacy?

2. Relevance: How pertinent was training for mission tasks and real operations?

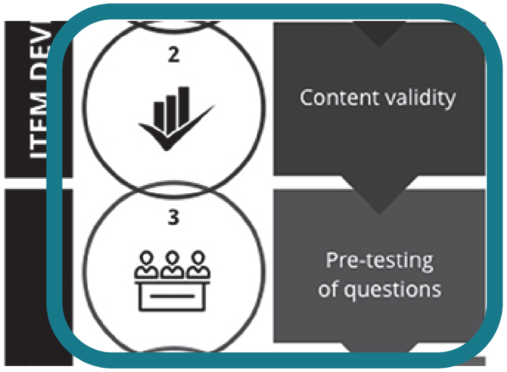


OATS – Operational Assessment of Training Scale

Figure adapted from Boateng et al., 2018

An initial 15 item survey was created with Likert-style responses

ITEM #	SUB	ITEM
1	R	I can see myself using what I learned in training during real operations.
2	R	All of the information covered was relevant to how I interact with the system.
3	R	Training accurately portrayed operations in the field.
4	R*	Training did not cover important ways I interact with the system.
5	R*	Training adequately covered all important ways I interact with the system.
6	R	I would not make changes to the course content.
7	R*	The course covered topics I don't think should have been covered.
8	R*	The training had a lot of information that wasn't relevant to me.
9	R	The course's level of difficulty was appropriate for someone in my position
10	E	I'd be confident using the system during real operations without additional training.
11	E*	I'd want additional training before using the system during real operations.
12	E	The training improved my understanding of how to interact with the system.
13	E	The training prepared me to properly interact with the system.
14	E	Training prepared me to solve common problems.
15	E	The training prepared me to easily use the system to accomplish my mission.



Subscales

E = Efficacy

R = Relevance

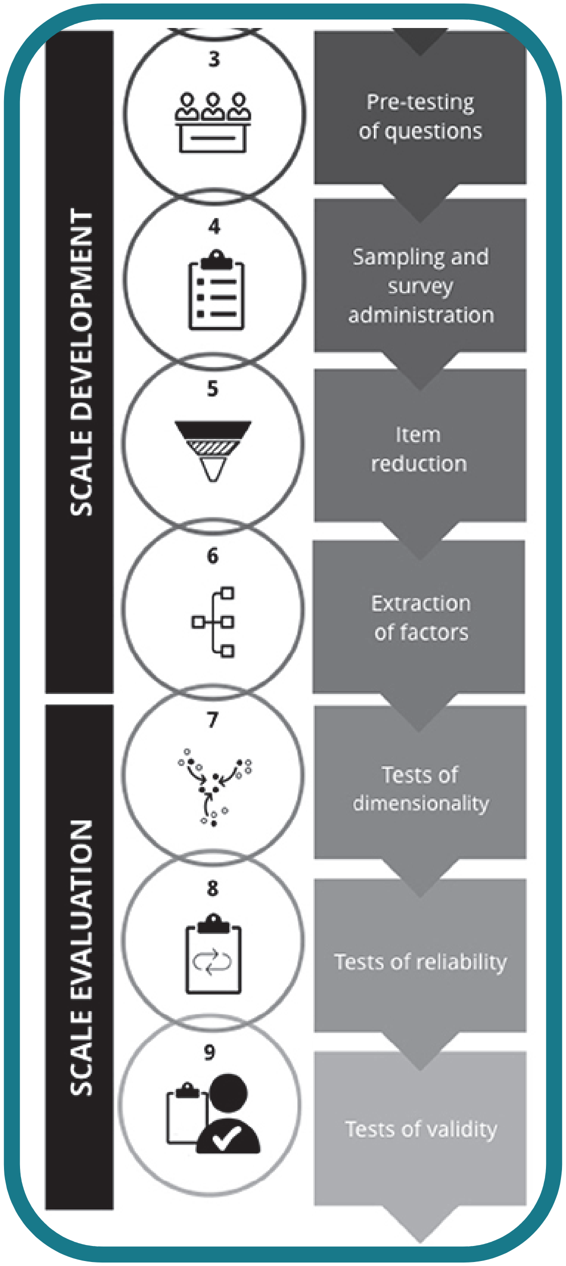
- Developed with scale experts, psychologists, system SMEs, trainers
- Items target "training" broadly
- Reverse-scored items denoted *

SME – Subject Matter Experts; Figure adapted from Boateng et al., 2018

“Structure validation” checks how much observed data matches proposed structure

- We proposed that a two-factor for initial structural validation.
  - Propose that 2-factor model “fits better” than a 1-factor model of training.[1]
  - Check for reliable, consistent responses within and between sub-scales.[2]

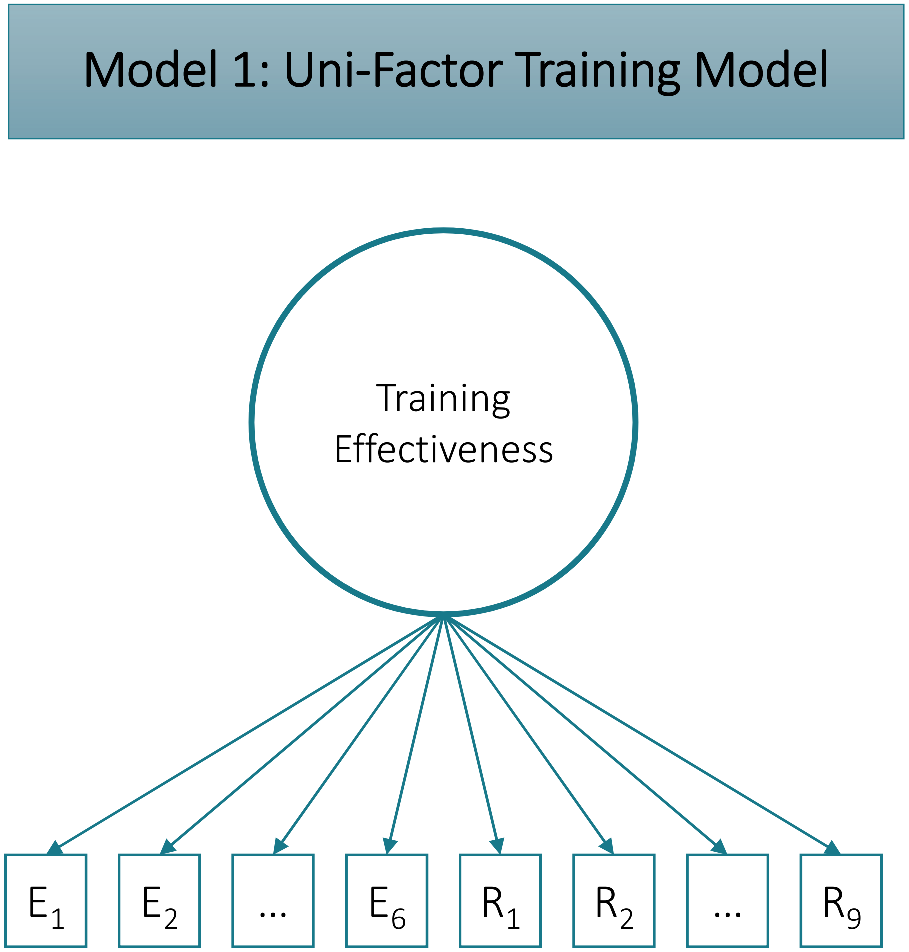
- Getting to a final model is iterative:
  1. Assess fit of proposed model.
  2. Compare to relevant alternatives.
  3. Refine best model based on fit statistics, SMEs.
  4. Return to 1. Repeat 2-4.

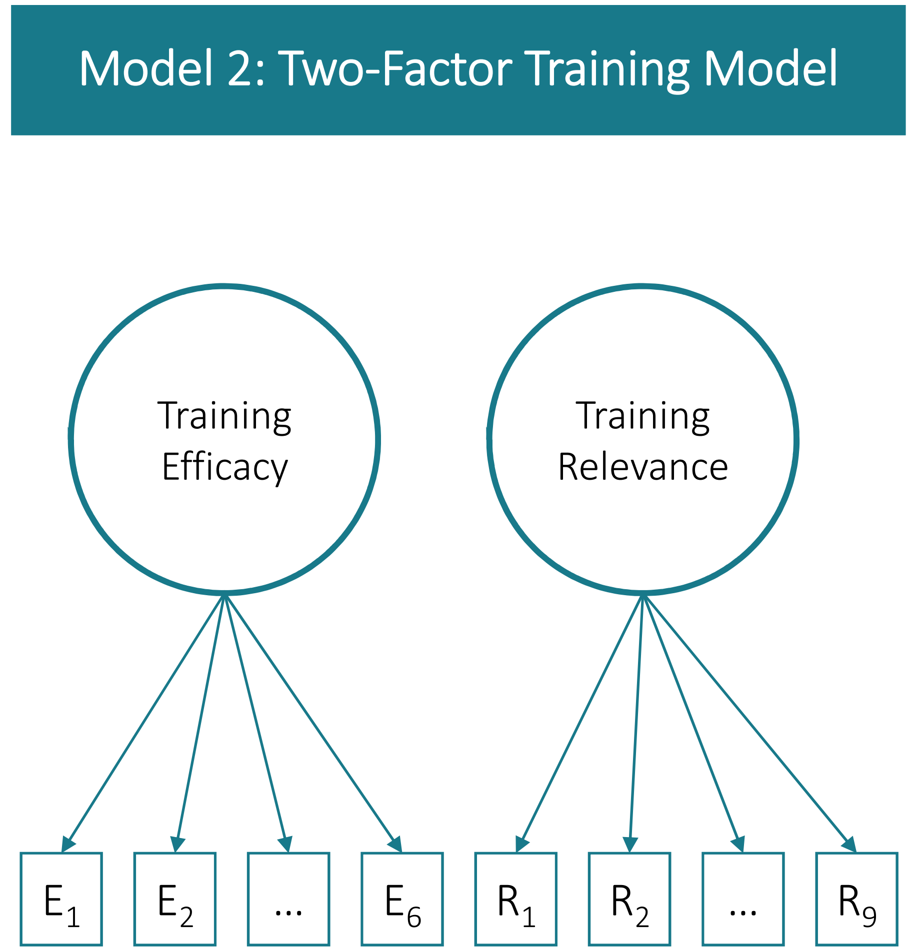


[1] Structure validity; [2] Inter-item and item-total correlations

Figure adapted from Boateng et al., 2018

We propose Model 2 is best, but we need to ensure observed (collected) data confirm that

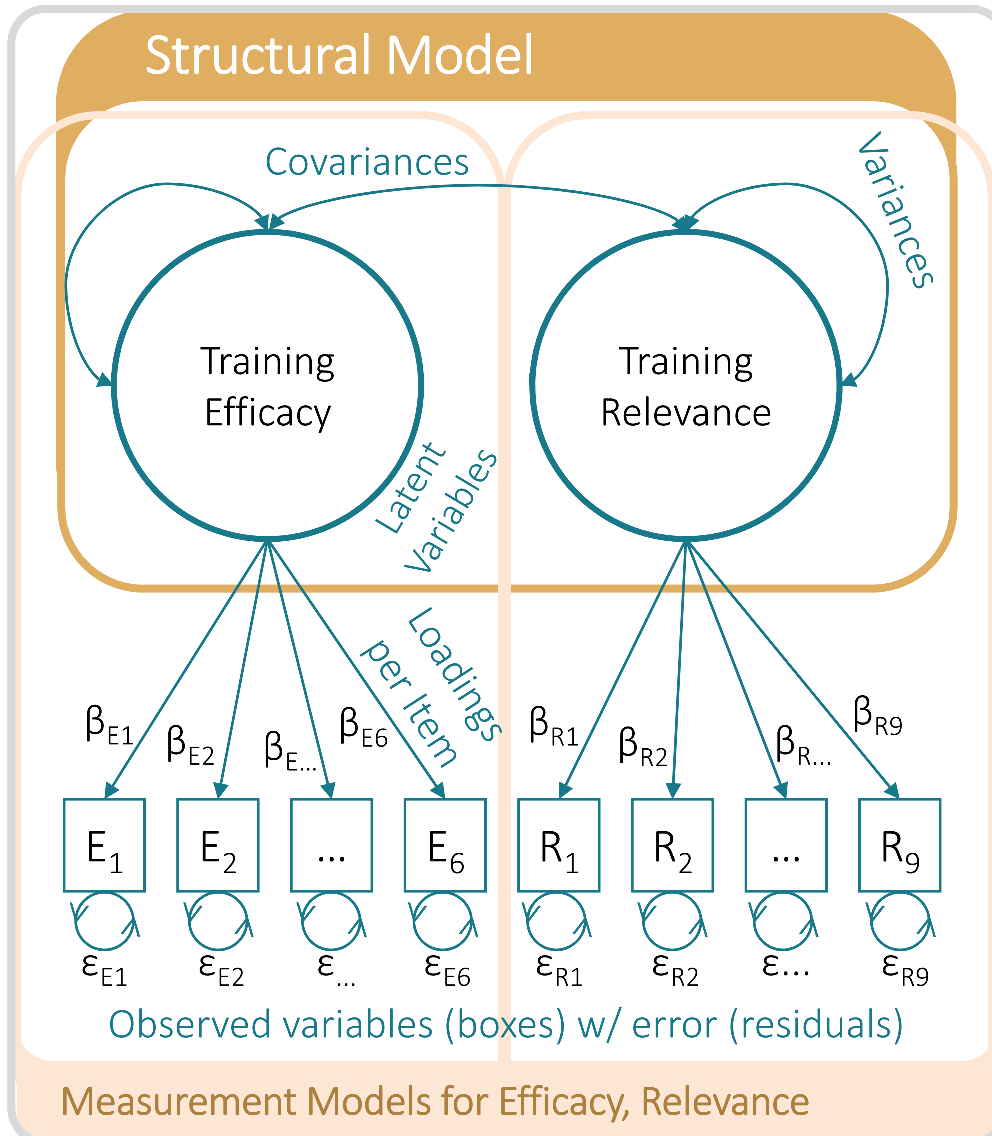




E_n – OATS Efficacy items 1-6; R_n – OATS Relevance items 1-7

Structural equation modeling (SEM) approach: Propose model, test against observed data

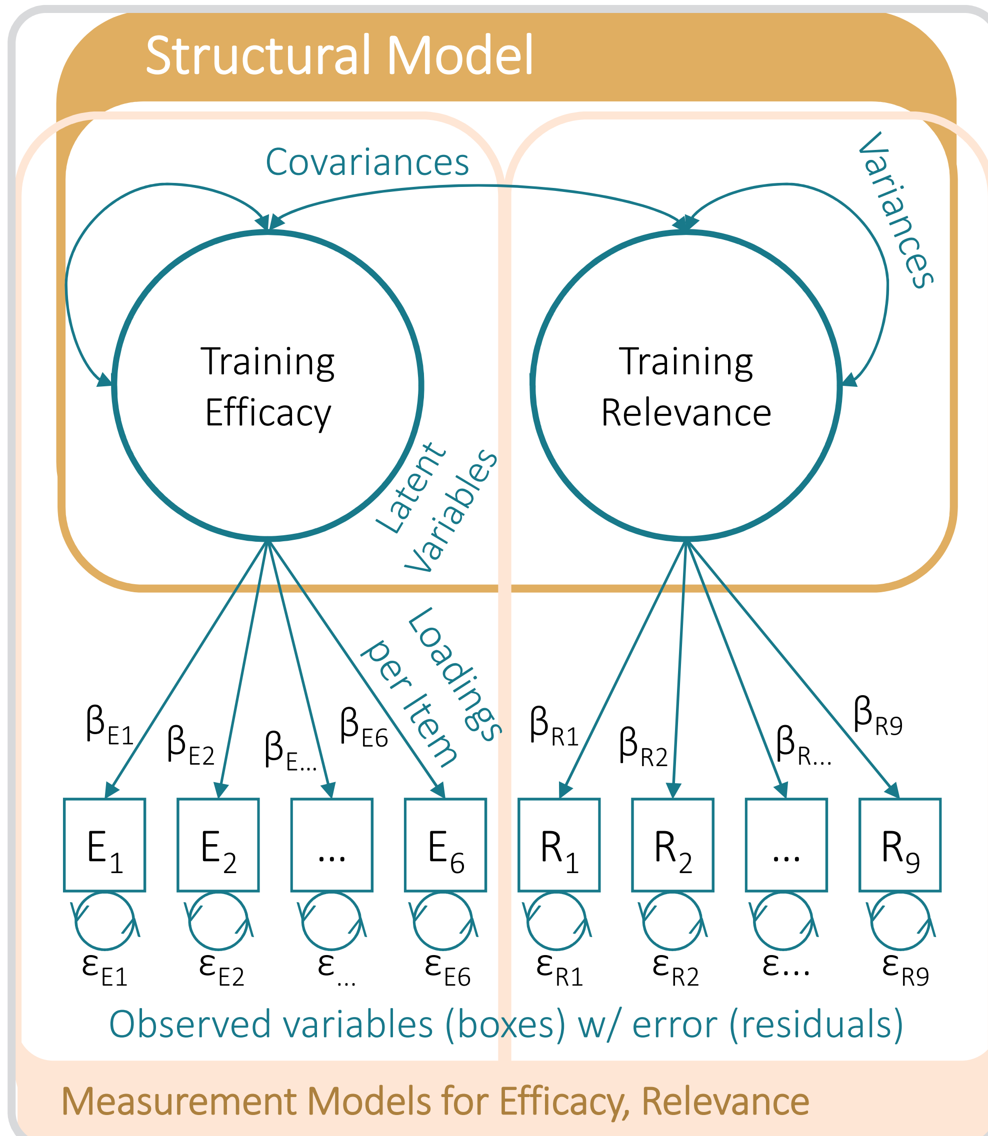
- Start with a theoretically motivated structural model and measurement model(s). Includes:
  - Observed variables
  - Latent variables
  - Paths
  - Variances
- Check fit of proposed model (w/ constraints) via bootstrapped ML; compare to observed data's relationships.
  - Does proposed `cov(X,Y)` fit look like observed data?



ML – Maximum Likelihood

Structural equation modeling (SEM) approach: Propose model, test against observed data

- Model fit assessed by:
  - Farther from “worst,[1]” closer to “best[2]” models: CFI, TLI (goal >.90)
  - Degree of misspecification, RMSEA (goal < .08)
  - Does assumed structure fit observed structure? χ^2 test (goal = small)
  - AIC, BIC (goal = lower)



[1] Worst = Baseline model (0 covariance); [2] Best = Saturated model

AIC – Akaike information criterion; BIC – Bayesian information criterion; CFI – comparative fit index; RMSEA – root mean square error of approximation; TLI – Tucker-Lewis index

Sampling: Administered across broad systems, trainings of interest

- Goal: Gather enough survey responses from relevant peoples to understand structure and reliability of scale.
  - Deployed across a large range of systems.
  - Sampled range of operators, system admins, maintainers, etc.

- Collected N = 812 responses across 24 systems.
  - Kept all responses for analysis.
  - Various filtering processes: Similar outcomes.

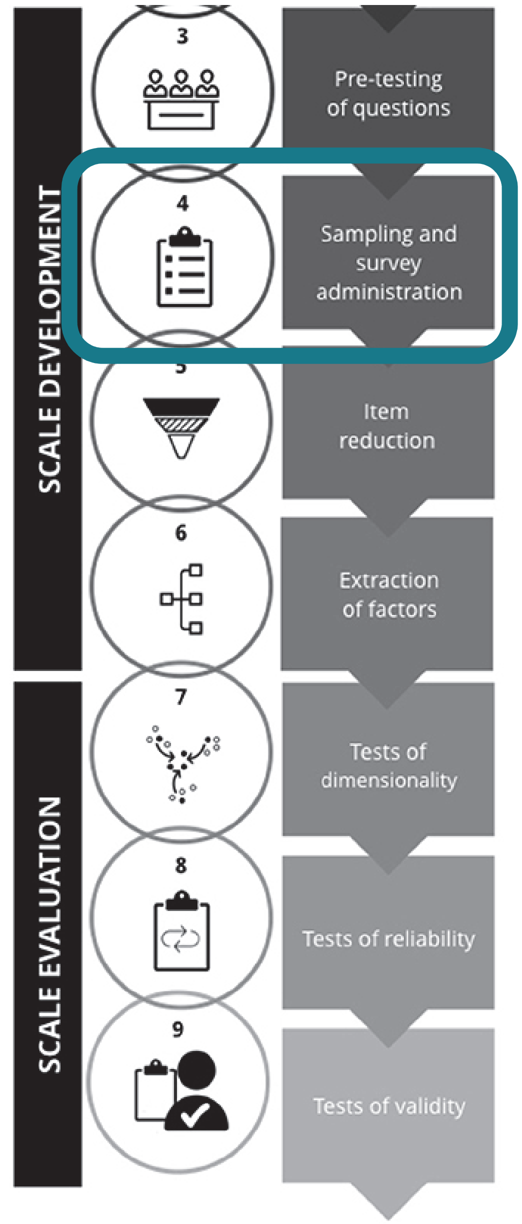
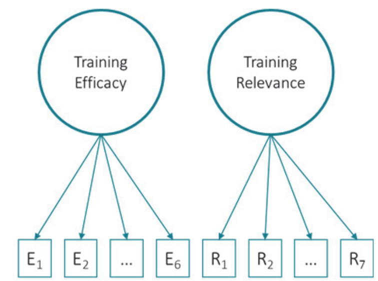


Figure adapted from Boateng et al., 2018

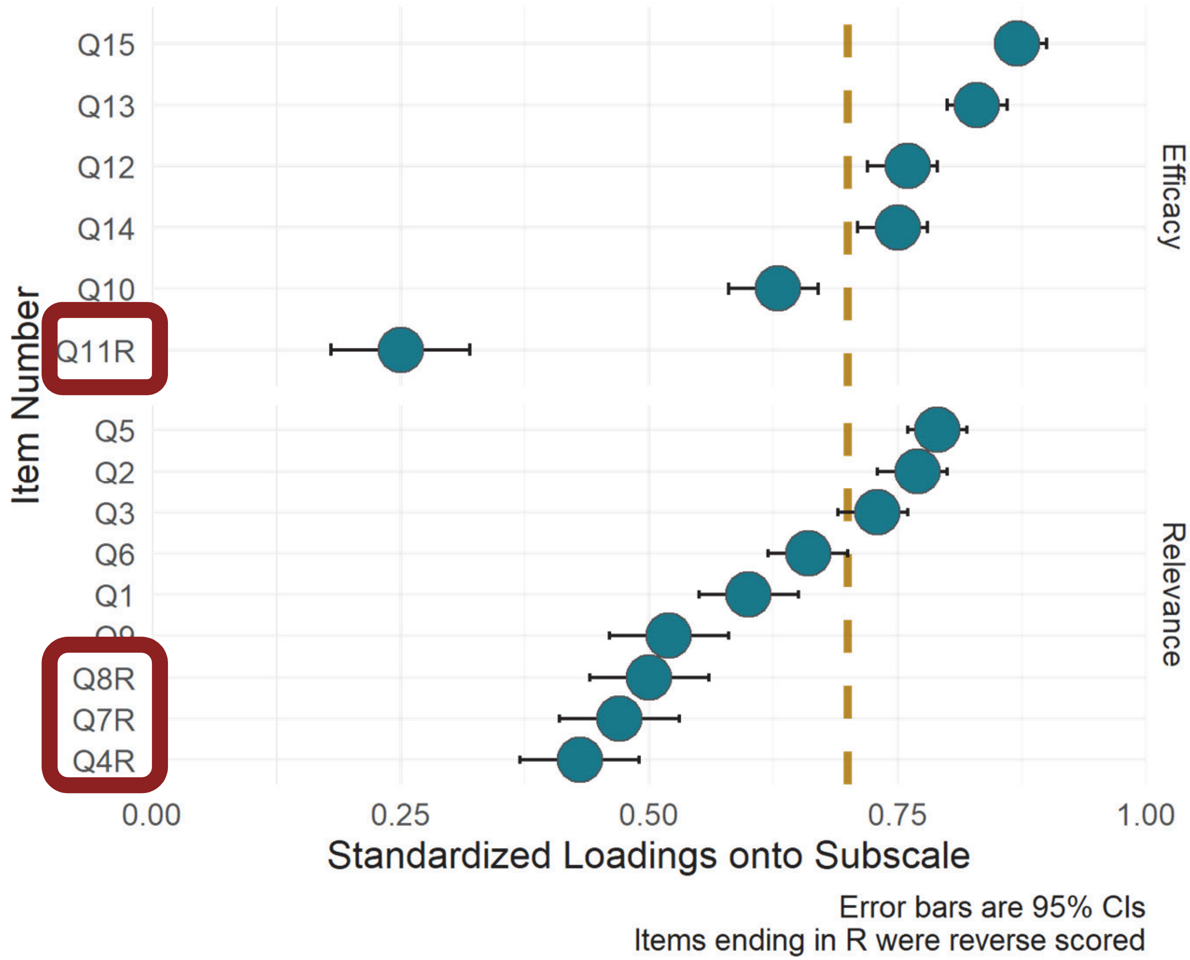
Initial model comparisons



Metric	Goal	Model 1: Uni-Factor Training Model	Model 2: Two-Factor Training Model
CFI	> .90	0.80	0.84
TLI	> .90	0.76	0.81
RMSEA	< .05	0.128, $p < .001$	0.116, $p < .001$
Model χ^2	Small	1202.03, $p < .001$	991.34, $p < .001$
df χ^2		90	89
Model 2 fits better than Model 1, $\chi^2(1) = 210.69$, $p < .001$			

Both models fit poorly.
Why?

Reverse-scored items load poorly onto their factors (Model 2 below)



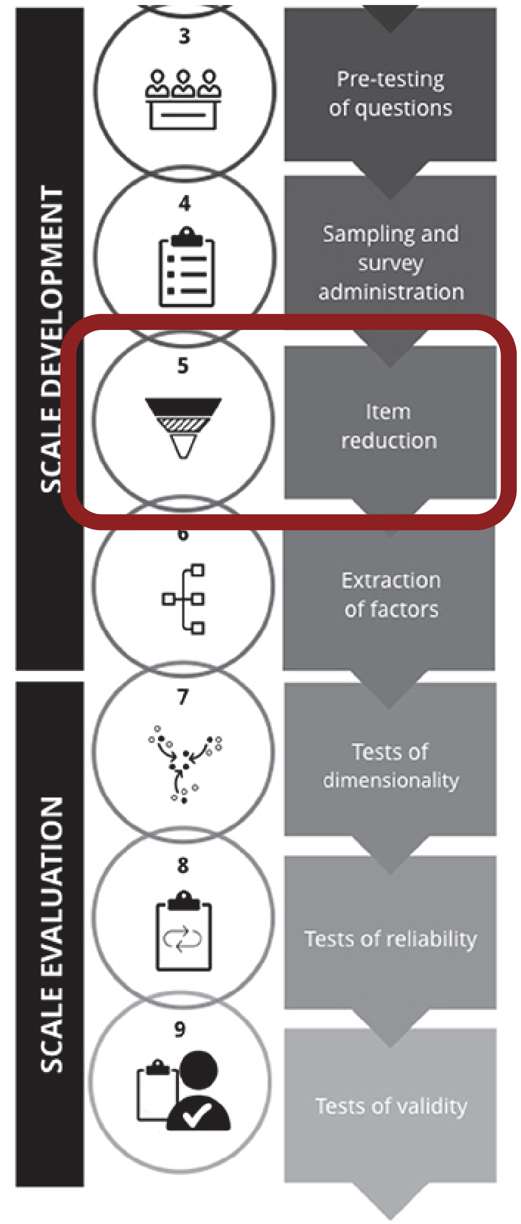
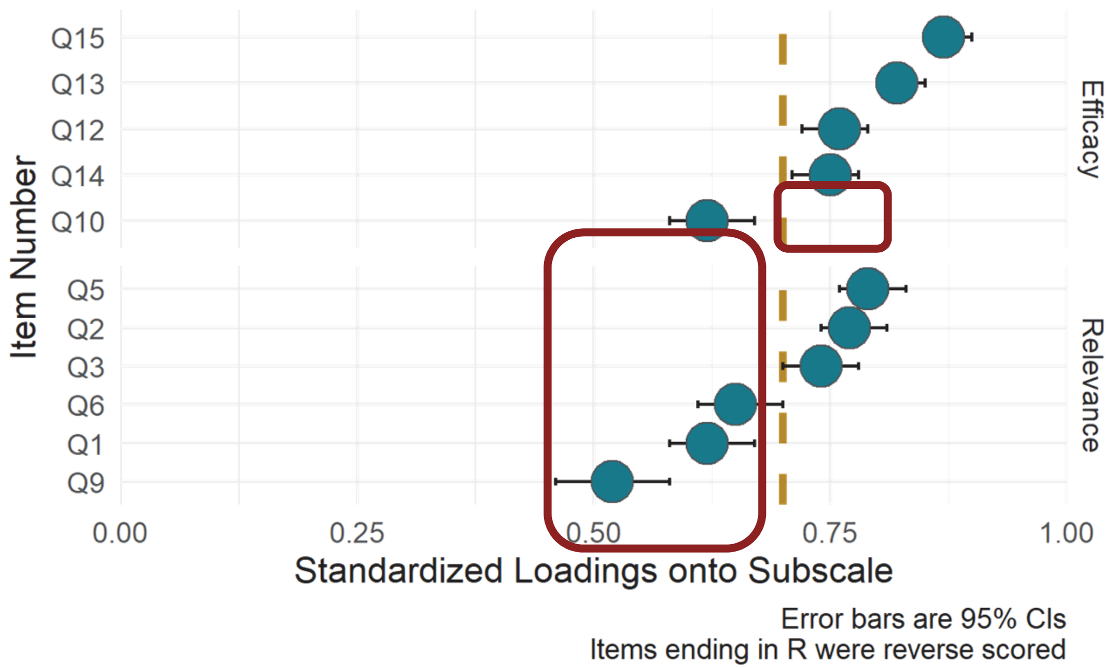


Figure adapted from Boateng et al., 2018

CIs – Confidence Intervals

Revised set of response items

- Removed all reverse-scored items.
  - Consistent with other military T&E results.[1]
  - Relatively common in practice.
- Cut items below a .70 cutoff
- Also cut Q14 to keep 3 items per subscale.



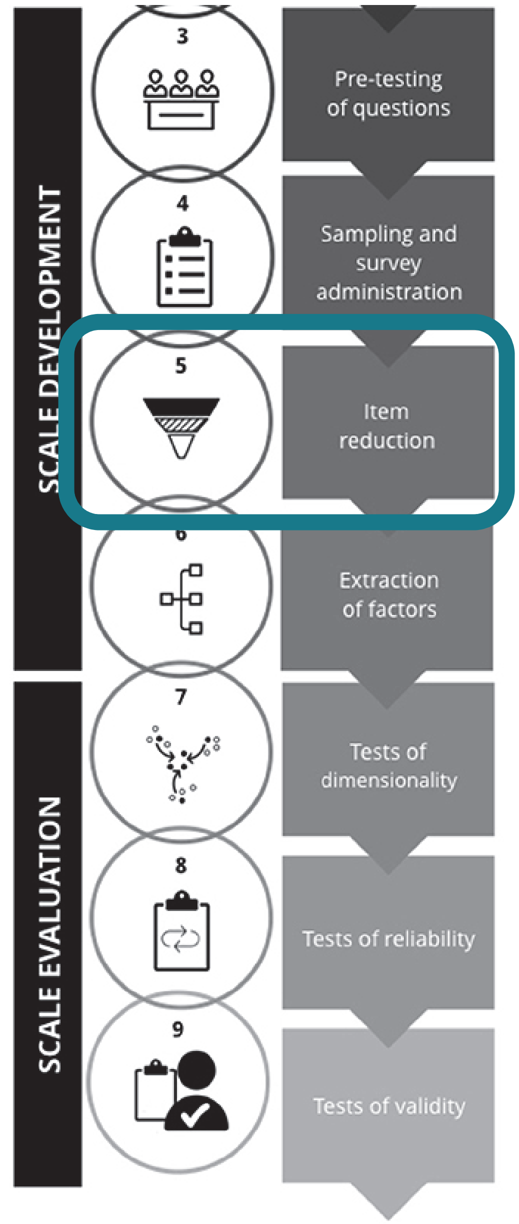
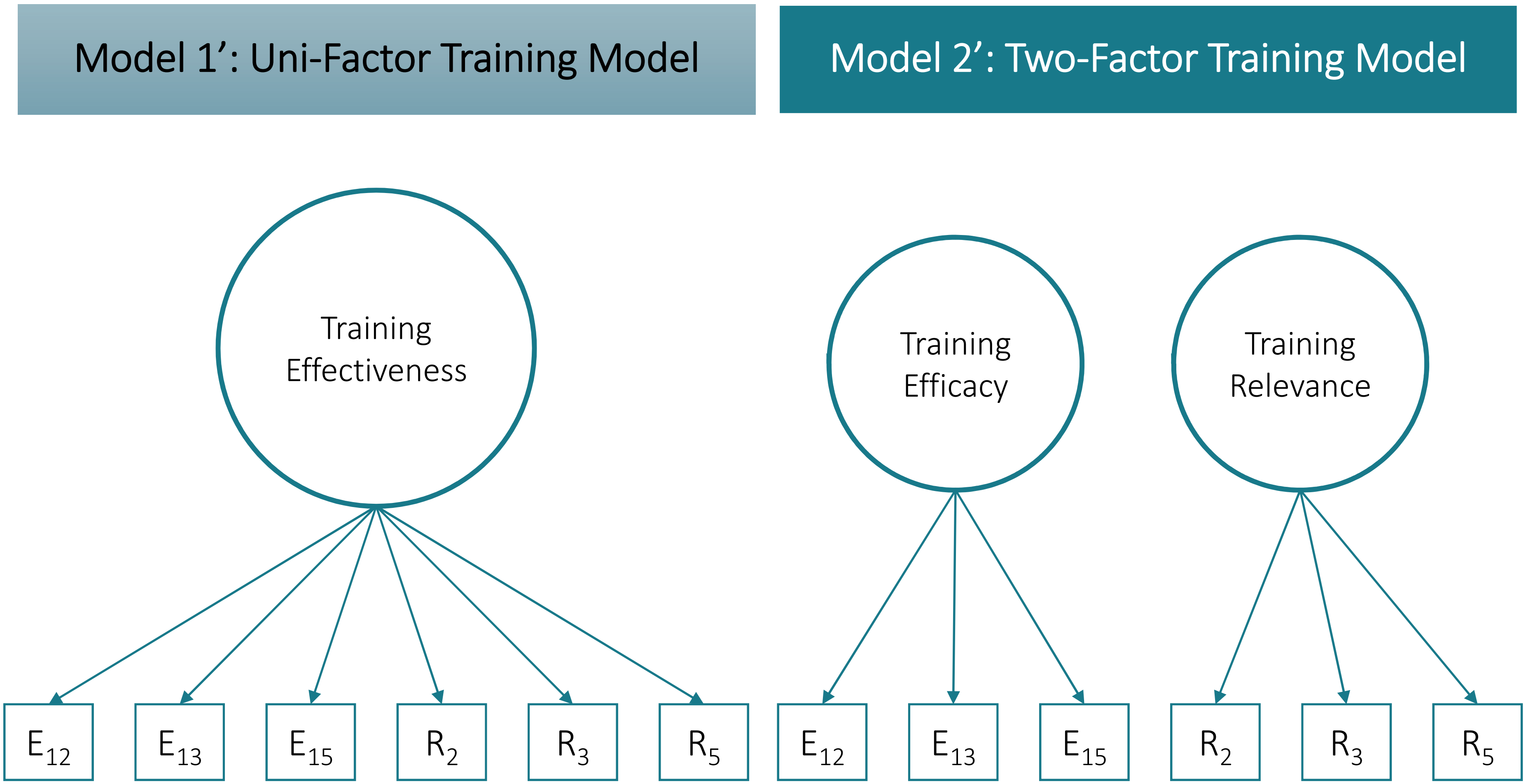


Figure adapted from Boateng et al., 2018; CIs = confidence intervals

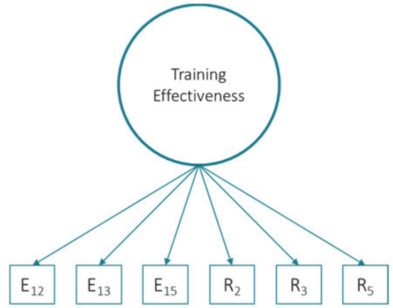
[1] Possibility that positively and negatively worded items measure different underlying constructs or have other issues, Weem & Onwuegbuzie, 2001; Dalal & Carter, 2014;Barnette (2000).

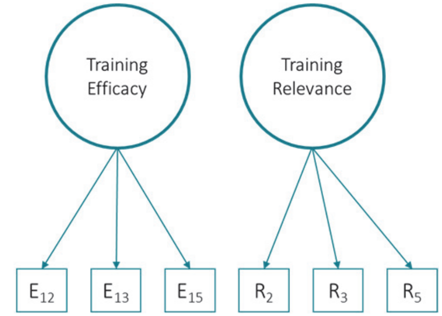
Revised 1- and 2-factor models for comparison



E_n = OATS Efficacy items 12, 13, 15; R_n = OATS Relevance items 2, 3, 5

Revised model comparisons

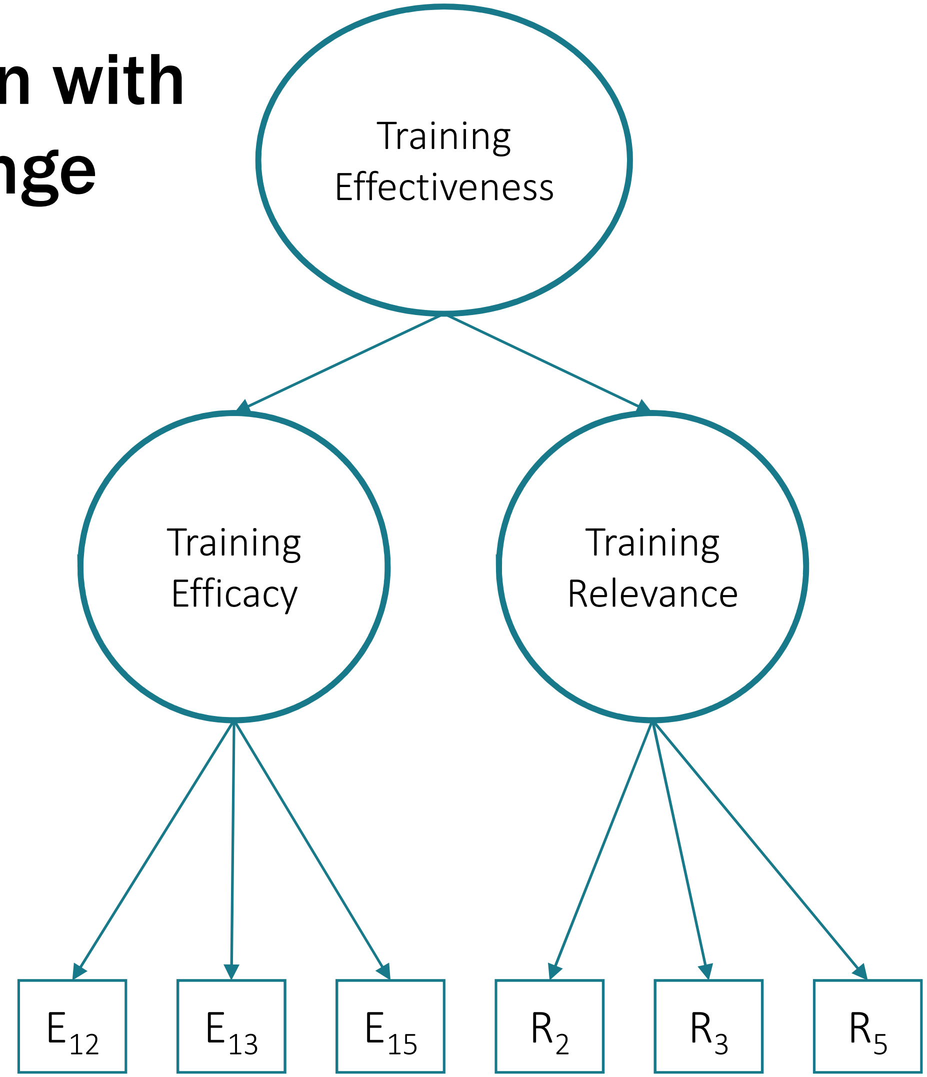




Metric	Goal	Model 1’: Uni-factor Training Model	Model 2’: Two-Factor Training Model
CFI	> .90	0.94	0.99
TLI	> .90	0.91	0.99
RMSEA	< .05	0.140, $p < .001$	0.052, $p = .401$
Model χ^2	Small	141.86, $p < .001$	24.41, $p = .002$
df χ^2	n/a	9	8
AIC	Lower	14,516	14,401
BIC	Lower	14,572	14,461
Model 2’ fits better than Model 1’, $\chi^2(1) = 117.45$, $p < .001$			

Re-standardizing the solution with a roll-up score does not change the solution fit.

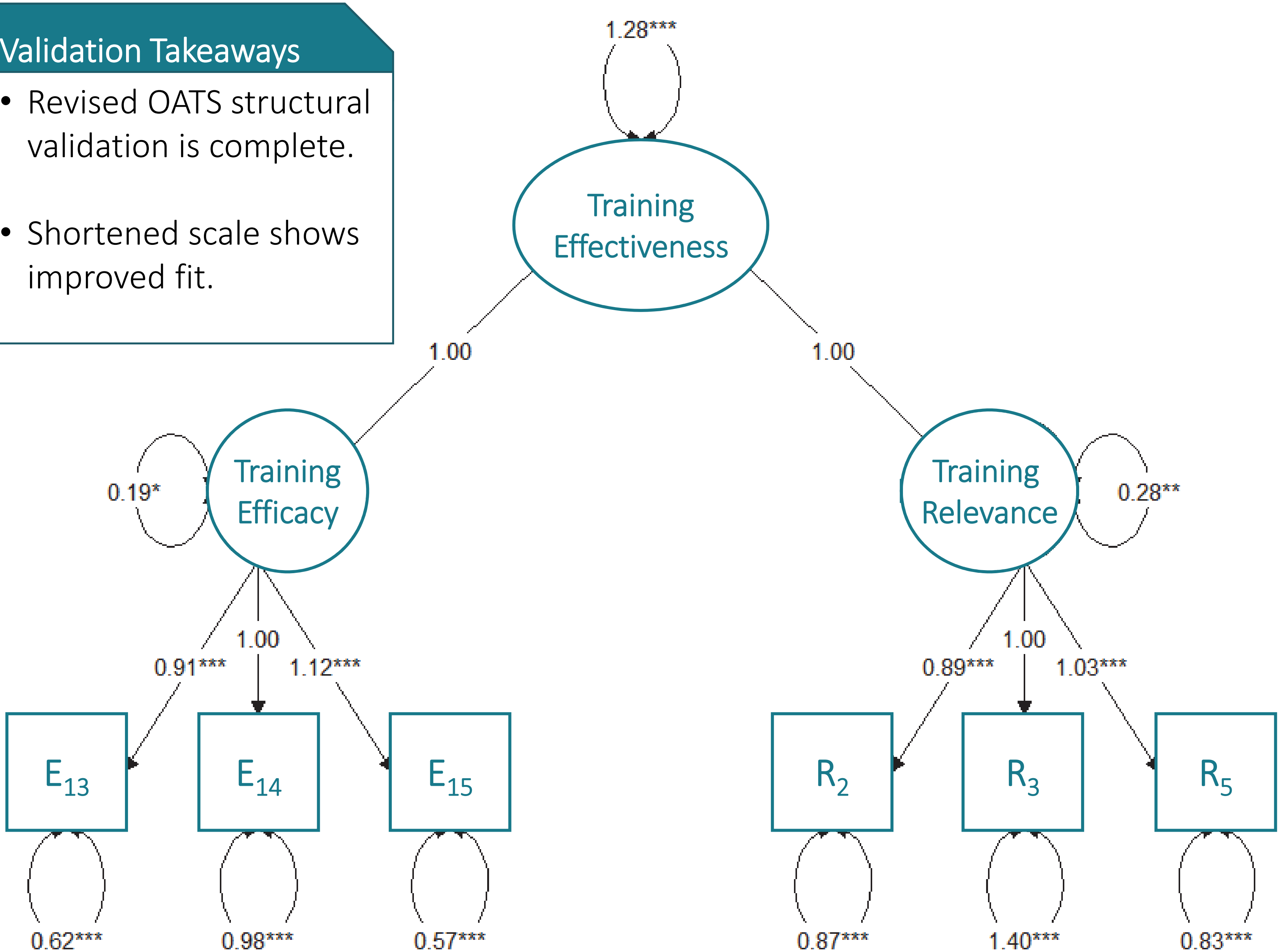
Metric	Goal	Model 3: Roll-up Model
CFI	> .90	0.99 ✔
TLI	> .90	0.99 ✔
RMSEA	< .05	0.052, $p < .401$ ✔
Model χ^2	Small	24.41, $p < .002$
df χ^2	n/a	8
AIC	Low	14,553
BIC	Low	14,609



Marker method used: Assumes both loadings from Training Effectiveness → Efficacy, Training Effectiveness → Relevance are equal.

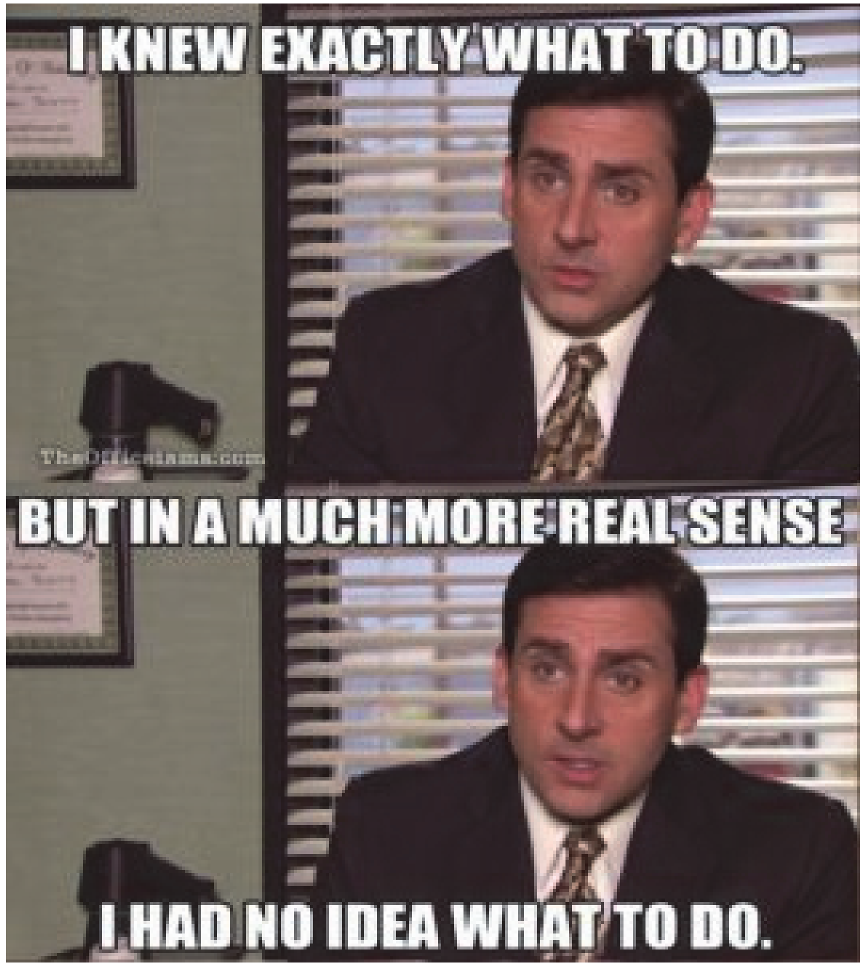
Validation Takeaways

- Revised OATS structural validation is complete.
- Shortened scale shows improved fit.



Note: This solution standardizes loadings onto the total factor.

What about administration schedules?



What about administration schedules?

With continued system use, operators can better assess how much training enabled efficacy:

- Across all relevant missions
- Across the system’s suite of capabilities
- With a wider range of teammates

Additional OATS data collection can help understand these changes over time.

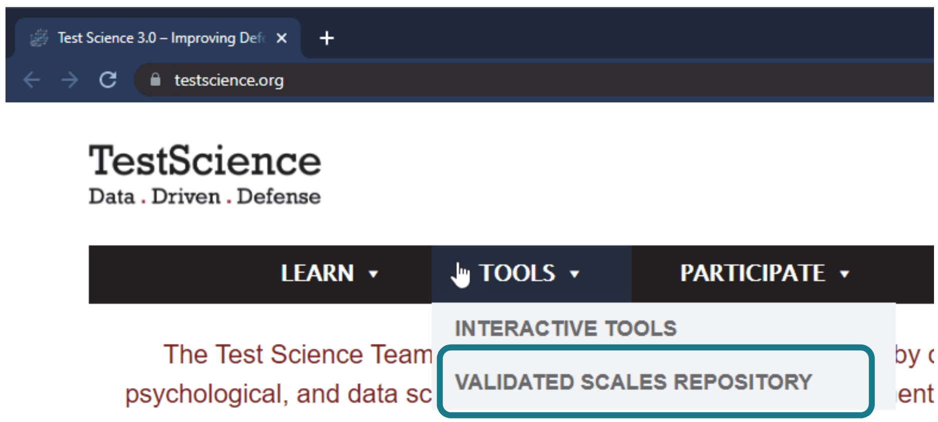
Takeaways

- Administer the shorter, six-item OATS moving forward.
- Report both subscales (plus roll-up)
- *Recommended*: Administer at least two (or more) times

ITEM #	SUB	ITEM
1	R	All of the information covered was relevant to how I interact with the system.
2	E	The training prepared me to easily use the system to accomplish my mission.
3	R	Training accurately portrayed operations in the field.
4	E	The training prepared me to properly interact with the system.
5	E	Training prepared me to solve common problems.
6	R	Training adequately covered all important ways I interact with the system.

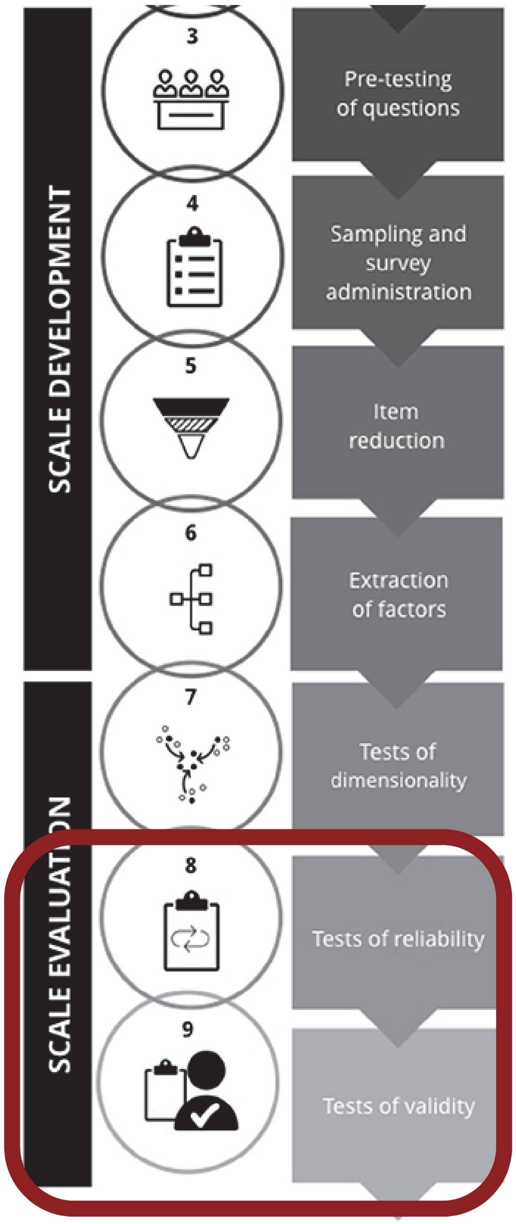
Questions?
Brian Vickers, bvickers@ida.org

Visit TestScience.org for more information on surveys for T&E.



Out of scope: Criterion validity, test-retest reliability

- Outside the scope of this presentation: additional types of reliability and validity
  - Reliability within individuals at different times[1]
  - Correlations with outcomes (e.g., suitability)[2]

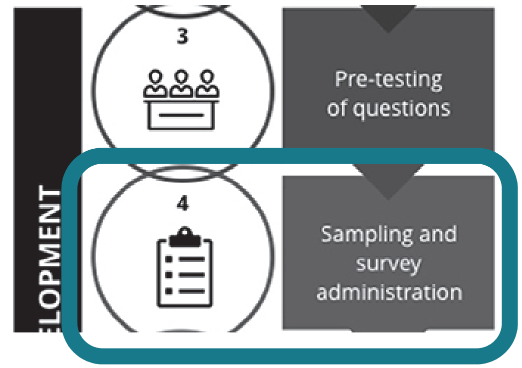


[1] Test, re-test reliability; [2] Criterion validity

Figure adapted from Boateng et al., 2018

Backups

Filtering Summary



Results briefed here use all data.

- Potential filters we looked at are included below.

Table. *Potential reasons to filter data.*

			Bad. Contradictory response[1]	Good. No contradictory response[1]	Total
Bad. Includes missing items			0	0	0
Good. No missing items	Bad. Straightlining.	Bad. All 1s or 7s.	25	0	25
		Good. Mix of responses.	28	0	28
	Good. Did not straightline.	Bad. All 1s or 7s.	0	118	118
		Good. Mix of responses.	26	615	641
Total			79	733	812

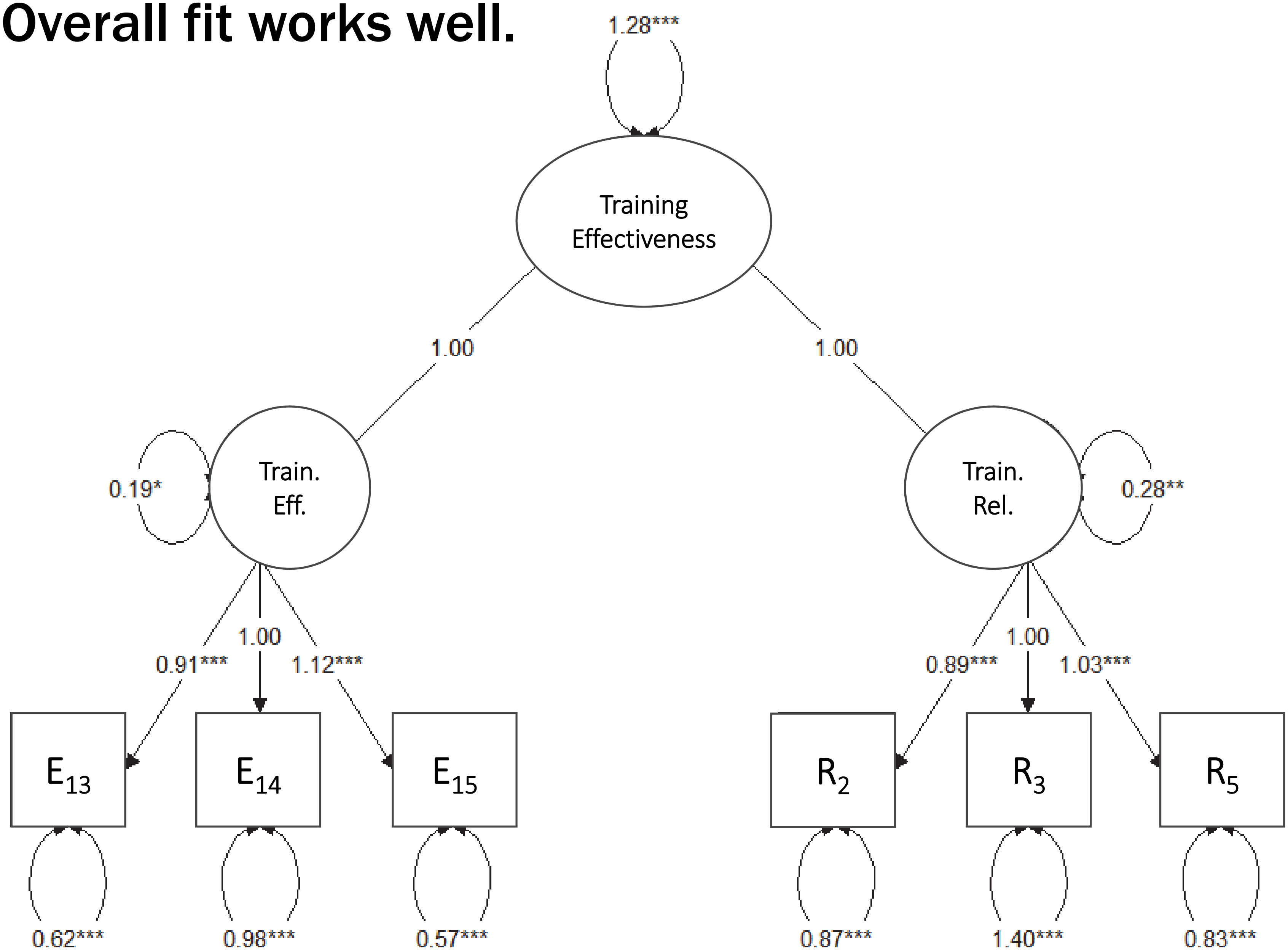
Figure from Boateng et al., 2018

[1] "Contradictory responses" indicates respondents gave opposite answers to at least one reverse-scored item (e.g., 1 and 7, 2 and 6).

Final model fit

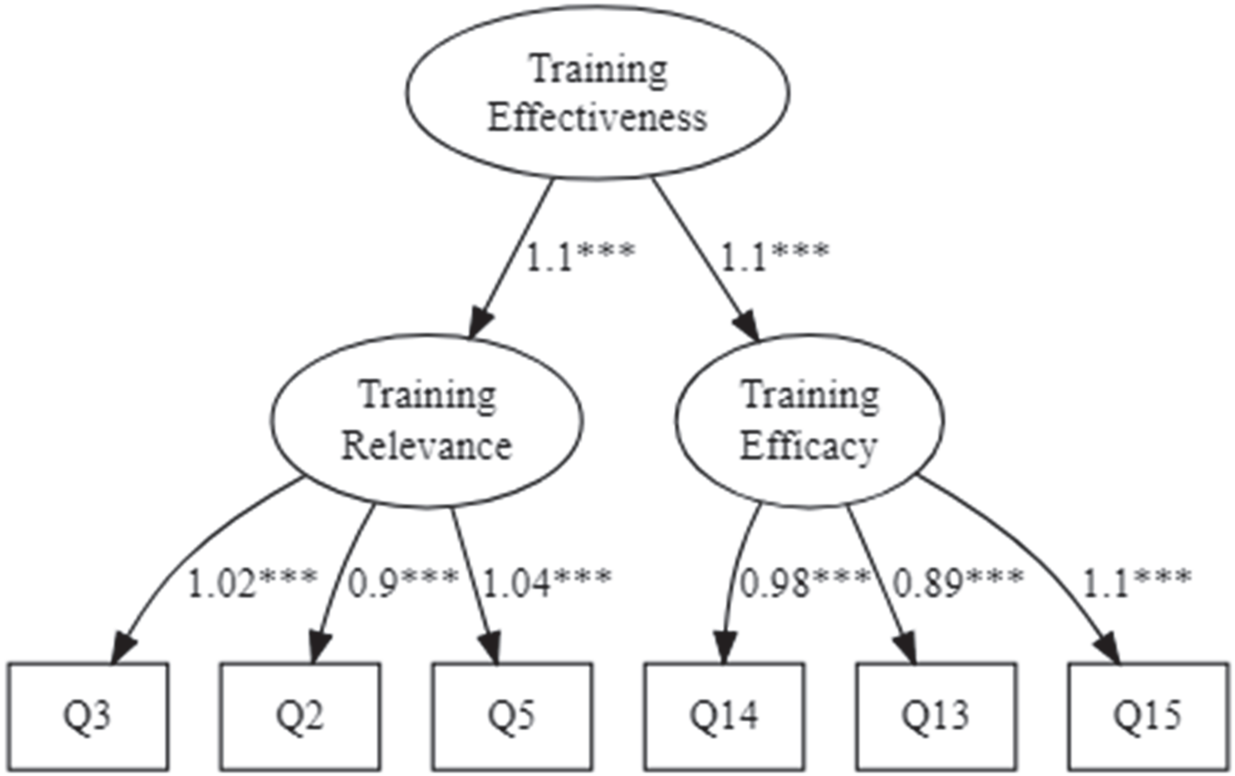
	lhs	op	rhs	est.std	se	z	pvalue	ci.lower	ci.upper
1	R	=~	Q3	0.726	0.021	34.605	0.000	0.685	0.767
2	R	=~	Q2	0.766	0.019	40.007	0.000	0.729	0.804
3	R	=~	Q5	0.816	0.017	47.771	0.000	0.782	0.849
4	E	=~	Q14	0.775	0.017	44.320	0.000	0.741	0.810
5	E	=~	Q13	0.814	0.016	52.097	0.000	0.783	0.845
6	E	=~	Q15	0.874	0.013	67.083	0.000	0.848	0.899
7	Total	=~	R	0.907	0.026	34.330	0.000	0.855	0.959
8	Total	=~	E	0.935	0.027	34.833	0.000	0.882	0.987
9	Total	~~	Total	1.000	0.000	NA		1.000	1.000
10	Q3	~~	Q3	0.473	0.030	15.537	0.000	0.413	0.533
11	Q2	~~	Q2	0.413	0.029	14.066	0.000	0.355	0.470
12	Q5	~~	Q5	0.335	0.028	12.010	0.000	0.280	0.389
13	Q14	~~	Q14	0.399	0.027	14.710	0.000	0.346	0.452
14	Q13	~~	Q13	0.337	0.025	13.260	0.000	0.287	0.387
15	Q15	~~	Q15	0.236	0.023	10.375	0.000	0.192	0.281
16	R	~~	R	0.178	0.048	3.707	0.000	0.084	0.272
17	E	~~	E	0.126	0.050	2.518	0.012	0.028	0.225

Overall fit works well.



Note: This solution standardizes loadings onto the total factor.

Overall fit works well.



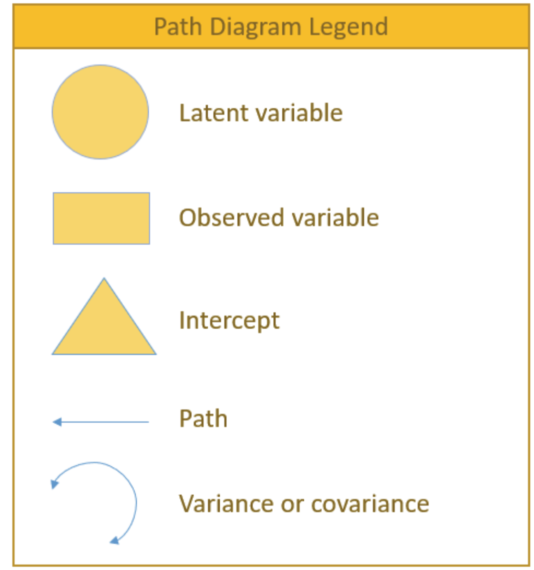
Note: This solution standardizes latent variances for relevance and efficacy at 1.

Belief in ability to perform a task (self-efficacy) impacts effort, persistent interest, and success at difficult tasks (Gist, 1987).

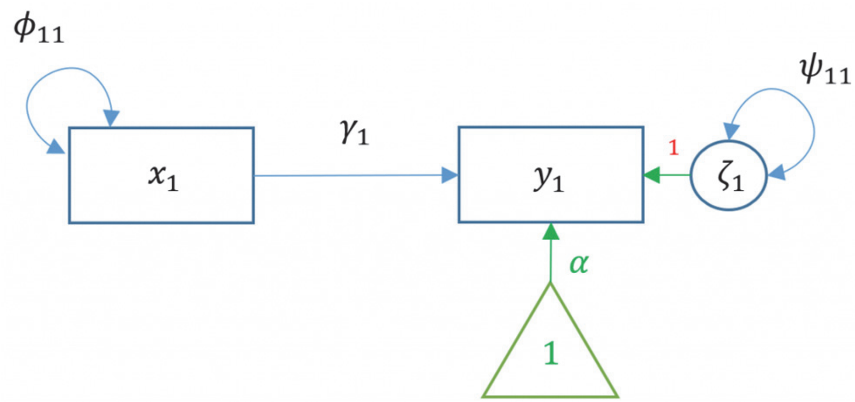
Can fall among three dimensions (Bandura, 1977, Bandura & Adams, 1977).

- Strength: Degree to which they believe they can achieve task.
- Generality: Degree to which it generalizes across situations.
- Magnitude: Ability to apply across all levels of difficulty.

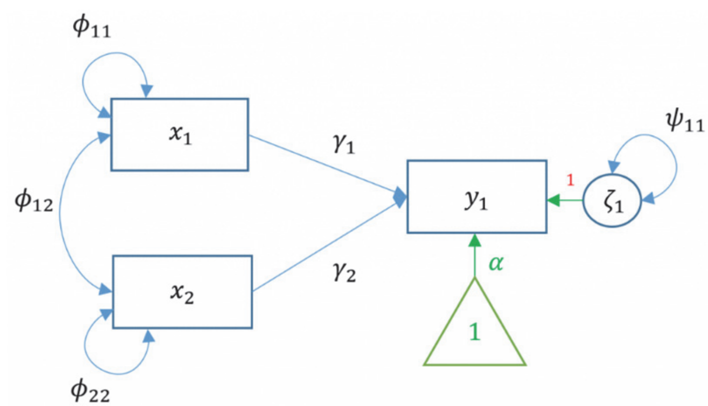
Example: SEM vs. Regression



Simple linear regression



Multiple regression



REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
30-03-2022	OED Draft	07-2021 - 03-2022

4. TITLE AND SUBTITLE

DATAWorks 2022: Measuring Training Efficacy: Structural Validation of the Operational Assessment of Training Scale (OATS)

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Brian Vickers, Daniel Porter, Rachel Haga, Heather Wojton

5d. PROJECT NUMBER

BD-09-2299

5e. TASK NUMBER

229990

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER

D-32972

H 2022-000049

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Public release approved. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: Vincent Lillard

14. ABSTRACT

Effective training of the broad set of users/operators of systems has downstream impacts on usability, workload, and ultimate system performance that are related to mission success. In order to measure training effectiveness, we designed a survey called the Operational Assessment of Training Scale (OATS) in partnership with the Army Test and Evaluation Center (ATEC). We confirmed that OATS exhibited a two-factor structure for training relevance and training efficacy. Additionally, a shortened, six-item measure of the OATS with three items per subscale continues to fit observed data well, allowing for quicker assessments of training. We discuss various ways that the OATS can be applied to one-off, multi-day, multi-event, and other types of training events.

15. SUBJECT TERMS

Human Systems Interactions (HSI), survey, survey development, survey validation, training

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Vincent Lillard
Unclassified	Unclassified	Unclassified	Unlimited	41	**19b. TELEPHONE NUMBER** *(Include area code)* (703) 845-2230

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18