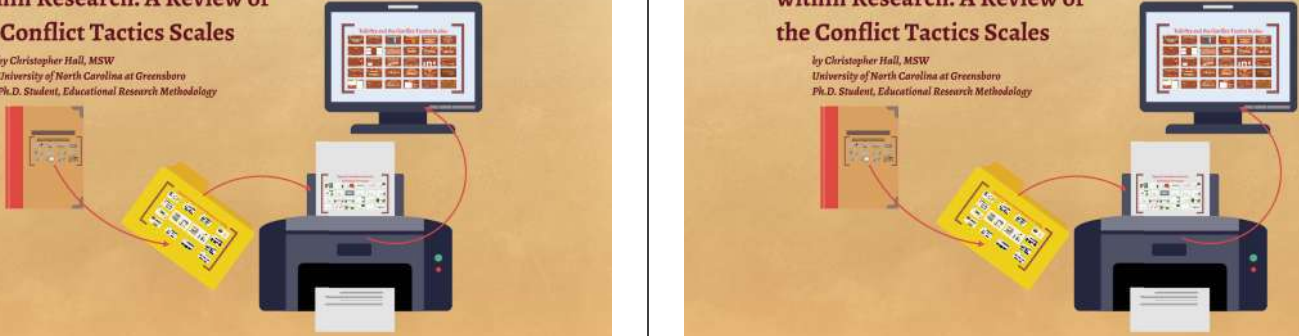


# Information on Validity within Research: A Review of the Conflict Tactics Scales

by Christopher Hall, MSW  
University of North Carolina at Greensboro  
Ph.D. Student, Educational Research Methodology



# Information on Validity within Research: A Review of the Conflict Tactics Scales

by Christopher Hall, MSW  
University of North Carolina at Greensboro  
Ph.D. Student, Educational Research Methodology



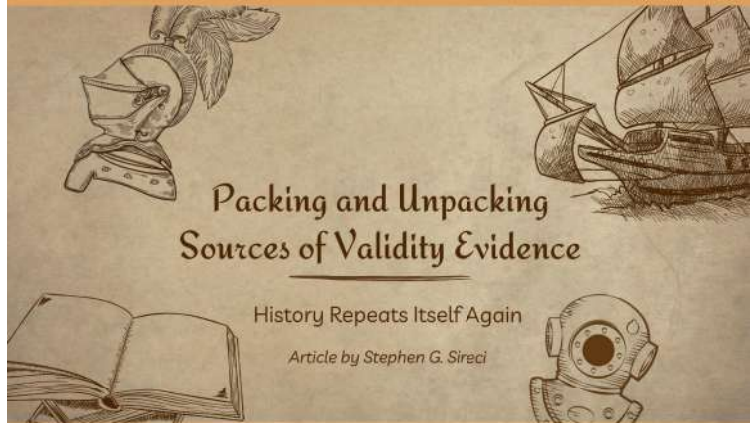
## The History and Context of Validity



## Packing and Unpacking Sources of Validity Evidence

History Repeats Itself Again

Article by Stephen G. Sireci



## Framework for Discussion

Newton and Shaw discuss methods of talking and thinking about validity for educational and psychological measurement (EPM). They identify the following considerations:

- Do not refer to the "validity of the test" - **validity is not a property** of testing instruments
- Do not use **validity modifier labels** - that has changed through history, and we will be discussing why that has shifted over the years
- Validity means different things to different communities - the **context is for EPM**
- **Consensus is important**, but also implies that there is ongoing disagreement and challenges to accepted standards
- Differences between standards, custom, and practice might be due to:
  - Intentional misuse
  - Lack of awareness or misunderstanding
  - Genuine divergence



Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*(3), 301-319. <https://doi.org/10.1037/a0032969>

"I believe validity:

- 1) is *not* an inherent property of a test,
- 2) refers to the interpretations or actions that are made on the basis of test scores, and
- 3) must be evaluated with respect to the *purpose* of the test and how the test is *used* (p.20)."

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (p. 19-37). IAP Information Age Publishing.

## Packing & Unpacking Components of Validity

"From the beginning, validity theory and practices have wavered between simple and complex notions, and multiple types of validity and associated modifiers have been proposed."

"The unitary conceptualization of validity is currently prominent, yet the various aspects or types of validity persevere and are used by test developers and practitioners (p.31)."



## Regardless of the Packaging

*The fundamental notions of the appropriateness of tests will persevere*

Tests Should ALWAYS:

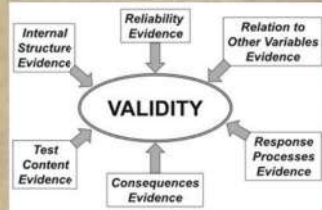
- Demonstrate predicted relationships with other measures of the intended constructs and unintended constructs,
- Contain content consistent with their intended uses,
- Be put to purposes that are consistent with their design and are supported by the evidence (p. 32)"



## Universal Conceptualization of Validity

The current consensus of validity is one of a universal conceptualization - that is validity is about a **unitary construct validity framework** with multiple evidence sources, not multiple types of validity.

This framework was developed to address challenges within research where individuals would argue over which type of validity was most important, and which should receive the most focus or weight in the validation process.



Visual representation of the Universal Conceptualization of Validity from: Peeters, M. J., & Harpe, S. E. (2020). Updating conceptions of validity and reliability. *Research in Social & Administrative Pharmacy*; 16(8), 1127-1130. <https://doi.org/10.1016/j.sapharm.2018.11.017>

## Origins of Validity as a Concept in Research (p. 21-22)



1899-1937

### Origins of Validity

Binet-Simon Scale was largely pragmatic, **defined validity in terms of correlation of test scores with some criterion** (Pearson, Kelley, Thurstone, Bingham, Spearman)



1917-1946

### Military Testing

Army Alpha & Beta tests classifying into most appropriate levels of service using incremental validity (**more variance in performance on a criterion**)


1946

### Challenge to Defining Validity


Rulon stated, "Validity is usually described as the extent to which a test measures what it is purported to measure. This is an unsatisfactory and not very useful concept of validity, because under it the **validity of a test may be altered completely by arbitrarily changing its "purport."**"

## Evolution of Validity to Standards & Frameworks (p. 23-25)


### Jenkins' Question

1946  Criterion data are likely to be **unreliable** due to failure of the criterion-measure to comprise a large and significant part of the total field of performance desired.


### Critiques Continued

1946  Rulon stated that "**validity on a test might be high for one use and low for another**, and the whole question is whether the test does what we are trying to do with it. Accordingly, **we cannot label a test as valid or not valid except for some purpose.**" Ask for proof that the test does its job, but not the same evidence fits all.

### APA Involvement

1952-54  The APA Committee on Test Standards offered four "categories/types" of validity: predictive, status (concurrent), content, and congruent (construct) - then **established the first test standards based on these.**

### Cronbach & Meehl

1955  Construct validity must be investigated whenever no criterion or universe of content is accepted as adequate to define the quality measured; and was introduced to **specify types of research required in developing tests where conventional views of validation are inappropriate**

## Argument Based Approaches to Validity (p. 28-31)

### Messick 1989

"Validation is a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what the test scores mean (p. 29-30)"

### Sireci's Opinion

"Although this validation framework acknowledges that validity can never be established absolutely, it requires evidence that the test measures what it claims to measure; the test scores display adequate reliability, and test scores display relationships with other variables in a manner consonant with the current Standards (p.29)"

### Kane 1992

Proposed argument based approach where the validator builds an argument that focuses on defending the use of a test for a particular purpose based on empirical evidence to support that use.

### 1999 Five Sources

- Validation framework:
- Test content
  - Response processes
  - Internal structures
  - Relations to other variables
  - Consequences of testing

### Kane's Description

"An interpretative argument is an approach to validity rather than a type of validity... a compromise between sophisticated validity theory and reality that at some point we must make a judgment (p. 29)."

### Criticisms

Issues with the argument-based approach includes lack of prescriptive rules, lack of clarity when sufficient evidence gathered leading to validity being about making a persuasive argument.

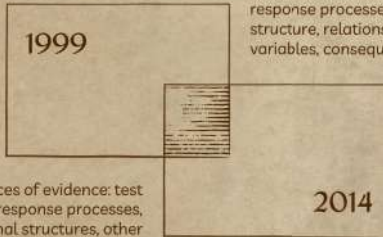
# Shifts in Language & Categories of Validity

## Publication

## Validity Nomenclature

Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal	1952: APA	<b>Categories:</b> predictive, status, content, congruent
Technical recommendations for psychological tests and diagnostic techniques	1954: APA	<b>Types:</b> construct, concurrent, predictive, content
Standards for educational and psychological tests and manuals	1966: APA	<b>Types:</b> criterion-related, construct-related, content-related
Standards for educational and psychological tests	1974: APA, ABECA, ILCTM	<b>Aspects:</b> criterion-related, construct-related, content-related
Standards for educational and psychological testing	1985: ABECA, APA, ILCTM	<b>Categories:</b> criterion-related, construct-related, content-related
Standards for educational and psychological testing	1999: APA, ILCTM	<b>Sources of Evidence:</b> content, response processes, internal structure, relations to other variables, consequences of testing

1999: Sources of evidence: content, response processes, internal structure, relations to other variables, consequences of testing



2014: Sources of evidence: test content, response processes, internal structures, other variables, convergent and discriminant evidence, test-criterion relationships, validity generalizations

In addition: evidence for validity and consequences of testing to include interpretation & uses of test scores intended by test developers, claims made about the test use not directly based on test score interpretations, and consequences that are unintended

# Packing & Unpacking Validity

PACKING	UNPACKING
A test is valid for anything with which it correlates.	Forms of validity to include predictive, status, content, congruent
Does the test measure what it purports to measure?	Validity arguments for clarity, coherence, or plausibility of assumptions
Validity is a unitary concept.	Five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and testing consequences

Brest, G. G. (2008). Packing and unpacking sources of validity evidence: History repeats itself again (PowerPoint presentation). The Centers of Validity Research, New Directions in Applications, University of Maryland, College Park, VA, United States. <https://www.elsevier.com/locate/jpr/packing-and-unpacking-sources-of-validity-evidence-history-repeats-itself-again>

# Intimate Partner Violence Research & Validity

# VALIDITY FOR INTIMATE PARTNER VIOLENCE PROFESSIONALS



## FOR IPV PROFESSIONALS:

Interpreting and applying the AERA, APA, NCME (2014) *Standards for Educational and Psychological Testing* to guide IPV professionals in developing a critical lens of research and literature in their field.

### CONTENTS



#### 01 VALIDITY & VALIDATION

Information providing insight into validity and validation within research

#### 02 RELIABILITY & MEASUREMENT

Measuring precision and consistency in research studies and instruments

#### 03 PSYCHOLOGICAL TESTING

Standards behind psychological testing instruments and strategies

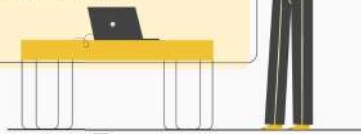
#### 04 OTHER DYNAMICS AND CONCLUSIONS

Exploration of additional considerations and overall conclusions

## 01: VALIDITY VS. VALIDATION

- Validity is how well evidence and theory reflect the interpretation of a measure score
- Validation is collective evidence for the basis of interpreting a test score

*Why does this difference matter for research?*



## THE VALIDATION PROCESS NEVER ENDS

Validation is the joint responsibility of the test developer as well as the group who is using the test in providing relevant evidence and rationale to support how the test scores are used as intended by the developer.

*"The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used."*

—STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING (2014), P. 13

### FOR IPV PROFESSIONALS:

The test user is you, the IPV professional. Common test settings include counseling/educational sessions, victim/survivor interviews, and community-based groups.

## CONSTRUCTS

Constructs are ultimately what is being measured by the test. Understanding the construct is critical to understanding the validity and reliability of the instrument itself. There are two specific issues that can arise with constructs on testing instruments:



### CONSTRUCT UNDERREPRESENTATION

The degree to which a measure fails to assess an important aspect of the construct.



### CONSTRUCT IRRELEVANCE

The degree to which scores are affected by processes that are external to the purpose of the test: itself.

#### EXAMPLES FOR IPV PROFESSIONALS:

If the construct is measuring abusive and violent behavior, you would expect the instrument to be measuring that dynamic. If the test is not directly asking about that it may underrepresent that construct.

If the test is measuring recidivism rates, but there are reasons beyond IPV that may lead to re-arrest and these are not accounted for within the testing instrument itself, the test may experience construct irrelevance.

## O2: RELIABILITY & MEASUREMENT

Two methods of using the term 'reliability':

Reliability as precision: more general notion of consistency of scores across instances of testing (standard errors, tolerance ratios, IRT)

Reliability Coefficient: More specific reference to the reliability of coefficients of classical test theory (score correlations between two equivalent forms)



## TYPES OF RELIABILITY COEFFICIENTS

When a testing instrument is analyzed for reliability, there are several methods of considering how consistent the testing results are within a population.

### O1: ALTERNATE FORM COEFFICIENTS

Derived from administering alternate forms in independent testing sessions

### O2: TEST-RETEST COEFFICIENTS

Derived from administering the same form at different times

### O3: INTERNAL CONSISTENCY COEFFICIENTS

Based on relationships among scores derived from individual items on a single administration - this often overlaps with construct validity

#### FOR IPV PROFESSIONALS:

Test-retest coefficients are common among IPV instruments and research, while alternate form coefficients are not, due to measurements of characteristics not scores.

## STANDARD ERROR OF MEASUREMENT

Standard Errors of Measurement (SEM) indicate the average error of measurement estimated over a population, and is an indicator of reduced consistency of scores.

### LARGE STANDARD ERROR OF MEASUREMENT

This will indicate that the scores in the testing instrument have low reliability and precision of results. This can inform that the scores might be suffering from construct irrelevance.

### CONFIDENCE INTERVALS OF SCORES

This is the measurement of error that often takes the form of  $\pm X$  where the number indicates how far the result might fluctuate from the mean.

### INDIVIDUAL VS GROUP MEANS AND SEM

SEM for individual scores are not appropriate measures of the precision of group averages - instead SEM for the estimates of group means should be used



### 03: PSYCHOLOGICAL TESTING

The results from tests and inventories used in assessment may help professionals to understand test takers more fully, and to develop more informed and accurate plans of action for the individual. **BUT CONTEXT IS IMPORTANT!**

Testing is not only about scores, but must include interviews, observations, and records from other sources to make the most thorough and accurate assessment of an individual.



### RESPONSIBILITIES OF PROFESSIONALS

Professionals, both psychological and other contexts, must be clear about the reasons a test taker is being assessed. These reasons guide the selection of appropriate tests, inventories, diagnostic procedures, and humanization of the individual themselves.



#### FAMILIARITY & MEANING

The professional must be familiar with the validation evidence for the intended uses of scores from the tests and inventories selected. It is important to verify that the construct being assessed has equivalent meaning across cultural contexts and individual situations.



#### UPDATES AND TRAINING

Professionals are responsible for guarding against reliance on test scores that are outdated. Interpreters of meaning must take the personal history of the individual and literature relevant to the test itself. All administrators of the test should be fully trained in their use, and must be able to provide the test taker with introductory information on the test itself.



#### RIGHTS OF TEST TAKERS

Professionals should share test scores and interpretations with the test taker when appropriate. Such information should be expressed in a language the test taker can understand. Test results should be kept confidential consistent with scientific, professional, legal, and ethical requirements.

### 04: OTHER DYNAMICS & CONCLUSIONS

While not always a direct connection to IPV work, discussion of educational testing, fairness in testing, program evaluation, policy studies, and accountability are useful to consider within research studies and instruments.



### EDUCATIONAL TESTING

Educational testing often is focused on achievement tests provided for K-12 and college students. While IPV professionals will not provide these kinds of tests to participants, these features can be important to understand within research and instruments developed for IPV.



#### EDUCATIONAL TESTING PURPOSES

Educational testing is designed to make inferences that inform teaching and learning that inform the individual or curricular level, or make inferences about outcomes for individuals and groups, or to inform decisions about individuals and their process through the program or educational setting.



#### MULTIPURPOSE TESTING

In educational settings, when a test is designed to serve multiple purposes, evidence of validity, reliability, and fairness should be provided for each intended use. It is incumbent on the user to provide evidence if they want to use a testing instrument outside of its intended purpose.



#### NORMING

Local norms should be developed when appropriate to support test users' intended interpretations.

#### FOR IPV PROFESSIONALS:

For IPV instruments, it is important to pay attention to how the scores were normed to similar populations. Sometimes there are differences in populations to consider to determine if the test has evidence of construct validity and is testing what it indicates to be testing.

## FAIRNESS IN TESTING

See ASBA, APA, NCEM, EDUAI Standards, Chapter 3

All individual cultural dynamics such as race, ethnicity, gender, language, culture, age, disability, and socioeconomic status can impact an individual's ability to respond to an assessment.

### ACCESSIBILITY & UNIVERSAL DESIGN

All test takers should have an unobstructed path to having their true attitude, ability, and behavior assessed. While context needs to be considered, the process should be similar for all test takers.

### RESPECTFUL AND CONSISTENT TREATMENT

All test takers should receive comparable treatment upholding individual dignity, agency, and respect for their individual humanity.

### PROMOTION OF VALID SCORE INTERPRETATIONS

All steps of the testing process should promote valid score interpretations for intended score users for the widest possible range of individuals and relevant subgroups of the intended population



### FOR IPV PROFESSIONALS:

Unless a testing instrument was specifically developed for a non-English speaking population, it is not recommended to use an instrument which has been translated into another language. Cultural norms present in United States English may not translate well to other languages or cultures, particularly those of a IPV nature.

## PROGRAM EVALUATION, POLICY, AND ACCOUNTABILITY

See ASBA, APA, NCEM, EDUAI Standards, Chapter 33



### CONTEXT & HIGH STAKES

Testing and assessments for IPV are high stakes for individuals and families. Contexts are critical to gain greater understanding.



### AGGREGATES & EFFECTIVENESS

Aggregates of scores and overall testing results can be used to determine the "effectiveness" of individual agencies and programs and may impact funding.



### TESTING VARIETY

Measuring things beyond IPV are very useful, and may add depth to understanding individuals and the contexts in which they live.



### COMMUNICATION

Important to consider how test results are explained to the community and referral sources, and coordinating approaches for individuals.



### ACCOUNTABILITY

It is important to focus not only on the accountability of the individual test taker, but also that of the professionals and agencies.

## Specific Considerations for Validation Processes



## Impacts of Time Upon Validity and Validation Evidence

## Categories of Time Impacts on **Validity**

### Time as it relates to Validity

When instruments are constructed, research conducted, and experiments analyzed, collecting validation evidence is a part of the process in determining how inferences might be made from the results. **Validation is an ongoing effort across time.**

The final part of this presentation will be addressing these issues and where they fit within the Conflict Tactics Scales, and how understanding specific analysis of research can benefit IPV professionals. **Throughout the common thread of time, the following topics will be covered here:**

1. Documenting Evidence Over Time
2. Response Shifts & Cognitive Appraisal
3. Shifting Focus of Validity for Test Use
4. Reconceptualizing Constructs Over Time

2

## 1. Documenting Evidence Over Time

Bowman, N. D., & Goodboy, A. K. (2020). Evolving considerations and empirical approaches to construct validity in communication science. *Annals of the International Communication Association*, 44(3), 219-234. <https://doi.org/10.1080/23808985.2020.1792791>

McEwan, B. (2020). Sampling and validity. *Annals of the International Communication Association*, 44(3), 235-247. <https://doi.org/10.1080/23808985.2020.1792793>

## Factor Analysis & Science

- Confirmatory Factor Analysis (CFA) results of measures are now commonly reported in publications to verify the factor their structure.
- Measurement development work can grow in parallel with statistical research.
- Bowman & Goodboy (2020) caution authors of “rampant respecification,” which is becoming common in publications and stifling measurement development.

4

Bowman & Goodboy (2020, p. 229)

## Sampling and **Validity**

- Readers and reviewers must always review manuscripts to consider the representativeness of samples.
- Crowdfunded sample sources provide new concerns for measures because they introduce the threat of bot responses.
- Thorough reporting of demographics in all studies is necessary to support measurement development.
- Measurement development is hindered when editors devalue replication work and non-statistically significant findings.

5

McEwan (2020, p. 239, 242-245)

## 2. Response Shifts and Cognitive Appraisal

McClimans, L., Bickenbach, J., Westman, M., Carlson, L., Wasserman, D., & Schwartz, C. (2013). Philosophical perspectives on response shift. *Quality of Life Research, 22*(7), 1873-8. <https://doi.org/10.1007/s1136-012-0300-x>

Sewatzky, R. (2019). Relating response shift and cognitive appraisal to measurement validation. *Quality of Life Research, 28*(10), 2633-2634. <https://doi.org/10.1007/s11136-019-02276-9>

## What are "Response Shifts"?

A concrete phenomenon where individuals shift their responses to certain questions over time, with an abstract quality related to how such individuals change their self-evaluation in ways that influence their responses. These might be due to "weak" or "strong" categories of self-evaluation.

## Why do these "Response Shifts" matter?

Changed responses can greatly impact validity evidence that seeks to measure constructs that are considered to be stable. While this concept has been explored within the context of "Quality of Life" research, there is applicability within other fields when exploring validity threats.

## "Weak" vs. "Strong" self-evaluations

"Weak" self-evaluations involve perspectives contingent on concrete circumstances and their impact on the self (time/age, environment, physical condition). "Strong" self-evaluations are considered to be "deeper" analysis based on the "vision of the good" which is a background picture that motivates perspectives of qualitative worth that informs decisions and choices.

7  
McClimans et al. (2012, p. 1872-1874)

## Illustrating Response Shift

### Strong Evaluations

Due to my injury I initially see this as an inability, but I see dignity in other hobbies so do not focus on my limitation and over time instead talk positively about other hobbies.

### Disability Paradigm

Assumptions are made about individuals with disabilities having poor responses to limitations, but can often have strong visions of the good.

### Social Constructivism

Within science, social facts play a role within scientific determinations of natural facts. Various scientific practices informed knowledge of DNA.

### Weak Evaluations

I enjoy running, but I tore a muscle and cannot run until I am healed. When asked about hobbies, I focus on my limitations and say I cannot run due to my injury.

### Vision of the Good

If my vision of the good is only tied to running, I will evaluate limitations as harms. If it is instead tied to a variety of activities, a limitation to one does not change overall standpoint.

### Scientific Realism

Things we might think do not determine reality - if DNA had not been discovered, it would still be a biological reality.

McClimans et al. (2012, p. 1872)

## Response Shift and Construct Validity

Measures often create instruments that assume respondents are weak evaluators. This assumption is why response shifts are important to analyze. Most cognitive based instruments are not designed to handle answers that result from consistency caused by a deeper vision of a "good life" through recalibration, reconceptualization, and reprioritization.

When respondents answer with strong self-evaluation, psychometric integrities can be threatened. Answers may not support hypotheses over time, facing researchers with the question whether the measure is invalid or the theory is incorrect.

McClimans et al. (2012, p. 1874)

## What about Response Shift as it relates to Self-Presentation?

Researchers often presuppose concepts of "scientific realism" that they are describing phenomenon that they believe is true regardless of whether it has been proven or not. However, in social science, education, and several other fields, "social constructivist" views suggest that **social and cultural contexts play a role in what counts as "fact."** This means that discoveries must be constructed through the mechanics of knowledge production that is reflexive, and that involve communication processes that may be more qualitative. **Researchers often communicate the "norm" within instruments,** which individual's self-representation might not reflect.

*This contributes directly to measurement bias.*

10 McJannet et al. (2012, p. 1874-1879)

## Cognitive Appraisal

As an important concept in understanding response shifts, cognitive appraisal is about the methods in which people's change in cognition lead to changing interpretations in how they might change responses to certain questions.

Response shift research on cognitive appraisal can be viewed as a form of measurement validity evidence with a **goal of arriving at justifiable inferences about the meaning of variability** in longitudinal change of measurement scores. Messick's description of the need to gain evidence on "response processes" feed directly into researching cognitive appraisal as a part of collecting validity evidence.

This means that response processes are mechanisms of what people, do, think, or feel when interacting with, and responding to, an item or task.

*This leads to a subjective element within instruments over time.*

11

Sawatzky (2019, p. 2633-2634)

## Response Processes are not Exclusively Cognitive

There are varied ways to potentially explain why people change responses to items.

- Emotional Responses / Dispositions
- Someone may change a response without changing their understanding of the construct
- Evidence of changes due to response shift require analysis of cognitive appraisal for the construct itself

12 Sawatzky (2019, p. 2633)

## Suggestions for addressing Response Shifts in Research

Use statistical methods for examining variability in residual differences between observed and expected measurement scores based on regression-based methods

View response shifts as a form of heterogeneity where some people may experience response shifts and others do not. Use Latent Class, Rasch, or IRT modeling

Test longitudinal measurement models within two or more latent models. Latent classes can be specified to represent hypothetical groups who are more homogenous and invariant over time.

13

Sawatzky (2019, p. 2634)

### 3. Shifting Focus of Validity for Test Use

Moss, P. A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy, & Practice*, 23(2), 236-251. <https://doi.org/10.1080/0969594X.2015.1072085>

We need a more complex theory of validity that can *shift focus* as needed between:

*intended* interpretations and uses of test scores, traditionally & *actual* interpretations and uses by professionals, in practice

15

Moss (2016, p. 236)

*A comprehensive validity theory in educational assessment needs to acknowledge the ways that education professionals —teachers, school and district leaders, and policy-makers— use tests and other evidence relevant to their students' learning in their ongoing work.*

Moss (2016, p. 247)

### Categories of Validity Evidence to Consider:

#### Working Knowledge

that supports meaningful interpretations of learning,

#### Coherent Organizational Norms

and routines that support learning at the classroom, school, district and professional organization levels.

#### Time & Resources

to seek patterns in available data, frame meaningful problems, explore explanations and solutions, and develop additional data reflecting appropriate timescales

#### Material Resources

that enable networking and sharing of knowledge

#### Flexible Access

to existing data and research to address problems, support decisions, and guide actions as needed

**Agentic Identities & Epistemic Culture** that seek explanations for learning outcomes in professional practice and organizational infrastructure

#### Attention to Cycles of Inquiry

that evaluate the outcomes of decisions and actions

#### Collaborative Roles & Relationships

for practitioners and researchers to enable mutual learning

#### Attention to Outcomes

that include learning by professionals, organizations, and students

17

Adapted from Moss (2016)

## Reconceptualizing Constructs Over Time

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31*(12), 1412–1427. <https://doi.org/10.1037/pas0000626>

## Why should we care about construct validity?

The quality of the real-world decisions that are made based on psychological measurements depends on the construct validity of the measurements on which they are based.

Practitioners must justify use of specific assessment procedures to third-party payers; must use psychological measures whose precision and efficiency are well supported by multiple types of empirical data

Progress in psychological science is critically dependent on measurement validity; the more validly and reliably we can measure experienced affects, behaviors, and cognitions, the more we can advance psychology and neuroscience

19

Clark & Watson (2019, p. 1413)

## Three components of construct validity:

*substantive*

conceptualization & development of an initial item pool

*structural*

item selection & psychometric evaluation

*external*

an ongoing process

20

Clark & Watson (2019)

*Good scale construction is an iterative process involving an initial cycle of preliminary measure development, data collection, and psychometric evaluation, followed by at least one additional cycle of revision of both measure and construct, data collection, psychometric evaluation, revision, and so forth.*

Clark & Watson (2019, p. 1420)

21

## Revising the target construct's conceptualization

*Commonly neglected, but extremely important!*

Scale developers too often assume that their initial conceptualization is entirely correct, considering only the measure as open to revision. However, it is critical to remain open to **rethinking** one's initial construct.

scale developers must  
**"listen to the data" not "make the data talk"**

22

Clark & Watson (2019, p. 1420)

## Main Takeaways

*Gathering validity evidence takes time*

*Test developers', test users', and test takers' perceptions of constructs can change over time*

*Timely reporting and sharing of data is critical*

*Best practices in validity & validation change over time*

## References

- Bowman, N. D., & Goodboy, A. K. (2020). Evolving considerations and empirical approaches to construct validity in communication science. *Annals of the International Communication Association, 44*(3), 219-234. <https://doi.org/10.1080/23808985.2020.1792791>
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31*(12), 1412-1427. <https://doi.org/10.1037/pas0000626>
- McClimans, L., Bickenbach, J., Westerman, M., Carlson, L., Wasserman, D., & Schwartz, C. (2013). Philosophical perspectives on response shift. *Quality of Life Research, 22*(7), 1871-8. <https://doi.org/10.1007/s11136-012-0300-x>
- McEwan, B. (2020). Sampling and validity. *Annals of the International Communication Association, 44*(3), 235-247. <https://doi.org/10.1080/23808985.2020.1792793>
- Moss, P. A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy, & Practice, 23*(2), 236-251. <https://doi.org/10.1080/0969594X.2015.1072085>
- Sawatzky, R. (2019). Relating response shift and cognitive appraisal to measurement validation. *Quality of Life Research, 28*(10), 2633-2634. <https://doi.org/10.1007/s11136-019-02276-9>

24

This template is free to use under [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).  
Presentation template by SlidesCarnival  
Plant illustrations from Kohler's Medicinal Plants in sarawagazette at BSM.

## Validity and the Conflict Tactics Scales



# Construct-Irrelevance and the Conflict Tactics Scales:

Threats to Validity Evidence & Suggestions for Mitigation



## Purpose:

Investigate the quality of validation evidence for the Conflict Tactics Scales (CTS) by examining ongoing arguments about validity threats to the instrument, and to discuss how the evolution of thinking on validity is reflected within this example.



### Context

Validity theory established within the educational and psychological measurement framework



### Direction

Apply best practices from these measurement frameworks to validation activities in the IPV field



## Foundations

The Conflict Tactics Scales (CTS) have a history of being controversial within the domestic violence / intimate partner violence professional setting, with multiple criticisms of its validity evidence. At the same time, the author and colleagues argue strongly against these criticisms and point to ongoing evidence collected on the scales.

To provide a context to this discussion, the elements of these arguments will be discussed briefly to frame this presentation and the validity evidence on the instrument itself.



## History of the CTS/CTS2

1973

First study reporting data using CTS published (Form A)



CTS findings article published with 2,143 couples (Form N) nationwide, as the sample size

1979

1985

Family Violence Resurvey uses Form R adding items for checking and burning to respond to criticisms



Revised Conflict Tactics Scales (CTS2) published incorporating new questions and further responding to criticisms of validity

1996





# Construct-Irrelevance

Refers to the degree to which test scores are affected by processes that are extraneous to the test's intended purpose. On a test designed to measure anxiety [physical aggression], a response bias to underreport one's anxiety [perpetrating or being harmed by physical aggression] might be considered a source of construct-irrelevant variance.

AERA, APA, NCME (2014, p. 12)



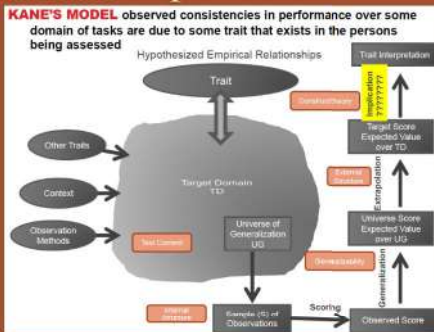
# Sources of Construct-Irrelevant Variance:

- Test items elicit varieties of responses other than those intended (AERA, APA, NCME (2014))
- Stereotype threat (Stricker & Ward (2008), Stricker & Ward (2004))
- Behavioral motives (Chapman & Gillespie (2019))
- Recall issues (Chapman & Gillespie (2019))
- Response styles (Damarin & Messick (1965))
- Items solved in ways that were not intended (AERA, APA, NCME (2014))
- Anxiety (French (1962); Powers (1988, 2001))
- Coaching (Messick (1981, 1982))

Such threats can result in *unfair tests with biased score interpretations*



# Kane's Interpretive Model (2006)



# Criticisms of CTS2 Validation Evidence

Chapman & Gillespie (2019)

"Specifically, respondents must be both willing to and able to admit to the act (i.e., they must be honest and they must be able to recall correctly" (p. 29).

"...the CTS2 includes behaviors that are socially unacceptable and punishable by law" (p. 29)

"others have argued that motives and meaning are vital for making sense of behavior, and thus determining whether the behavior constitutes IPV" (p. 31)

"factor analyses have frequently shown that the weapon items form a distinct factor from the other items" (p. 32)

"However, given the low coefficients reported for the sexual coercion scale in female samples and, to a lesser extent, the injury scale in male samples, these scales may benefit from being supplemented with interviews to provide context and further information pertaining to the items when used with the respective samples" (p. 30).

"It is not possible to make inferences about the concurrent validity of the CTS2" (p. 31)

"The authors acknowledge that more normative groups are needed to compare data obtained from respondents who fall outside of the reference sample for the CTS2" (p. 33)



# Validation Evidence Via Messick

"Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13).

MESSICK'S PROGRESSIVE MATRIX OF VALIDITY (1989)	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Value Interpretations	Social Consequences



# Our Dilemma

Even if test producers **design** a fair test, and test users **administer** the test fairly, some test takers may provide socially desirable responses rather than accurate responses, resulting in test scores **compromised by construct-irrelevance.**



# Social Desirability

Refers to respondents' *tendency to admit to socially desirable traits and behaviors and to deny socially undesirable ones.*

Respondents *present themselves in a positive light, independent of their actual attitudes and true behavior.*

Can also be conceptualized as *respondents' temporary social strategies coping with the different situational factors in surveys (e.g., presence of interviewer, topic of question).*

Krumpal (2013)



# Social Desirability is more than just Lying versus Honesty

Test-taker has defensive responses to items to avoid personal harm

Test-taker interprets all questions in positive frameworks



Test-taker provides responses they perceive are desired by the test-user



## Let's Test Construct Validity

Take a moment to relax. Clear your mind.  
I'm going to give you a quick one-question quiz.  
This quiz measures the accuracy of your  
memory, and your overall intelligence depending  
on if you get this correct or not.



**How many acorns were in the slide  
discussing Social Desirability is more  
than just lying vs. honesty (using a  
Venn Diagram)?**

**Remember - your score will determine  
your intelligence!**



## Examples of Socially Desirable Responses

Items asking the test taker to self-report...



their own socially  
undesirable  
attitudes

- anti-Semitism
- xenophobia
- chauvinism



the number of times  
engaging in criminal and/or  
illegal activity

- shoplifting
- driving drunk
- consuming illegal drugs



experiences of criminal  
victimization

- battered by a partner
- scammed by fraud
- sexually victimized



## Construct Validity Via Messick & Cronbach

*"Cronbach (1971) distinguishes between using a test to describe a person and using it to make decisions about the person." Messick (1989)*

*"Cronbach (1984) referred to content, criterion-oriented, and construct validation as 'methods of inquiry.' This usage serves to highlight the fact that test validation is a process of inquiry into the adequacy and appropriateness of interpretations and actions based on test scores." Messick (1989)*



# Messick's Message

Using a test to describe a person requires evidence of content or construct validity:

Content	Criterion	Predictive	Concurrent	Construct
"... is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are drawn." (Messick, 1995, p. 18)	"... is evaluated by comparing the test scores with one or more external variables considered to provide a direct measure of the characteristic or condition." (Messick, 1995, p. 18)	"... indicates the extent to which an individual's future level on the criterion is predicted from prior test performance." (Messick, 1995, p. 18)	"... indicates the extent to which the test scores estimate an individual's present standing on the criterion." (Messick, 1995, p. 18)	"... is evaluated by investigating what qualities a test measures, and to the degree to which certain explanatory variables account for performance on the test." (Messick, 1995, p. 18)
Professional judgment	Correlations and Regressions	Correlations and Regressions	Correlations and Regressions	The degree of fit of the information with the theoretical scheme underlying score interpretation is empirically evaluated
Reliability Messick, 1995, p. 23	Validity, 1995 Messick, 1995, p. 21			Reliability, 1995, p. 22



Using a test to describe a person as a perpetrator of intimate partner violence is:

- High stakes
  - Potential legal consequences
  - Potential social consequences
- Needs very strong construct validity evidence

# Construct-Irrelevance & Validity (Kane)

*"In most assessment contexts, the question is not whether an assessment measures the trait or some alternate variable but rather the extent to which the assessment measures the trait of interest and is not overly influenced by sources of irrelevant variance."*  
Kane & Bridgeman (2017)

*"Test scores that consistently underestimate or overestimate the variable of interest for a subgroup for any reason are said to be biased, and standardization tends to control this kind of bias, whether it is inadvertent or intentional."*  
Kane & Bridgeman (2017)

*"A prime threat to fair and valid interpretation of test scores comes from aspects of the test or testing process that may produce **construct-irrelevant variance** in scores that systematically lowers or raises scores for identifiable groups of test takers and results in **inappropriate score interpretations** for intended uses."*

AERA, APA, NCME (2014, p. 54)

**Standards for Educational & Psychological Testing: Construct Irrelevance**

Standard 3.2: Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.

Standard 3.3: Those responsible for test development should include relevant subgroups in validity, reliability, precision, and other preliminary studies used when constructing the test.

Standard 3.3a: Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from those subgroups.

Standard 3.3b: In testing individuals for diagnostic and/or special program placement purposes, test users should not use test scores as the sole indicators to characterize an individual's functioning, competence, attitudes, and/or predispositions. Instead, multiple sources of information should be used, alternative explanations for test performance should be considered, and the professional judgment of someone familiar with the tests should be brought to bear on the decisions.

Standard 3.3c: Reports of group differences in test performance should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of the differences. If appropriate contextual information is not available, users should be cautioned against misinterpretation.

AERA, APA, NCME (2014)

## What do the social scientists say?



### McEwan (2020)

All evidence cannot come from solely undergraduate samples

### Bowman and Goodboy (2020)

A measure must be used consistently across studies



### Swartzky (2019)

The risk of response shift

"The difference in the response processes that we are examining refers to the effects of gender socialization on the ways of symbolising violence, and its differential legitimization as a function of sex."

Álvarez (2014, p. 20)



*"Others have argued that motives and meaning are vital for making sense of behavior, and thus determining whether the behavior constitutes IPV"*  
Chapman & Gillespie (2019)

*"Factor analyses have frequently shown that the weapon items form a distinct factor from the other items"*  
Chapman & Gillespie (2019)

*"...Respondents must be both willing to and able to admit to the act (i.e. they must be honest and they must be able to recall correctly) ... the CTS2 includes behaviors that are socially unacceptable and punishable by law."*  
Chaoman & Gillespie (2019)

*"If the scales assessing violence in dating couples are attempting to compare the violence suffered from and perpetrated by women and men, then they must necessarily ensure the equivalency of the scores derived from the responses of both groups."*  
Álvarez (2014, p. 19)



## Kane on Validity

"Validation has a contingent character; the **evidence required to justify a proposed interpretation** or use depends on the proposed interpretation or use" (Kane, 2006, p. 60)

"Validation focuses on interpretations, or meanings, and on decisions, **which reflect values and consequences**"  
(Kane, 2006, p. 18)

"The evaluation of test score uses requires an evaluation of the consequences of the proposed uses, and **negative consequences can render a score use unacceptable.**" (Kane, 2014, p. 46)







## Dilemma Resolution:

What do we propose to limit the potential for social desirability to create construct-irrelevant responses to instrument items?



## Proposed Enhancements to the Standards

 Revised Standard 3.2	Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, social, or other characteristics.
 New Standard 3.6b	Where credible evidence indicates that test scores <b>may differ in authenticity</b> for relevant subgroups in the intended examinee population, test developers are responsible for examining supplemental evidence to support the score.
 New Standard 3.9b	When construct-irrelevance is a likely result of the test due to <b>serious potential consequences</b> such as physical danger or legal charges, test developers and/or users should, where feasible, <b>remove perceived threat</b> to the test taker.
 New Standard 3.10b	When testing individuals in which <b>physical harm is a potential consequence</b> , test users should not use test scores as the sole indicators. Instead, <b>multiple sources of information</b> should be used, including <b>biometric data</b> where possible.

## Proposed Enhancements to the Standards

 Revised Standard 4.12	"Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications." In this section, <b>include exploratory component of test development</b> to account for <b>possible construct-irrelevant dynamics</b> that may exist as a part of the target domain. Include within instrument plan relevant evidence gathering to this effect.
 New Standard 4.13b	Test developers <b>must provide credible evidence of mitigation of any social desirability components</b> within the testing instrument. This may include specific social desirability measurement scales, or scales that focus on potential deception in a test taker's responses.

## References Page One

- AERA, APA, NCME. (2014). *Standards for educational and psychological testing*. AERA.
- Álvarez, C. D. (2014). What do the dating violence scales measure? *Procedia-Social and Behavioral Sciences*, 161, 18-23. <https://doi.org/10.1016/j.sbspro.2014.12.004>
- Chapman, H., & Gillespie, S. M. (2019). The Revised Conflict Tactics Scales (CTS2): A review of the properties, reliability, and validity of the CTS2 as a measure of partner abuse in community and clinical samples. *Aggression and Violent Behavior*, 44, 27-35. <https://doi.org/10.1016/j.avb.2018.10.006>
- Damarin, F., & Messick, S. (1965). *Response styles and personality variables: A theoretical integration of multivariate research (Research Bulletin No. RB-65-10)*. Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.1965.tb00967.x>
- French, J. W. (1962). Effect of anxiety on verbal and mathematical examination scores. *Educational and Psychological Measurement*, 22(3), 553-564. <https://doi.org/10.1177/001316446202200313>
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). American Council on Education/Praeger.
- Kane M., Bridgeman B. (2017) Research on validity theory and practice at ETS. In R. Bennett & M. von Davier (Eds.) *Advancing human assessment: Methodology of educational measurement and assessment*. Springer. [https://doi.org/10.1007/978-3-319-58689-2\\_16](https://doi.org/10.1007/978-3-319-58689-2_16)

## References Page Two

- Kothgauer, O. D., & Felnhofner, A. (2020). Does virtual reality help to cut the Gordian knot between ecological validity and experimental control? *Annals of the International Communication Association*, 44(3), 210-218. <https://doi.org/10.1080/23808985.2020.1792790>
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Qual Quant*, 47, 2025-2047. <https://doi.org/10.1007/s11135-011-9640-9>
- Messick, S. (1981). The controversy over coaching: Issues of effectiveness and equity. *New Directions for Testing and Measurement*, 11, 21-53.
- Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing policy. *Educational Psychologist*, 17(2), 69-91. <https://doi.org/10.1080/00461528209529246>
- Messick, S. (1989). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). American Council on Education/Praeger.
- Powers, D. E. (1988). Incidence, correlates, and possible causes of test anxiety in graduate admissions testing. *Advances in Personality Assessment*, 7, 49-75.

## References Page Three

- Powers, D. E. (2001). Test anxiety and test performance: Comparing paper-based and computer-adaptive versions of the Graduate Record Examinations® (GRE) General Test. *Journal of Educational Computing Research, 24*(3), 245–273. <https://doi.org/10.2190/680W66CR-QRP7-CL1F>
- Straus, M. A. (1979). Measuring intrafamily conflict and violence: The conflict tactics (CT) scales. *Journal of Marriage and the Family, 41*(1), 75–88. <https://doi.org/10.2307/351733>
- Straus, M., Hamby, S. L., Boney-McCoy, S., & Sugarman, D. B. (1996). Revised conflict tactics scales (CTS2): Development and preliminary psychometric data. *Journal of Family Issues, 17*(3), 283–316. <https://doi.org/10.1037/102126-000>
- Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34*(4), 665–693. <https://doi.org/10.1111/j.1559-1816.2004.tb02564.x>
- Stricker, L. J., & Ward, W. C. (2008). Stereotype threat in applied settings re-examined: A reply. *Journal of Applied Social Psychology, 38*(6), 1656–1663.



## Information on Validity within Research: A Review of the Conflict Tactics Scales

by Christopher Hall, MSW  
University of North Carolina at Greensboro  
Ph.D. Student, Educational Research Methodology

