education sciences

MDPI

Commentary

Measurement Fragmentation in Educational and Psychological Research: The TTCT as a Case Example

Kyung Hee Kim

Department of School Psychology and Counselor Education, Graduate School of Education, William & Mary, Williamsburg, VA 23187, USA; k@7musesinstitute.org

Abstract

Measurement is fundamental to scientific inference, yet educational and psychological research often relies on instruments that are shortened, partially administered, or otherwise modified without evidence that the resulting scores remain valid. This article examines measurement fragmentation, defined as the use of selected parts, abbreviated forms, or derivative versions of an instrument as though they were interchangeable with the validated form. The Torrance Tests of Creative Thinking (TTCT) are used as a case example because their strong full-version psychometric foundation and frequent fragmented use in published research make the consequences of non-equivalent measurement unusually visible. The full TTCT is supported by stronger reliability and validity evidence than fragmented TTCT uses, whereas derivative forms such as the Abbreviated Torrance Test for Adults have weaker support. When these non-equivalent forms are treated under a single instrument label, apparent theoretical inconsistency and attenuated meta-analytic estimates may reflect measurement artifacts rather than properties of creativity itself. The article concludes that stronger educational and psychological research requires explicit reporting of test versions, version-specific reliability and validity evidence, and closer editorial scrutiny of altered measurement procedures.

Keywords: measurement fragmentation; test version; reliability; validity; cumulative evidence synthesis; Torrance Tests of Creative Thinking (TTCT)

1. Introduction

Scientific conclusions in educational and psychological research depend fundamentally on the quality of measurement. Without adequate measurement, even rigorous study designs and transparent analytic practices cannot ensure sound conclusions, because weak measurement distorts score meaning and undermines the inferences drawn from research findings (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME] (AERA, APA, & NCME, 2014); Higgins et al., 2026; Stefana et al., 2025). The *Standards for Educational and Psychological Testing* make this point explicit by grounding validity in the interpretation and use of scores from a specific test version administered under specified conditions, rather than in the instrument name alone (AERA, APA, & NCME, 2014). Two psychometric concepts—reliability and validity—are central to measurement quality. Reliability concerns the consistency or precision of scores across items, raters, occasions, or forms, whereas validity concerns the extent to which those scores support the interpretations and uses intended by the researcher (AERA, APA, & NCME, 2014). Reliability is therefore necessary for validity, because scores burdened by substantial measurement error cannot sustain

Academic Editor: Luca Tateo

Received: 22 January 2026
Revised: 1 May 2026
Accepted: 3 May 2026
Published: 7 May 2026

stable interpretation, but it is not sufficient; scores may be highly consistent without accurately representing the intended construct (AERA, APA, & NCME, 2014; Higgins et al., 2026; Stefana et al., 2025).

Yet researchers often treat an instrument's name as if it guarantees validity across uses, even when the administered version departs materially from the validated form. This assumption is not psychometrically justified. Recent work has stressed that validity should not be treated as a blanket property of a test in the abstract but as something tied to the measurement outcomes produced in particular uses and supported by context-specific evidence (Higgins et al., 2026). Likewise, best-practice guidance for scale development and validation emphasizes that measurement quality depends on clear construct definition, justified instrument selection, and explicit evidence for the scale actually used, rather than on convenience, familiarity, or prior reputation alone (Stefana et al., 2025).

This problem becomes especially acute when researchers administer only selected parts of an instrument, rely on abbreviated or derivative versions, or apply norms and validity evidence from a full, validated test to a modified form. Such practices are referred to herein as *measurement fragmentation*. Measurement fragmentation may improve efficiency, but it also risks altering construct representation, narrowing behavioral sampling, increasing measurement error, and weakening comparability with findings based on the validated version. General psychometric research shows that shortened tests and extracted subscores can reduce construct coverage, weaken reliability, and attenuate associations with external criteria unless the modified form is independently validated (Credé et al., 2012; Sijtsma & Emons, 2011). Partial, abbreviated, or derivative forms therefore should not be presumed to inherit the evidentiary support of the parent instrument automatically.

This concern is consistent with broader critiques of contemporary research practice. Paredes and Carré (2024) observe that reform efforts in psychological science have focused heavily on statistical procedures—such as preregistration, transparency, and replication—while often neglecting the earlier stages of measurement, where constructs are defined, instruments are selected, and test versions are chosen. The *Standards for Educational and Psychological Testing* similarly make clear that meaningful changes to test content, structure, or administration require their own support, because altered versions cannot simply be assumed to sustain the same score interpretations as the validated form (AERA, APA, & NCME, 2014; Higgins et al., 2026).

2. The Torrance Tests of Creative Thinking as a Case Example of Measurement Fragmentation

The TTCT serves as the focal case in this article. It is available in more than 35 languages and has been described as the most widely used and researched creativity test (Kim, 2021). Its broad use, strong full-version psychometric record, and frequent use in fragmented forms in published research make it an especially revealing illustration of measurement fragmentation. The concern here is not that the full TTCT lacks psychometric support, but that partial and derivative uses are often interpreted as retaining the evidentiary support of the validated instrument. Poor reporting or altered use should not be conflated with infirmity of the parent measure itself (Murphy & Hall, 2024).

Developed by Torrance (1966) to identify and support creative potential, the TTCT assesses multiple dimensions of creative thinking through two complementary batteries. The TTCT–Figural, the more widely used and more predictive form, includes Picture Construction, Picture Completion, and Lines or Circles, yielding the subscores Fluency, Originality, Elaboration, Abstractness of Titles, Resistance to Premature Closure, and the 13 Creative Strengths. The TTCT–Verbal includes Asking, Guessing Causes, Guessing Consequences, Product Improvement, Unusual Uses, and Just Suppose, yielding Fluency,

Flexibility, and Originality. Together, these tasks reflect a broad and multifaceted conception of creative thinking (Kim, 2011, 2017, 2021).

The strongest reliability and validity evidence for the TTCT pertains to administrations that use the complete set of activities within a battery and follow the test's standardized scoring and normative procedures (Kim, 2006, 2008, 2017, 2018). For the Figural composite scores, KR-21 internal consistency coefficients typically range from .78 to .91 across grades and ages, and interrater reliability for Figural subscores and the Creativity Index generally falls between .90 and .99 when standardized scoring procedures are followed (Scholastic Testing Service, 2017). Test–retest reliability ranges from .50 to .93, reflecting the influence of motivational conditions on performance stability (Kim, 2006).

The full TTCT is also supported by substantial evidence of validity. Meta-analytic findings based on full TTCT use show that TTCT scores predict real-world creative achievement, $r = .30$ (Kim, 2018). Predictive validity strengthens when TTCT–Figural and TTCT–Verbal core subscores are combined ($r = .46$), and the complete six-score TTCT-Figural predicts creative achievement more strongly still ($r = .53$; Kim, 2018). Cross-sectional studies further show that full TTCT scores distinguish individuals with documented creative accomplishments from those without (Chávez-Eakle et al., 2006; Wechsler, 2006). Longitudinal evidence is especially important: Childhood TTCT scores predict personal or public creative accomplishment four decades later (Cramond et al., 2005) and five decades later (Runco et al., 2010). These findings indicate that the complete TTCT supports meaningful inferences about creative thinking and long-term creative accomplishment.

Despite this psychometric foundation, complete administrations have become relatively uncommon. Among 523 published TTCT studies from 1968 to 2018, only 57.55% analyzed TTCT scores; and among those 301 studies, only 51.83% used the full TTCT. By contrast, 36.54% relied on only one or two TTCT activities, and 11.63% used the Abbreviated Torrance Test for Adults (ATTA), a derivative form introduced by Goff and Torrance (2002). Kim (2018) reported its use in that proportion of the TTCT studies reviewed. These practices illustrate measurement fragmentation: Modified, shortened, or derivative forms are treated as though they were interchangeable with the validated instrument.

The TTCT's design makes it especially vulnerable to fragmentation. Each activity assesses distinct creative-thinking processes—such as perceptual openness, tolerance for ambiguity, abstraction, cognitive flexibility, and originality—and these processes are not interchangeable. The meaning of TTCT scores emerges from combined sampling across tasks rather than from any single activity alone. Extracting Fluency, Originality, or Elaboration scores from one task narrows behavioral sampling to a limited slice of creative thinking, increases the proportion of measurement error, and weakens reliability. As a result, fragmented administrations can attenuate relations with external criteria even when the underlying construct has not changed.

A brief comparison makes this problem concrete. Full TTCT administrations have shown predictive validity coefficients ranging from $r = .30$ to $r = .53$, whereas partial TTCT administrations predict creative achievement much more weakly ($r = .18$), and ATTA scores are weaker still ($r = .13$–$.15$; Kim, 2018; Said-Metwaly et al., 2022). If these non-equivalent forms are all treated simply as "the TTCT," the resulting literature can appear theoretically inconsistent. Yet that inconsistency is better understood as a measurement artifact: Weaker effects arise when abbreviated or fragmented forms reduce construct coverage and increase measurement error, not because the phenomenon under study is itself inconsistent.

Partial administrations and abbreviated forms are often motivated by genuine practical constraints, such as limited testing time, participant burden, and the realities of large-scale or applied research. Those constraints, however, do not resolve the psychometric problem. A shorter or altered procedure may be expedient, but expedience does not establish

equivalence to the validated full TTCT. When researchers administer only part of the test or substitute a derivative form, they create a materially different measurement procedure, and the resulting scores require their own evidentiary support under the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

Those standards are clear on the relevant principle: Validity applies to interpretations of scores from the specific test version used; altered or alternate versions must provide their own evidence of reliability, validity, and interpretive comparability; and partial-test scores cannot simply inherit the support of the full instrument (AERA, APA, & NCME, 2014). Published partial TTCT studies have generally not provided the version-specific reliability and validity evidence required by these standards. Instead, validity evidence from the full TTCT has often been extended to fragmented forms without adequate justification.

A similar problem appears in the widespread use of the ATTA, which has often been described as a standardized short version of the TTCT. The available documentation, however, does not establish the ATTA as a validated equivalent of the full Figural or Verbal TTCT. It derives from the Brief Demonstrator Form of the TTCT (Goff & Torrance, 2000) rather than from the validated full batteries, and according to Bonnie Cramond, Torrance did not participate in its development (B. Cramond, personal communication, 16 March 2024). The ATTA manual does not adequately document the rationale for selecting its three tasks, independent evidence of reliability or validity, demographic or methodological details of norming samples, or evidence of comparability with the TTCT (Athanasou & Bugbee, 2007; Kim, 2018). These omissions are directly relevant to standards for version differences, norming procedures, and score comparability (AERA, APA, & NCME, 2014, pp. 86, 102, 126). Empirically, ATTA scores show weak predictive validity (r = .13–.15), well below the full TTCT (Kim, 2018; Said-Metwaly et al., 2022).

In TTCT research, measurement fragmentation occurs through two main pathways: partial administrations that fail to sample the full range of creative-thinking behaviors, and unvalidated substitutions such as the ATTA. Both practices weaken reliability and validity, depart from professional testing standards, and distort the interpretation of creativity research. When different research groups unknowingly use different versions of "the TTCT," discrepancies in findings can be mistaken for theoretical disagreement rather than recognized as consequences of non-equivalent measurement.

3. Implications and Recommendations for Educational and Psychological Research

The TTCT case illustrates a broader methodological problem in educational and psychological research. Its structure and psychometric history make departures from the validated form relatively easy to detect, but similar fragmentation may be less visible in other areas of assessment (AERA, APA, & NCME, 2014; Higgins et al., 2026). When instrument modifications are unreported, poorly justified, or treated as inconsequential, score meanings can shift, comparability across studies can erode, and theoretical claims can be built on unstable measurement foundations (Paredes & Carré, 2024; Sijtsma & Emons, 2011; Sinharay et al., 2011).

Concerns about shortened, partial, or otherwise modified instruments have been raised across multiple domains, including educational measurement, personality and selection testing, psychosomatic assessment, and psychology and psychiatry (AERA, APA, & NCME, 2014; Credé et al., 2012; Kruyen et al., 2012, 2013; Sijtsma & Emons, 2011; Sinharay et al., 2011). Taken together, this body of work suggests that shortened tests and extracted subscores generally have narrower construct coverage, weaker reliability, altered internal structure, and weaker associations with external criteria unless the modified form is independently validated. Measurement fragmentation therefore undermines not

only individual studies but also comparability across studies, replication, and cumulative evidence synthesis.

These consequences extend to meta-analysis. When full TTCT administrations, partial TTCT uses, and ATTA scores are pooled under a single label, the resulting estimate no longer summarizes one psychometrically coherent instrument. Instead, stronger effects from the validated full TTCT are blended with weaker effects from fragmented and derivative forms, attenuating pooled estimates and increasing the risk that theoretical conclusions will reflect measurement heterogeneity rather than the construct itself (Kim, 2018; Said-Metwaly et al., 2022). In this way, fragmentation becomes not merely a technical issue, but a threat to cumulative science.

Improving research quality, therefore, requires attention not only to analytic procedures but also to the earlier stages of measurement, where constructs are defined, instruments are selected, versions are chosen, and administration fidelity is determined. At a minimum, the exact test version used should be reported explicitly. Task composition, scoring procedures, administration conditions, and any deviations from standard protocol should be documented so that readers can determine whether findings are comparable across studies (Paredes & Carré, 2024).

Reliability and validity evidence must also correspond to the version actually administered. Psychometric support for a full test cannot simply be transferred to partial administrations, extracted subscores, or derivative forms. As the *Standards for Educational and Psychological Testing* make clear, evidentiary support is tied to the interpretations of scores produced by a specific version under specific conditions (AERA, APA, & NCME, 2014). Single tasks and extracted subscores should not be interpreted as stand-alone measures unless they have been independently validated for that purpose. Subscores detached from a multi-task instrument often do not retain sufficient reliability, construct representation, or interpretive value (Sinharay et al., 2011; Sijtsma & Emons, 2011). Short forms should be validated as short forms; equivalence with the full version cannot be assumed.

Researchers who shorten, restructure, or substitute instruments should therefore explain what was changed, why the change was made, and what evidence supports the intended interpretation of the resulting scores. Journals and reviewers should evaluate such measurement decisions as carefully as analytic ones. Convenience or widespread use does not establish psychometric equivalence.

Measurement transparency should likewise be treated as part of open science. Reporting test versions, task changes, scoring rules, and administration details is essential for evaluating construct fidelity, interpreting findings, and enabling meaningful replication (Paredes & Carré, 2024). Scientific rigor depends not only on stronger analyses but also on stronger measurement decisions.

4. Conclusions

Measurement fragmentation threatens the validity, comparability, and cumulative value of educational and psychological research. The TTCT case illustrates how partial administrations and unvalidated derivatives can weaken reliability, attenuate observed associations, and create the appearance of theoretical inconsistency when the underlying problem is methodological rather than substantive. These distortions arise not because the full TTCT lacks psychometric support, but because altered forms are often treated as though they were interchangeable with the validated instrument.

More broadly, reforms that focus solely on statistical techniques cannot remedy problems created by weak or insufficiently documented measurement. Stronger research requires transparent reporting of test versions, careful attention to administration fidelity, and evidence of reliability and validity for the specific version used. Without such safeguards,

empirical conclusions may rest on scores whose meaning has shifted without acknowledgment. Addressing measurement fragmentation is therefore essential for producing findings that are interpretable, replicable, and theoretically meaningful.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable. No new data were created.

Conflicts of Interest: The author declares no conflict of interest.

References

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Athanasou, J. A., & Bugbee, A. C. (2007). Review of the Abbreviated Torrance Test for Adults. In B. S. Plake, J. C. Impara, & R. A. Spies (Eds.), *The fifteenth mental measurements yearbook*. Buros Institute of Mental Measurements.

Chávez-Eakle, R. A., Lara, M. D. C., & Cruz-Fuentes, C. (2006). Personality: A possible bridge between creativity and psychopathology? *Creativity Research Journal*, *18*(1), 27–38. [CrossRef]

Cramond, B., Matthews-Morgan, J., Bandalos, D., & Zuo, L. (2005). A report on the 40-year follow-up of the Torrance Tests of Creative Thinking: Alive and well in the new millennium. *Gifted Child Quarterly*, *49*(4), 283–291. [CrossRef]

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology*, *102*(4), 874–888. [CrossRef] [PubMed]

Goff, K., & Torrance, E. P. (2000). *Brief demonstrator form of the Torrance Test of Creative Thinking: Training/teaching manual for adults with norms-technical data*. Scholastic Testing Services.

Goff, K., & Torrance, E. P. (2002). *Abbreviated Torrance Test for Adults*. Scholastic Testing Services.

Higgins, W. C., Kaplan, D. M., Gillett, A. J., Sutton, J., & Ross, R. M. (2026). Rethinking psychological measurement: Validity potential versus realised validity. *Studies in History and Philosophy of Science*, *116*, 102123. [CrossRef] [PubMed]

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity Research Journal*, *18*(1), 3–14. [CrossRef] [PubMed]

Kim, K. H. (2008). Meta-analyses of the relationship of creative achievement to both IQ and divergent thinking test scores. *Journal of Creative Behavior*, *42*(2), 106–130. [CrossRef]

Kim, K. H. (2011). The creativity crisis: The decrease in creative thinking scores on the Torrance Tests of Creative Thinking. *Creativity Research Journal*, *23*(4), 285–295. [CrossRef]

Kim, K. H. (2017). The Torrance Tests of Creative Thinking: Figural or verbal? Which one should be used? *Creativity. Theories–Research–Applications*, *4*(2), 302–321. [CrossRef]

Kim, K. H. (2018, August 9–12). *The Torrance Tests of Creative Thinking: Figural and verbal—Domain-general or domain-specific creativity?* [Paper presentation]. 126th Annual Convention of the American Psychological Association, San Francisco, CA, USA.

Kim, K. H. (2021). Creativity crisis update: America follows Asia in pursuing high test scores over learning. *Roeper Review*, *43*(1), 21–41. [CrossRef]

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, *12*(4), 321–344. [CrossRef]

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, *13*(3), 223–248. [CrossRef]

Murphy, B. A., & Hall, J. A. (2024). How a strong measurement validity review can go astray: A look at Higgins et al. (2024) and recommendations for future measurement-focused reviews. *Clinical Psychology Review*, *114*, 102506. [CrossRef] [PubMed]

Paredes, J., & Carré, D. (2024). Looking for a broader mindset in psychometrics: The case for more participatory measurement practices. *Frontiers in Psychology*, *15*, 1389640. [CrossRef] [PubMed]

Runco, M. A., Millar, G., Acar, S., & Cramond, B. (2010). Torrance Tests of Creative Thinking as predictors of personal and public achievement: A fifty-year follow-up. *Creativity Research Journal*, *22*(4), 361–368. [CrossRef]

Said-Metwaly, S., Taylor, C. L., Camarda, A., & Barbot, B. (2022). Divergent thinking and creative achievement: How strong is the link? *Psychology of Aesthetics, Creativity, and the Arts*, *18*(5), 869–881. [CrossRef]

Scholastic Testing Service. (2017). *Torrance Tests of Creative Thinking: Norms-technical manual: Figural forms A and B*. Scholastic Testing Service.

Sijtsma, K., & Emons, W. H. M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, *70*(6), 565–572. [CrossRef] [PubMed]

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29–40. [CrossRef]

Stefana, A., Damiani, S., Granziol, U., Provenzani, U., Solmi, M., Youngstrom, E. A., & Fusar-Poli, P. (2025). Psychological, psychiatric, and behavioral sciences measurement scales: Best practice guidelines for their development and validation. *Frontiers in Psychology*, *15*, 1494261. [CrossRef] [PubMed]

Torrance, E. P. (1966). *Torrance Tests of Creative Thinking: Norms technical manual (Research edition)*. Personnel Press.

Wechsler, S. (2006). Validity of the Torrance Tests of Creative Thinking to the Brazilian culture. *Creativity Research Journal*, *18*(1), 15–25. [CrossRef] [PubMed]