THE NUGGET REPRINT

Wargaming and Validity

Author: David Burden, david.burden21@bathspa.ac.uk

This article originally appeared in The Nugget, Issue 349, Dec 2022, the Journal of the Wargame Developments Group. I wrote it to start to get my head around the idea presented, and comments are welcome as I try and work it up into a more polished article/paper for more formal publication.

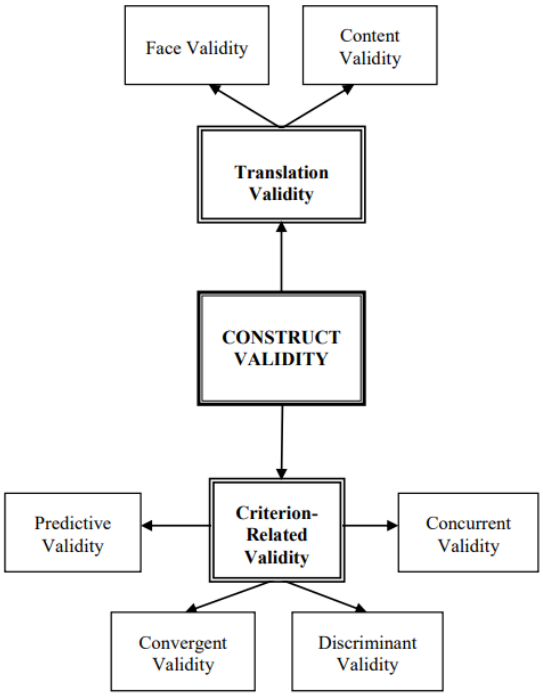
A challenge from a client recently got me thinking (perhaps a bit late) about how wargames should be validated. Graham Longley-Brown (in Successful Professional Wargames) notes that wargamers (and system engineers) distinguish between verification (does it follow the design) and validation (is it true to reality), and it was certainly the latter definition that I was interested in. Graham notes that other communities (including the MOD and his book!) exchange the definitions. A recent talk by Kate Kuehn on "Valid and Meaningful Assessment of Wargames" for the King's College Wargaming Network (Kuehn, 2022) also seemed to use validation more in the "does it meet the design" sense. A Working Group led by Stephen Downes-Martin at the MORS Wargaming Special Meeting in October 2017 on the "Validity and Utility of Wargaming" (Downes-Martin , 2017) had lots of interesting discussion, but didn't really provide anything that I felt was actionable in my research context.

A common model that is often used (e.g. Jiménez-Buedo & Miller, 2010) is that of Internal Validity (is there cause and effect, have other factors been excluded?) and External Validity (can this be applied to other situations?). McDermott (2011:27) notes that the relative importance of these measures of validity changes between disciplines. Whilst the idea of External Validity is highly useful as it is measuring how well we can apply the wargame and its findings to other comparable situations, Internal Validity seems more of a challenge as it suggests trying to identify that the model is measuring potentially all of the correct cause-and-effect relationships within an engagement – achievable at a lower level perhaps, but unlikely at higher levels.

In a 2020 talk to UK Fight Club (Mason, 2020) Roger Mason talked about the fact that the scientific validation of wargames was impossible, but that they provided an opportunity to explore and understand possible future outcomes – which chimes with the discussion after John Curry's talk at COW2022 about wargames providing contour lines on a map of the future. But I was after something a bit more "precise". Roger, though, also talked about face, construct and content validation. A quick scurry around the Internet found that these, and other associated terms like criterion validity and predictive validity seemed to suffer from even a worse problem than validity and verification, with everyone having their own (often conflicting) definitions and fancy diagrams.

In the end I found what I think is a really good paper and model by Ellen Drost on "Validity and Reliability in Social Science Research" (Drost, 2011) which seemed to lay out everything quite clearly, and I think I could make a pretty good case for wargaming being a branch of social science research! It's also got 2111 citations on Google Scholar – so at least others seem to see merit in it as well.

Drost's core diagram is this:



Based on her thoughts, each of these elements can be defined as follows:

- **Construct validity** – How well has the concept, idea, or behaviour – that is the construct – been transformed into an "operationalisation" across all measures.
- **Translation validity -** Does the operationalisation reflect the understood meaning of the construct – a more *a priori* assessment?
- **Face validity** - A subjective "on face value" judgment on the operationalisation of a construct – and as such is often seen as a weak form of construct validity.
- **Content validity –** Is the content appropriate to the requirement, often involving the judgment of Suitable Qualified and Experienced Personnel (SQEP).
- **Criterion-related validity –** The degree of correspondence between a test measure and one or more external real-world metrics – a more *a posteriori* assessment?
- **Concurrent validity** - When the real-world metric exists at the same time as the test measure, i.e. the ability of a test to match events in the present.
- **Predictive validity -** When the real-world metric occurs in the future.
- **Convergent validity** – Measuring the same thing in different ways should give the same or similar results (i.e. are the tests measuring what we think they should test)
- **Discriminant validity** – Measuring something different but in the same way as for Convergent validity should yield no or low correlation results (i.e. are the tests not measuring something completely different).

So, putting this in the context of a wargame, what might this give us?

- **Construct validity** – A top level "measure" of how well the wargame reflects a military reality.
- **Translation validity -** Standing alone, has the essence of the military reality been captured by the wargame (before we start comparing it to other data)?
- **Face validity** – On a face-value basis, does the wargame seem to reflect the military reality?
- **Content validity –** In the opinion of a SQEP does the wargame cover the things it ought to cover to reflect a particular military reality, and does it give reasonable results?
- **Criterion-related validity –** Do wargame results align with real-world results?
- **Concurrent validity** – If you run the wargame concurrent with real-world activities (e.g. the Ukraine or a training exercise) do you get similar results?
- **Predictive validity -** If you run a game to look at what might happen in the future, does that future actually materialise?
- **Convergent validity** – Do different players, or minor changes in the scenario, or completely different wargames on the same topic all tend to the same result?
- **Discriminant validity** – Does the outcome seem independent of how many times you run it, or who with or how you change the scenario or even underlying mechanics? If so something else might be at play!

Of course, measuring convergence/discrimination might be hard if the system is fairly chaotic (like most battles), and it may be that these are better considered as a spectrum of convergent/discriminant validity – tempered by the reality being modelled.

What immediately seems to be missing from this list from a wargames point of view is "**historic validity**" – to what extent does the construct (the wargame) align with real-world metrics from the past realities.

Drost and others (e.g. Findley, 2021:368) also place Construct Validity within a broader context, which also brings in Internal and External Validity discussed earlier, with the top-level forms of validity being identified:

- **Statistical conclusion validity** – Does a relationship exist between the two variables?
- **Internal validity** - Given that there is a relationship, is the relationship a causal one, or are biases involved?
- **External validity** - How generalisable is this relationship across persons, settings, and times?
- **Construct validity** - Can the variable be operationalised such that it corresponds to the larger theoretical concept of interest?

Interestingly, and possibly since the roots of the discussions about validity are in science, there is nothing here about whether the model/wargame meets its "purpose" or the needs of the user.

And so in wargame terms:

- **Statistical conclusion validity** – Are we able to discern relationships between variables in the game – for instance force mix and success?
- **Internal validity** – Are any relationships causal, or are they result of biases in specification, design, play, players, adjudication or other factors?

- **External validity** – Can we apply the results of the wargame to similar operations other than the one(s) explicitly modelled by the game?
- **Construct validity –** Does the wargame reflects a military reality?

As a wargames researcher and designer obtaining reasonable translation validity seems to be a minimum bar I'm aiming for. As discussed at the start predictive validity (and possibly concurrent validity) is not what we are really focussed on (although our sponsors may be!), but historic validity should certainly be important. Convergent validity should be something that comes through multiple play-tests and real world use, and the idea of playing the same scenario with multiple different game systems is certainly an appealing one as a researcher. Divergent validity is probably a real warning sign during play-testing – if everything you do gives exactly the same result then something in the game may well be broken. As long as we then have decent construct validity the next step would then be to show good external validity.

I think that this is a model that I'll certainly explore further to see how it works "in anger" through the experimentation phase with my urban wargames. Of course, the question for you is whether there are other models that people are using to structure any discussion about the validity of wargames, and which measures (apart from "having fun") do you think are important?

Many thanks to Nick and Evan for comments on an early draft.

References:

Downes-Martin, S. et al. (2017). Validity and Utility of Wargaming. Workshop at MORS Wargaming Special Meeting October 2017. Available at: https://paxsims.files.wordpress.com/2017/12/validity-and-utility-of-wargaming-working-group-report-final-rev.pdf [Accessed 27 July 2022]

Drost, E.A. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, 38(1), pp.105-123. Available at https://www3.nd.edu/~ggoertz/sgameth/Drost2011.pdf. [Accessed 27 July 2022]

Findley, M. G., Kikuta, K., & Denly, M. (2021). External validity. *Annual Review of Political Science*, 24, 365-393. Available at: http://www.michael-findley.com/uploads/2/0/4/5/20455799/arps_2021_external-validity.pdf

Jimenez-Buedo, M., & Miller, L. M. (2010). Why a trade-off? The relationship between the external and internal validity of experiments. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 25(3), 301-321. Available at: https://www.redalyc.org/pdf/3397/339730813003.pdf [Accessed 9 Sep 2022].

Kuehn, K. (2022). Valid and Meaningful Assessment of Wargames. [Video]. King's College Wargaming Network. Available at https://www.youtube.com/watch?v=v3WkYxU4IUM [Accessed 27 July 2022]

Longley-Brown, G. (2019). *Successful Professional Wargames: A Practitioner's Handbook*. Curry, J. (Ed). UK:The History of Wargaming Project.

Mason, R. (2020). Wargaming Hybrid Warfare. [Video]. UK Fight Club. Available at: https://www.ukfightclub.co.uk/webinar-2 [Accessed 27 July 2022].

McDermott, R. (2011). Internal and External Validity. In Druckman, J.N., Green, D. P., Kuklinski, J. H. & Lupia, A. (Eds). *Cambridge Handbook of Experimental Political Science*. Cambridge, UK: Cambridge University Press. Available at: https://app.oarklibrary.com/file/2/b80f0820-3e0d-49de-8467-2b62758beb55/c7bc48f0-1160-4140-bdab-11140eaca397.pdf. [Accessed 9 Sep 2022].

