The Generations Myth

Chris Baird, Chair, NCCD Board of Directors
About 25 years ago, Andrews, Bonta, and others began describing different “generations” of risk assessment instruments. Although this terminology may have initially been intended only to delineate differences in development methods, it quickly gained footing as a means to promote new risk assessment models.

Culturally, “generations” terminology implies an improvement in design and functioning from earlier models to the latest model. After all, a generation-6 smartphone must offer advantages over the generation-5 smartphone, or there would be no inducement to buy the new model.

Developers of risk assessment tools for adult and juvenile corrections have, in fact, claimed that later-generation models are superior to the earlier instruments already in use (see, for example, Andrews, Bonta, & Wormith, 2006). But a review by this author found no real evidence that these risk assessment models perform better than older models. Instead, our review of efforts to promote the generation-3 and-4 instruments uncovered inconsistencies in labeling, a failure to account for all components of some models, flawed logic behind cited evidence, and claims of enhanced predictive power that are not supported by evidence.

Although there is some variance in how each generation is defined, generation-1 decision making is most often described as clinical judgment or the absence of instrumentation. Generation-2 instruments are defined as those relying only on static factors—those factors that do not change over time. Generation-3 instruments introduced the inclusion of dynamic risk factors, also referred to as “criminogenic” needs. Developers claimed that inclusion of these factors allowed risk levels to change over time based on changes in an individual’s circumstances.

Generation-4 models purportedly address “responsivity,” meaning that the system provides information on individual capacities and learning style, identifying programs and strategies that are likely to produce success in the community.

As cited in the manual that accompanies the LSI-CMI, a generation-4 tool, “Andrews, Bonta, and Wormith (2006) examined the predictive validity of different generations of risk instruments. The second-generation risk assessments (such as the PCL-R, Wisconsin, and SFS) had predictive validity in the range of .26 to .46. The third-generation assessments (LSI-R) had an average predictive validity for general recidivism of .36 and the fourth-generation [instrument] (LSI/CMI) had a predictive validity of .41. Accordingly, with the improvement in each generation there was improvement in the predictive power of the instrument” (p. 3).

Putting aside the fact that the authors compare ranges to averages and present no information regarding the power of the analyses cited, the data presented do not support their conclusion. The best results (.46) came from a generation-2 assessment; the generation-3 average falls squarely in the middle of the range cited for generation-2 assessments, and the one generation-4 instrument correlated with outcomes (not described) at a lower rate than at least one of the generation-2 instruments. In sum, these comparisons do not provide any evidence of improved predictive power over generations.
Claims supporting the validity of other higher-generation instruments are equally questionable. Independent studies have found that some have little demonstrated validity (see, for example, Flores, Travis, & Latessa, 2004; Skeem & Eno Louden, 2007). In 2010, the Canada Department of Justice reported, “For the most part, these instruments have been adopted without proper validation and reliability studies” (Hannah-Moffat & Maurutto, 2003). The Illinois Juvenile Justice Commission found little data to support another risk assessment model, the YASI, issuing a report stating, “While the YASI is often cited as highly valid and reliable, the author of this report was unable to substantiate such claims and could not locate any peer-reviewed articles in which these properties were assessed” (Illinois Criminal Justice Information Authority, 2010). In essence, these models were adopted by agencies in the United States and Canada before their validity was established, sometimes replacing well-validated tools already in place.

Recent research comparing results from simple actuarial instruments—those likely to be defined as generation-2 models—found that these instruments actually separated high-, moderate-, and low-risk offenders more accurately than risk assessment systems claiming to be generation-3 or -4 (Baird et al., 2013). A review panel of researchers, including developers of later-generation risk assessments, agreed in a response to the NCCD report that later-generation instruments were not more valid than simple, actuarial models—a significant step back from earlier claims (Andrews et al., 2006).

Adoption of these tools appears to be, in essence, based more on promotion than evidence. The following presents a brief synopsis of what occurred.

When the LSI instruments emerged, there was a concentrated effort to demonstrate their superiority over the most widely used system at that time, the National Institute of Corrections (NIC) model, which was based on a system developed for the state of Wisconsin.

As articulated in the NIC manual, agencies adopting the NIC model were encouraged to validate the risk instrument on their own system-involved population and make all appropriate changes as soon as available data permitted. Many validation studies were conducted, some of which resulted in changes to the initial risk instrument to reflect differences in law, policy, practice, and populations in each jurisdiction.

The NIC manual also explained, in detail, the role of an “assaultive offense” item on the initial risk scale. This item was not a risk factor, but included a weighting that, in effect, established a matrix that considered both risk and prior violence to establish an initial level of community supervision. In Wisconsin, every person convicted of an assaultive offense, regardless of risk level, was placed at the highest level of community supervision for the first six months of probation or parole. Each agency using the NIC model was free to adopt this policy or to delete the assaultive offense item from the scale.

The NIC model also included a reassessment risk instrument that focused on behavior observed since the last assessment. This instrument allowed individuals rated high or moderate risk at admission to move to lower supervision levels over time. Over 50%...
of all cases rated moderate or high risk moved to lower risk levels over the course of their supervision period (Baird, Heinz, & Bemus, 1979).

Finally, the NIC model included a separate needs assessment. The three objectives of this assessment were (1) to ensure that specific needs were considered for every case; (2) to add consistency to the manner in which needs were assessed; and (3) to provide direction for case planning. Finally, as a 2004 National Institute of Justice survey noted, most agencies adopting the NIC model included CMC (Case Management Classification), a component of the Wisconsin system devoted to what is now called “responsivity” (Hubbard, Travis, & Latessa, 2001).

Despite the fact that the NIC model included all of these elements, virtually all comparisons made by LSI supporters focused solely on the initial risk instrument, which LSI proponents labeled a generation-2 instrument. This was a serious misrepresentation, given that the model also included a risk reassessment and a needs assessment. Moreover, this misrepresentation encouraged perceptions that the Wisconsin model was “not dynamic” and that the LSI and other later-generation risk assessment instruments offered improvements over the Wisconsin model.

Another error is more concerning. In comparing the relative validity of the NIC initial risk assessment to generation-3 and -4 instruments, nearly every study included points generated by the policy factor on the Wisconsin scale, “prior assaultive convictions.” Hence, they rated all individuals with a prior assault as high risk, when, in fact, many were not. This incorrect analysis seriously diminished the relationship between risk scores on the Wisconsin scale and recidivism. (For an excellent discussion of this effect, see Eisenberg, Beryl, & Fabelo, 2009). As a result, other instruments appeared to produce higher correlations with recidivism, allowing LSI supporters to claim greater predictive capability.

Other errors were made as well. In comparing results from their own Ohio risk assessment and the Wisconsin system, Latessa and colleagues (Latessa, Smith, Lemke, Makarios, & Lowenkamp, 2009) used a version from a Canadian province that combined scores from the Wisconsin risk reassessment and the needs assessment. Comparing data from an initial risk assessment and a combination of a reassessment and needs assessment is neither meaningful nor useful. The developers’ selection and labeling of this as the “Wisconsin model” seems to demonstrate that little care was taken to ensure that comparisons were legitimate. Despite being alerted to this mistake over three years ago, the Ohio report is, at this writing, unchanged and available online.

In the decade following development of the LSI family of instruments, the corrections field was deluged with articles on the LSI. Nearly all repeated the generations language in their introductions. Most of these studies were based on small samples of cases, limiting their value, and nearly all used correlations as the only measure of validity (Vose, Cullen, & Smith, 2008). Furthermore, any level of correlation, no matter how modest, was presented as evidence of validity. Even when important reviews questioned the level of evidence behind generation-3 and-4 models, their use continued to expand (Skeem & Eno Louden, 2007; Illinois Criminal Justice Information Authority, 2010).
LSI developers also published papers emphasizing the predictive power of “dynamic risk factors” (sometimes called criminogenic needs), often using obtuse logic to support the contention that they were, in fact, better “predictors” than static factors such as prior measures of criminal behavior. When studies found poor relationships between LSI classifications and recidivism, they were generally downplayed; the lack of validity was often attributed to problems with training and/or implementation that resulted in a lack of fidelity with the model as defined. A few studies that analyzed the relationship of individual risk factors to outcomes, however, indicated that the lack of validity may well be due to design issues, as a number of scale factors simply were found to have little relationship to recidivism (Flores et al., 2004, Austin, Coleman, Peyton, & Johnson, 2005; Baird et al., 2013). Despite these problems, the LSI evidence opened the door for other commercially available assessment models, some of which had even less evidence of validity or reliability.

The LSI promotional effort was very successful. By 2014, the NIC model all but disappeared from the correctional landscape, despite the fact there is little, if any, evidence that the LSI produced results equal to or better than those produced by the NIC model.

Although generations labeling seems entrenched in corrections lexicon, it is clear that it has been used to imply superiority where none exists. So-called generation-3 and -4 instruments are not more dynamic, claims of greater validity are simply not true, and claims of additional capacities are, at best, highly suspect. The developers of the YLS/CMI, for example, claim the system addresses “responsivity” and that this makes it a generation-4 instrument (Andrews et al., 2006). Responsivity means the system matches interventions with an individual’s characteristics and learning styles. But there is nothing in the assessment model that does this. Responsivity could perhaps be addressed in training, but the model itself does not contain anything that identifies the learning style or capacities of an individual. While this information would indeed be of value to case planning, it requires far more in-depth analysis than that provided by the YLS/CMI.

There is nothing inherently wrong with using generations terminology to identify real advancements in practice or products. But when differences are not improvements and the terminology is used principally to promote a product, real damage is possible. Resources may be wasted on making unwarranted changes, staff expectations may be raised unrealistically, and case plans may be less effective, meaning that we miss a real opportunity to create a positive impact in our communities and improve public safety. It is especially troubling that some of these systems are finding their way into sentencing and release decision making, assuming a far greater role than is warranted given their limitations. There is a pressing need to clean up the morass of flawed concepts, inconsistencies, false claims, and marketing jargon that permeates corrections. We must be absolutely clear about what risk assessment is and what it can and cannot do for the justice field.
References


