AVMetrics
Flexible Solutions. Precision Results.
AUTOMATED VALUE METRICS

August 21, 2023

Clinton Jones, General Counsel,
Attention: Comments/RIN 2590–AA62,
Federal Housing Finance Agency, Fourth Floor,
400 Seventh Street SW, Washington, DC 20219.

RE: Quality Control Standards for Automated Valuation Models, (RIN) 2590–AA62

Dear OCC, Board, FDIC, NCUA, CFPB and FHFA Agencies:

AVMetrics, a data, technology, and consulting company, was founded in 2005 by Lee Kennedy after twenty years of appraisal analytics and risk management background. The company's primary mission is to leverage AVM testing analytics to improve the prudent use of Automated Valuation Models within the housing finance industry. AVMetrics is not a lender, but our clients include Banks, Lenders, Credit Unions of all sizes, as well as almost all the AVM developers in the United States. We are responding for ourselves, as industry participants and subject matter experts.

For context, Lee has published three peer-reviewed journal articles on the theory and practice of AVM testing and validation. Lee is the current Chair of the AVM task force of the Industry Advisory Council to The Appraisal Foundation. That task force was formed for the express purpose of making recommendations to regulators. It is our understanding that The Appraisal Foundation will provide comments separately in this rulemaking process.

AVMetrics' primary focus remains the independent testing and measurement of AVMs for use in residential housing, so these Quality Control Standards are of utmost importance to our work.

The following pages include an executive summary and AVMetrics' responses to selected questions from the Agencies' Quality Control Standards for Automated Valuation Models.

Sincerely,

Lee Kennedy,
CEO/Managing Director
AVMetrics, LLC

2655 FIRST STREET, SUITE 250, SIMI VALLEY, CA, 93065
T (805) 421-5056 WWW.AVMETRICS.NET

Executive Summary

- We are experts in AVM testing based on decades of extensive practical testing experience as well as academic expertise.
- We support the focus on specific AVM applications such as underwriting decision-making.
- We caution that validating AVMs is a difficult task for most users of AVMs. We believe that many AVM users are not currently doing adequate AVM validation, and they are not likely to ensure compliance with nondiscrimination laws without specific guidance or a simplified way of complying. One alternative is to prescribe a means of certifying AVMs, which would allow AVM users to rely on AVM certification and reduce the uncertainty and cost of compliance.
- Regarding the question of a principles-based approach vs a prescriptive approach, we favor a blended approach that would provide principled guidance that occasionally includes some prescriptive requirements. We believe that in places, more clarity and specificity with respect to the guidelines will help with compliance, which will help with the achievement of the policy goals of advancing fair lending and reducing discrimination.
- We have included references to a detailed write-up on how to test AVMs for bias, and we offer it as a helpful example of how such testing could be accomplished.

Q5:	Question 5. Please address the feasibility of mortgage originators performing quality control reviews of the AVMs that secondary market issuers use to evaluate appraisal waiver requests. What, if any, consequences would such an approach have for mortgage originators' use of appraisal waiver programs?

It would be very difficult for originators to perform such quality control reviews, because normally the AVM is opaque to the originator. In general, most originators have little ability to perform quality control reviews of AVMs that they use themselves, because performing detailed quality control reviews requires a significant investment in data and technology to conduct ongoing testing at scale. The exception is the case of the originator who outsources such quality control reviews, as this is the only practical way to accomplish this objective for all but the largest originators.

Q 9:	Question 9. Are the compliance obligations of lenders and securitizers clear under this proposed rule?

We believe there is ample room to improve the clarity of compliance obligations. With the proposed non-prescriptive guidelines, many lenders will not know how to evaluate the fair lending compliance of their AVM. However, until any method or approach is deemed inadequate, many will do the absolute minimum with the understanding that they cannot be found to be out of compliance until standards are established, and so anything that they do, however flimsy and ineffectual, will be defensible at least once, and that may be enough to postpone the burden of actual fair lending quality control for several years.

Q 15:	Question 15. What, if any, alternate definitions would be more suitable than the proposed definition of control systems? What challenges, if any, would be involved in integrating control systems for AVMs into existing control systems?

The proposal's text describes control systems as the functions (such as internal and external audits, risk review, quality control, and quality assurance) and information systems that are used to measure performance, make decisions about risk, and assess the effectiveness of processes and personnel, including with respect to compliance with statutes and regulations.

This definition of control systems seems well-tuned for appraisals created individually by people. However, AVMs require a highly mathematical, database-dependent, statistically-based "control system." While those descriptions could be contained within "information systems" we believe that the definition would be more useful if it emphasized the analytical and statistical nature of control systems designed for an AVM.

Q 30:	Question 30. Is additional guidance needed on how to implement the quality control standards to protect the safety and soundness of financial institutions and protect consumers beyond the existing supervisory guidance described in part I.A of this SUPPLEMENTARY INFORMATION? Should such additional guidance explain how a regulated entity would implement quality control for an AVM used or provided by a third party?
Q 31:	Question 31. In what ways, if any, would a more prescriptive approach to quality control for AVMs be a more effective means of carrying out the purposes of section

	1125 relative to allowing institutions to develop tailored policies, practices, procedures, and control systems designed to satisfy the requirement for quality control standards? If so, what would be the key elements of such an alternative approach?

Additional guidance is needed in order for regulated entities to effectively apply the first four quality control factors. Our experience throughout the recent decade is that only the most sophisticated regulated entities can effectively establish quality control standards that ensure AVM accuracy and fairness. The vast majority of regulated entities would benefit – as would the safety and security of the nation's financial institutions and consumers – from increased clarity and specificity about AVM quality control.

We've identified three opportunities for additional guidance that could improve quality control standards for AVMs. This content is consistent with our work with the IAC/ TAF task force.

1) Standardized Reporting Elements
2) Universal Confidence Score
3) Certification

Standardized Reporting Elements

This committee provides suggestions for reporting elements, including client and intended users, intended use of the valuation, property identification, property interest, type and definition of value, effective date, models and algorithms used, user's role, reconciliation of current and prior sales prices, property characteristics changes, current and intended use of the property, optimal use conflicts, assumptions, point value and reliability measures, and certification. Which we understand is a long list and may not be obtainable as a first effort but could be explored and reached over time.

For AVM standards, the focus should be on models' reliability (measures) and market analysis (data and outputs) rather than point-value credibility. Specific report standards proposed include tables and graphics representing the data selection criteria, explanation of predictive methods, mapping of the competitive market area (if used within the models programing), bias-variance tradeoff (as a quantifiable measure), and acceptance/rejection criteria that defines the acceptable of an AVM for a particular Use Case.

Universal Confidence Score

Previously, a number of others have tried to simplify AVM evaluation by proposing a universal confidence score metric. The metric that was proposed early on was Forecast Standard Deviation, more simply known as FSD. Though implementing FSD as a standard had good intentions, the difficulty for all to understand it and assumptions for its use may not always be optimal. After evaluating a number of metrics that could be used for the universal confidence score metric, Mean Absolute Error and P10 seem to have a clear advantage over all other options. And, because P10 is much more widely adopted and it is simple and easy to convey, we recommend the adoption of P10 as a standardized "Universal Confidence Score" for AVMs.

How does an AVM vendor implement this new UCS? Ultimately this is a choice that is left up to the AVM developer but to help illustrate what this might look like here are some outlines of possible ways to implement this that follow statistically sound approaches.

Probably the most straightforward approach to implementing a UCS score would be to look at the AVM errors in a given area from a recent time period. In this empirical error approach, one could look at all homes in a given ZIP code or MSA and calculate the AVMs P10 for all sales from the last six months. In regions with 200-300 recent sales, the estimate of UCS from empirical errors, let's call this empirical error the P10region should be a reliable estimate of short-term AVM accuracy. One could then directly calculate the UCS of the AVM in that region where UCS = P10region. The AVM developer would need to periodically refresh the UCS calculation as time passes in order to account for changes in AVM accuracy driven by either market developments or AVM model improvements. One could further refine this approach by sub-dividing by additional property characteristics such as home types (SFR vs. Condos) or by living area or bedroom and bathroom counts - subject to not cutting the sales data sample too thin. Even better would be if the sales errors used to derive UCS are based upon an identical test set of properties across AVMs, any AVM metrics derived from these standardized properties would be comparable. These metrics might consist of the FSD, P10, or UCS as described above. However, if vendors use different test sets, they will be reporting different, non-comparable AVM metrics.

In the end, each vendor will need to decide the best approach to implement the UCS. And the implementation could take several cycles of model-code-deploy-test to determine whether the implementation was a success. For example, if it was discovered that the UCS was overstating the P10 accuracy, then the revised version of UCS would need to make the UCS's smaller. Eventually, a version of UCS would be achieved that aligned with the correct P10 accuracy. Over time, we hope the industry will move to using UCS directly, but we anticipate that there will be some transition time.

Likewise, the AVM vendor may choose to continue to use its existing confidence score in addition to the new UCS in its valuation reports.

Certification

The initial Task Force report broached the topic of certification by stating that we should not rely on self-reported model quality and performance from model vendors and a few independent testing firms that operate without standards. Instead, we must develop both industry-wide standards, as well as create the impetus for centralized testing units and entities, ensuring proper comparative metrics and instilling faith in the numbers for users. Related to this, we see the possibility for credentialization or certification in the AVM space. This Task Force applies the term certification to two separate but related concepts. First, certification refers to the process of evaluating the reliability of AVMs themselves. The initial focus will be on AVM Certification as a properly tested and certified AVM. Second, certification can also refer to the communication of the certification status and appropriateness to an individual use case.

Certification of an AVM could be conducted at different levels of granularity such as certifying use by geography, price tiers, property types, or model metrics such as a Universal Confidence Score (UCS) and prediction interval bands. For example, a third party might certify an AVM only in cases where its UCS is above 70. Combinations of these characteristics could also be applied to further narrows where an AVM is determined to be competent enough to be used for this task.

Additionally, AVMs could be certified at different levels (similar to a residential license or residential certification of appraisers). More accurate and precise AVMs might be certified, with qualifications, for use with more consequential transactions. Models testing with less accuracy or precision could still be acceptable for less consequential use cases such as marketing or portfolio valuation. An appropriate certification could indicate to model users what the most appropriate use of the model could be. For example, UCS >80 could be certified for one particular use case, while those only reaching >70 could be certified for a lower risk use case.

Q 32:	Question 32. What are the advantages and disadvantages of specifying a fifth quality control factor on nondiscrimination? What, if any, alternative approaches should the agencies consider?

The advantage would be a set of minimum standards and key elements that would ensure that a discriminatory factor(s) is *not* present in any model and that every model's performance metrics are acceptable across neighborhoods of every demographic group.

The disadvantage of creating new guidance under this rule would be the possible redundance or conflict with existing fair lending guidance or laws. The alternative approach might be to enhance fair lending guidance to ensure valuation bias is explored when or where fair lending violations are suspected.

Q 33:	Question 33. To what extent is compliance with nondiscrimination laws with respect to covered AVMs already encompassed by the statutory quality control factors requiring a high level of confidence in the estimates produced by covered AVMs, protection against the manipulation of data, and random sampling and reviews? Should the agencies incorporate nondiscrimination into those factors rather than adopt the fifth factor as proposed? Would specifying a nondiscrimination quality control factor in the rule be useful in preventing market- distorting discrimination in the use of AVMs?

Incorporating nondiscrimination factors into the existing four quality control factors would require being more prescriptive about what those factors include. At a minimum, the guidance would have to include the fact that nondiscrimination was a requirement for promoting confidence, but it would be more effective to explicitly include guidance on what kind of testing would be required, how extensive it should be and what must be tested for.

Our response to question 35 describes how we believe such testing should be accomplished. That level of detail may be too explicit for guidelines, but the right answer clearly lies somewhere in between the current non-prescriptive guidance and the very prescriptive detail provided in our answer to question 35 below.

Q 34:	Question 34. What are the advantages and disadvantages of a flexible versus prescriptive approach to the nondiscrimination quality control factor?

A flexible regime allows organizations with different internal capabilities and different dependencies on AVMs to perform different levels of validation. A highly prescriptive "one size fits all" approach might make AVM validation more burdensome than it would be worth, possibly making AVMs impractical for small businesses and allow for an unfair competitive advantage to larger institutions.

However, with only a general admonition to "establish controls," many originators may do the bare minimum, and those controls will likely be qualitative and non-specific. Furthermore, prescriptive guidance doesn't necessarily have to be "one size fits all." It can guide AVM users to adapt their validation practices to their situation and needs.

We believe an effective approach to implementing the nondiscriminatory quality control factor must be prescriptive to some extent. Between the two extremes of "completely flexible" and "completely prescriptive," the proposed approach is 100% at the "completely flexible" end, whereas we believe that a point on the continuum closer to the middle would be optimal.

Q 35:	Question 35. Are lenders' existing compliance management systems and fair lending monitoring programs able to assess whether a covered AVM, including the AVM's underlying artificial intelligence or machine learning, applies different standards or produces disparate valuations on a prohibited basis? If not, what additional guidance or resources would be useful or necessary for compliance?

Robust AVM testing is a process that very few lenders have the capability to perform. Moreover, AVM testing for fair lending bias is very new and undeveloped everywhere. We strongly believe that without some prescriptive guidelines along the lines of the methods or principles described below effective analysis of prohibited bias in AVMs will not be broadly adopted.

We would refer you to the TAF IAC task force paper II for a thorough description of an approach to bias testing. One of the TAF / IAC's AVM Task Force's goals was to develop an evaluation framework for measuring and identifying valuation fairness. The Task Force is proposing a methodology to satisfy fair lending law and guidance. Those six pages give a detailed description of how AVMs could be tested for prohibited bias.

When quantifying and testing for bias in valuations, the Task Force recommends explicitly examining valuations in relation to observed prices in the open market. Best practice is to test for all three types of model errors found; 1) Outlying Valuations; 2) Statistical Bias; and 3) Statistical Variance. We believe that the level of impact to consumers roughly follows in this same order, with the most tangible impacts to consumers of a protected class coming from outlying valuations and the least from statistical variance.

We also agree with the Task Force findings that valuation errors, particularly on transactions for first-time home buyers and credit decision s for protected class borrowers can be devastating for the nation's housing industry and communities. The Task Force's comprehensive evaluation framework can ensure valuations are fair and unbiased and comply with applicable law. In the Task Force's report's appendix, there is a full example of how to test for these types of bias – a methodology that can be applied to all three metrics.

Q 36:	Question 36. What, if any, other approaches should the agencies consider for incorporating nondiscrimination requirements in this proposed rule?

AVM users could be required to demonstrate evidence of nondiscriminatory valuations or rely on nondiscriminatory certification from a third party that performs more extensive testing than a small entity is practically capable of.

Q 37:	Question 37. In addition to providing time for implementation, in what other ways should the agencies facilitate implementation for small entities?

Establishing a certification process for AVMs would enable small entities to simply shop for a certified AVM, eliminating the complications of doing their own validation and uncertainty of knowing whether their compliance measures were adequate.

In the absence of a certification process for AVMs, guidance that is more prescriptive would be helpful to small entities, reducing their uncertainty over what compliance measures will be deemed to be adequate. Uncertainty can impose its own burden by leaving small entities stuck between an unbounded compliance effort and an uncertain liability for non-compliance.

Q 39:	Question 39. Is the number of hours estimated to establish policies, procedures and control systems to comply with the rule realistic for small institutions. If not, what number is hours would be more appropriate?

The number of hours estimated is not realistic. The estimate of 40 hours over three years plus 5 more hours per year might be appropriate for documentation of policies and procedures. But, validating compliance with nondiscrimination laws for AVMs will require significant ongoing efforts.

Our answer to question 35 describes a statistically-based, rigorous process for testing AVMs for non-discrimination. This process requires the acquisition of a lot of data on transactions, AVM value estimates, etc. It requires building a database, cleaning data, carefully building samples and running regression tests. And, it must be performed regularly, because models and market conditions change. A project like that cannot be accomplished in 13 to 18 hours per year.

Of course, different companies might approach the project differently, and some might pursue shortcuts. If a company were to outsource their validation of AVMs, then the original hourly estimate might be adequate, but of course, there would be a cost to outsourcing. Our belief is that a rigorous analytical approach would require between 100 and 400 hours per year, depending on the choices that the entity makes regarding frequency, investment in technology, etc.

Addendum 1:

AVMetrics' Current Testing Methodology Overview

AVMetrics' testing starts with the identification of an appropriate sample of properties for which benchmark values have very recently been established. These are the actual sales via arm's-length transactions between willing buyers and sellers — the best and most reliable indicator of market value. We clean the raw data we obtain from multiple sources to eliminate duplicates, bad addresses and other data discrepancies. We standardize differences such as "No.", "#" and "Number." We assemble this testing data into bimonthly test files to make the testing process as continuous as possible and to ensure that the data is as fresh as possible.

To properly conduct a "blind" test, these benchmark values must be unavailable or "unknown" to the model(s) being tested. AVMetrics currently provides between 750,000 and 1.2 million test records per quarter to AVM vendors (without information as to their benchmark values). The AVM vendors receive the bimonthly test files simultaneously, run these properties through their model(s) and return the predicted value of each property within 48 hours, along with a number of other model-specific outputs. These outputs are received by AVMetrics where the results are evaluated against the benchmark values. A number of controls are used to ensure fairness, including the following:

- ensuring each AVM vendor receives the exact same property list (so no model has any advantage)
- ensuring each AVM is given the exact same parameters (since many allow input parameters that can affect the final valuation)
- ensuring through multiple checks that no model had access the recent sale data, which would provide an unfair advantage

AVMetrics has recently introduced a novel testing methodology designed to tackle the issue of models becoming overly influenced by listing prices in arm's length transactions. This influence can lead to skewed analysis results. The new methodology addresses this concern by preventing the model from anchoring itself to the listing price. As a result, the analysis yields outcomes that are more precise and reflective of actual market conditions compared to situations where the model has direct access to the listing price. This innovative approach has garnered widespread acceptance within the AVM (Automated Valuation Model) community. The majority of AVM vendors and models now consider this methodology to be a fairer and more accurate way of conducting tests and making comparisons among AVM results.

In addition to quantitative testing, AVMetrics circulates a comprehensive vendor questionnaire twice annually. Vendors that wish to participate in the testing process complete, for each model being tested, roughly 100 questions regarding parameters, data, methodology, staffing and internal testing details. In the most recent questionnaire, we included questions about the use of protected class demographic information in models.

These questionnaires enable AVMetrics, and more importantly our clients, to understand model differences within both testing and production contexts, and it enables us and our clients to satisfy certain regulatory requirements describing the evaluation and selection of models (see OCC 2010-42).

AVMetrics then performs a variety of statistical analyses on the results, breaking down each individual geographic market, each price range, and each property type, and develops results which characterize each model's success in terms of error, precision, usability and accuracy. AVMetrics analyzes trends at the national, market and individual model levels, identifying where there are strengths and weaknesses, and improvements or declines in performance.

The last step in the process is for AVMetrics to provide a comprehensive comparative analysis for each model, showing where models stack up against other models in the test; this invaluable information facilitates the continuous improvement of each vendor's model offerings and informs AVM users on prudent selection and use of model outputs.

Addendum 2:

Articles on AVM Testing Authored by AVMetrics

1. How AVMetrics Tests AVMs Using our New Testing Methodology
2. AVMetrics Responds to FHFA on New Appraisal Practices
3. Four Points to Consider Before Outsourcing AVM Validation
4. In the World of AVMs, Confidence Isn't Overrated
5. The Proper Way to Select an AVM
6. Cascade vs Model Preference Table® - What's the Difference?
7. How AVMetrics Tests AVMs