

UNDERWRITING & CREDIT RISK

A bureau score is built to rank everyone. No lender lends to everyone.

Why a published performance statistic is a property of the test that produced it — and how a GCC lender measures the same score on its own book.

SEPTEMBER 2026 DUBAI & RIYADH

01 Our position

A bureau score's measured performance is a property of the test it was measured in, not only of the score. Measured the way a lender actually uses it — forward, on the population that applies, over the horizon the exposure runs — the same score is answering a harder question, and answers it less well.

This is not a complaint about a bureau. It is an observation about measurement, and it holds for every generic score in every market. The number a lender is quoted came from a specific test, on a specific population, over a specific window. Change any of the three and the number changes while the score does not.

Three filters stand between the population a score was measured on and the population a decision is made on. None is neutral with respect to the score.

WHO APPLIED → WHO WAS APPROVED → WHOSE OUTCOME WAS ALREADY RESOLVED

The first is set by brand, channel and geography. The second by credit policy and overrides. The third by a decision about who to include in the test — one that is rarely disclosed and almost never

discussed. They compound, and each can change the measured relationship between score and outcome.

Three propositions sit underneath this, and they are not equally proven.

One — a score's measured performance is determined by how the validation was constructed. Well supported. The developers of the two most-documented generic scores in the world publish the design choices that set the number, and one publishes a validation in which the same score records a Gini of 81.9 measured one way and 70.6 measured another.

Two — a validation sample in which some outcomes are already resolved at the observation point poses an easier problem than origination. Supported in principle and conceded in vendor methodology: excluding accounts already delinquent when the clock starts is a documented, optional design decision. What is not public is what that single choice is worth.

Three — the gap is widest on the thin-file, high-turnover populations GCC lenders are now competing for. Untested in public. A hypothesis, and the last third of this paper is how a lender tests it on its own book rather than taking our word for it.

02 What the number actually means

Every performance number is the result of a test, and before it means anything a reader has to know what the test was. The part that matters most is also the part least often stated: who was in it.

A validation drawn from a snapshot of an existing book contains accounts that were already delinquent on the day the clock started. Their outcome is not being predicted; it is being read. The score is being asked to separate customers who had already failed from customers who had not — a much easier problem than the one origination poses: rank a set of applicants who all look current today, whose outcomes are twelve months away, and none of whose files contains the answer. Same score, two different questions, two different numbers.

This is not our inference. FICO's guide raises it as a decision for the lender: *"It may also not be appropriate to include accounts that are no longer 'actionable'... (e.g. 90-days delinquent at observation) — that is, the accounts are past the point where you can prevent further losses."* VantageScore lists *"Bad at Observation (90+ dpd)"* as a named category in its sample-design rules. Both treat it as a design choice. A choice that is available is a choice that moves the number.

FIGURE 1 · TWO VALIDATION SET-UPS FOR THE SAME SCORE
SCHEMATIC — NOT MEASURED DATA

A · Validated on a book snapshot

Population: accounts already on the book · Some outcomes exist before the test begins



B · Validated the way the score is used

Population: the applicants the score was used on · No outcome exists at t_0



Above, the sample is drawn from the existing book and part of it carries a delinquency that predates the observation point, on a population already approved once and still standing. Below, every case is unresolved at the observation point and the outcome is revealed only at the end of the performance window. Nothing about the score has changed between the two.

So before a performance number means anything, six things have to be on the table.

WHAT A PERFORMANCE STATISTIC REQUIRES BEFORE IT CARRIES A CLAIM

SIX THINGS

The population sampled. The observation point at which the score was taken. The performance window over which the outcome was watched. The definition of the outcome. The exclusions applied. And any weighting of the sample.

WHAT IT MEASURES

Rank ordering — how well the score sorted the sample it was tested on. Not the probability of default at any given score, which is the odds relationship and a separate thing entirely.

THE NUMBER ITSELF

Gini, AUC, KS and divergence are all sample-dependent measures of model performance. None is a property of a model alone.

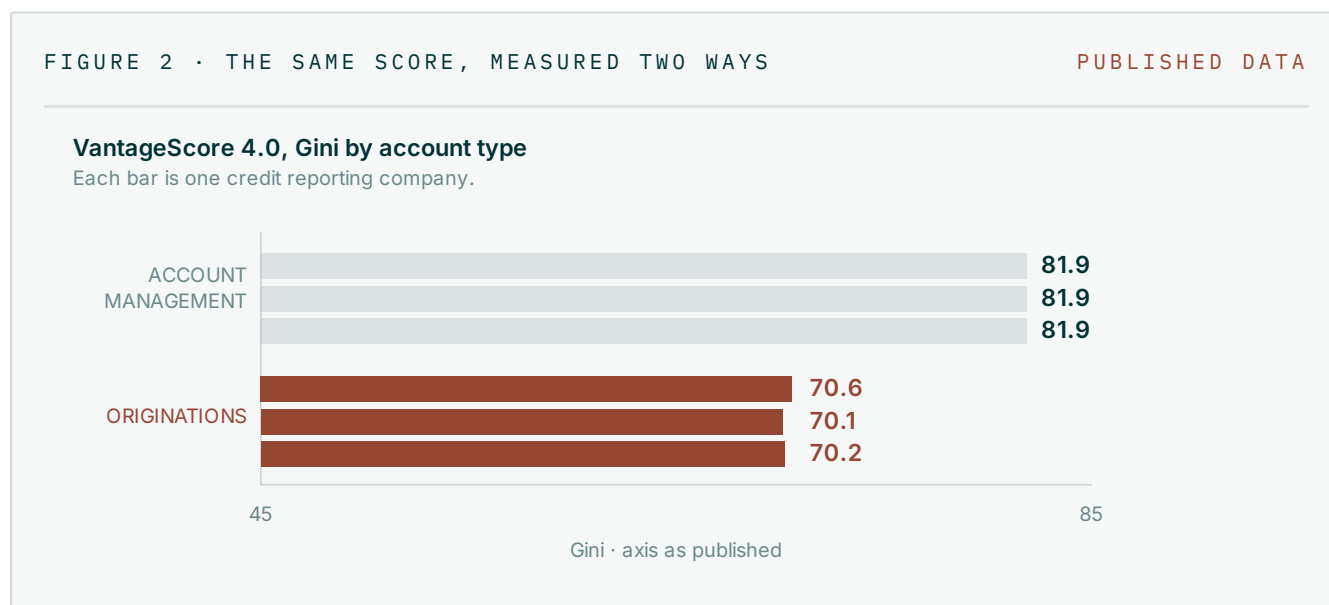
That list is not ours. It is drawn from what score developers and supervisory guidance require or disclose about validation design. FICO's validation guide sets out the windows it recommends — 18 to 24 months of performance for new accounts — defines the outcome as 90 or more days past due, and instructs that policy declines, policy accepts and overrides be excluded or clearly identifiable, which is the second filter named by the vendor. VantageScore publishes its development sample at similar depth, down to the observation date, the window measured from it, and the weighting applied.

This is what disclosure looks like. Six items, published, before any performance number is quoted. It is a reasonable standard to hold any score provider to, anywhere.

None of this requires new science. It requires that the six items be written down and dated before the result is examined. Where they are not fixed in advance, the window, the outcome definition or the exclusions can end up being changed in response to the result — sometimes deliberately, more often simply because the analysis is rerun when the first specification produces an unexpected number. At that point the design of the validation has become endogenous to its result, and validation has become a search.

03 The number moves when the test moves

The clearest evidence comes from a score developer reporting on its own model. In one chart, the same score is measured against two account types. On account management it records a Gini of 81.9 at each of the three credit reporting companies. On originations it records 70.6, 70.1 and 70.2.



Roughly eleven points of Gini, for the same score, in the same document. *VantageScore 4.0 validation study, Figure 10, "Performance Consistency Overall."* The study defines neither population, so the gap conflates sample composition, file depth, base rates and prior selection. It is cited as evidence that the set-up drives the number, not as a measure of any single mechanism.

We will not claim more than that. The study defines neither population, so the gap conflates several things at once and isolates none of them. What it establishes is the first proposition, quantified, by a source with no interest in establishing it: **one score, two differently-constructed set-ups, materially different measured performance**. A lender handed a single number and no conditions has been handed a figure from somewhere on that range and has not been told where.

The distinction underneath it is between the booked population and the application population. A booked population has already been filtered by the policy that approved it; measuring on it tells you how the score ranks survivors. The application population is the population on which the decision is actually made — but its declined outcomes are unobserved. That is precisely why applicant-level validation requires more care than booked-account validation, not less. Choosing the booked population because the data exists is the most common single error in the field.

WHAT A NUMBER IS

The number is a property of
the **test**, not only of the score.

04 Why it matters in the Gulf

Three structural facts, then the expensive version of the problem.

The population is different. A credit history that begins on arrival and stops on departure is not the same object as one accumulated over a working life. Score developers elsewhere have had to define this population explicitly in order to model it at all — new entrants, infrequent users, and consumers with no trades. In the GCC this is not a niche segment. It is central to the populations many lenders are trying to acquire.

The population is changing quickly. Books here expand through new emirates, new employer segments, new nationality mixes, new digital channels and new distribution partnerships. Each changes who applies. A validation vintage is a photograph of an applicant population that has already moved on, and the faster the growth, the staler the photograph.

There is no local measurement to appeal to. No published validation of any bureau score on a representative UAE or Saudi consumer population exists. Not by a bureau, not by a regulator, not by an academic. That absence cuts both ways and we will not use it selectively: nobody can show that a Gulf bureau score underperforms its headline, and nobody can show that it does not. Which means a lender cannot outsource the question of applicability. It has to be answered on a book.

The regulatory framework already asks for this. The CBUAE's Artificial Intelligence and Machine Learning Guidance requires training data *"sufficiently accurate, relevant and representative of the*

customer populations to which the models will be applied." A bank supervisor's examination manual makes the same demand in almost the same words: *"The sample must also be characteristic of the population to which the scorecard will be applied."* Two supervisors, two jurisdictions, nineteen years apart. This is an old requirement, not a new one.

We have watched this happen more than once, and it takes the same shape each time. A retail bank whose branch network and brand sit in Dubai opens in Abu Dhabi and applies its existing cut-offs there. The applications are materially worse than at home. The response is to raise the cut-off for Abu Dhabi.

Two things have happened, and neither is a ranking problem. The first is adverse selection: applicants who expect to be declined elsewhere may concentrate wherever acceptance looks likelier, and a new entrant with no history looks exactly like that. The second is larger, and requires the applicant to know nothing at all. A bank's distribution footprint decides who enters its funnel in the first place. Without brand, branch density or an existing relationship base, the customers who arrive through referral, payroll and existing custom never arrive — they have somewhere better to go.

WHAT ACTUALLY CHANGED

The bank is not seeing a worse sample of the same population. It is seeing a **different population.**

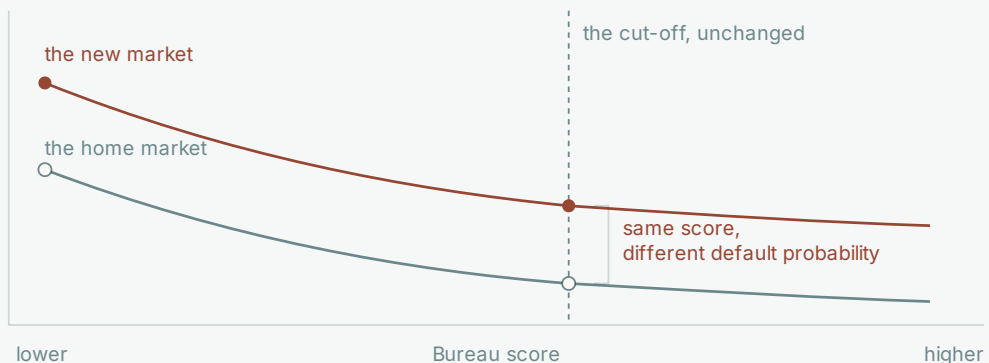
So the same score can carry a different probability of default in Abu Dhabi than it does at home. The cut-off was not transferable, because the odds relationship underneath it belonged to a population that did not exist in the new market. Raising the cut-off treats a population problem as a ranking problem: it moves a boundary along a curve calibrated somewhere else, and it tends to deepen the selection it was meant to correct, because the applicants with alternatives leave first.

A pattern we have observed in this market. It carries no figures and is not offered as evidence.

The same cut-off, two doors

One score axis, two applicant populations, one unchanged cut-off.

Default rate at a given score



Two applicant populations over one score axis — the lower curve the home market, where the odds were calibrated. The cut-off does not move between them; the default rate at that cut-off does. Which is why a cut-off calibrated on one population is not portable to another, and why raising it does not repair the relationship underneath.

05 The score is general, not wrong

A credit function cannot act on a critique of a number it has no alternative to. The constructive position is narrower than “the score is weak.”

A generic bureau score is fitted to rank the entire credit-active population on a generic outcome. It is therefore the best available single ranking across all sub-populations at once — which makes it the best ranking for none of them in particular. That is not a defect. It is the design specification. A universal instrument is exactly what a bureau should build, because it serves every subscriber and none of them exclusively.

No lender is a universal lender. Each operates on a narrower universe: one product set, one price point, one channel mix, one employer and nationality profile. Against that narrower universe, three distinct gains are available, and they are worth separating because they are won in different ways and cost different amounts.

Re-weighting. A variable strongly predictive in one segment and inert in another receives a compromise weight in a generic model, optimal for neither. A narrower universe carries fewer such compromises.

Re-defining the outcome. The generic bad is delinquency on any trade. A lender’s bad is default on this product, at this tenor, at this price, under this collections capability. Not the same event, and not the same drivers.

Re-calibrating the odds. The cheapest of the three and the most often skipped. The mapping from score to probability of default was estimated on a population that is not the lender's. The supervisory formulation is direct: *customized odds charts are more predictive than odds charts that are available from the bureaus*, because a bank's particular market has different characteristics and therefore different odds. Re-calibration needs no new model — only the lender's own outcomes against the score it already buys.

The practice supports this. In the US card market, issuers have built internal models on bureau attributes and application data for decades and found them materially better on their own populations than the generic score alone. That was our experience at Capital One. We attach no figure to it; the mechanism travels even where the magnitude does not.

The detail that matters is what those models did with the bureau score. **They consumed it. They did not replace it.** It entered as an input alongside attributes and application data, and it earned its place there.

THE CONSTRUCTIVE POSITION

A bureau score is not a decision. It is a **variable**.

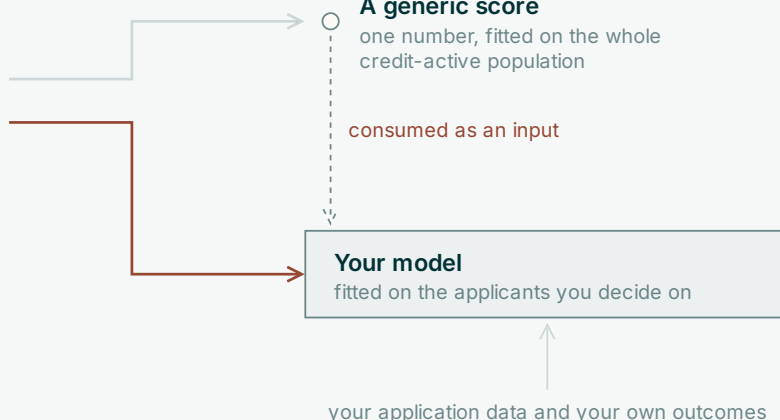
A subscription also delivers more than the variable. The file underneath the score is data a lender can model directly, and a score is one compression of that file, optimised for someone else's population. What is available at attribute level, at what granularity and with what history, is a question for each lender to put to its bureau — the same class of question as the six disclosure items, and it should be asked at the same time.

The same file, two compressions

A generic score is one fitting of the bureau file. A lender can fit another, on its own applicants, with the score inside it.

THE BUREAU FILE · available under subscription

trade history
balances and limits
utilisation and its trend
enquiries and their timing
account ages
payment patterns
delinquency depth



Seven kinds of information on the left; one number on the right. The compression is what makes a generic score portable between subscribers, and it is also what makes it general. The same underlying bureau data can support a second fitting — on the applicants a lender actually decides on, against an outcome it actually cares about — with the generic score consumed as an input rather than replaced.

06 What the credit function should measure

The chain runs in one direction. Only the last link shows up in the P&L.

MEASUREMENT → RANKING → DECISION → ACTION → ECONOMICS

Two things are conflated at the first link and should not be. **Rank ordering** is how well the score sorted the sample. **Calibration** is what a given score implies about the probability of default. A score can continue to rank correctly while the odds at each band move — through the cycle, through portfolio mix, through a change in acquisition channel. Rank-order statistics will not detect this. Population stability tells you the population has moved; a periodic odds chart tells you whether the score-to-outcome relationship has moved with it. A credit function monitoring Gini alone is monitoring only one half of the problem — and often the half least likely to break first. And the Abu Dhabi failure above is a calibration failure that was diagnosed as a ranking one.

Better ranking is not an economic result. The honest version of the question this paper raises is therefore not *how much worse is the score than advertised*. It is:

At your existing loss threshold, how many decisions would change if the score were calibrated and validated on your own applicant population — and what is that worth?

Every term is observable on a lender's own book:

```
incremental value
= (decisions changed × contribution per decision)
  - incremental expected loss
  - incremental acquisition and servicing cost
  - cost of capital on the incremental exposure
```

Recalibration is the cheapest term to establish, because it needs no new data — only the lender's own outcomes against a score it already pays for. If the answer is small, the measurement critique is academic on that book and should be dropped. If it is large, it was never a statistical question.

07 Four levels a credit function can run now

None of this requires a programme. Four levels, each producing something usable before the next begins.

One — ask for the conditions. Before accepting any statistic from any score provider, ask for the six — including how accounts already delinquent at observation were treated. A procurement question, not a technical one, and the answer decides whether the number can be compared to anything.

Two — plot your own odds chart. Take twelve to twenty-four months of your own originations, the score as pulled at application, and the outcome as observed. Plot bad rate by score band — a mapping, not a model, and most credit functions can produce it in a fortnight. Then run it separately by emirate, channel and segment. The divergence between those is the finding.

Three — validate the score the way you use it. Fix the windows first and date the document: observation point, performance window, application population, outcome definition, exclusions, and the policy in force at the time. Then run it, and run it again on the booked population alone. The difference between the two is your own version of Figure 2, and the most informative number in the exercise.

Four — build on the file, not only on the score. Fit a model on bureau attributes, the generic score as an input, and your own application data, on your own applicants. You then own the population, the window and the exclusions because you chose them — which is also the easiest position from which to satisfy a representativeness requirement.

One rule governs all four. Hold the incumbent and the challenger to the same population, the same window and the same economic outcomes. A new model is routinely asked for out-of-time, out-of-sample evidence on a representative population while the incumbent score, which has been setting cut-offs for years, is asked for nothing. A challenger should not have to clear a bar the incumbent was never held to.

Level four has a limit worth naming rather than discovering. Re-weighting, re-defining and re-calibrating all work on the file that exists. Where the file is thin — a new entrant, a first-time borrower, a customer six months into their first job in the country — there is less to re-weight, and better fitting recovers less. At that point the remedy is not a better fit to the same observation. It is a different observation of the same customer, and that is the subject of Paper II.

08 The credit officer's question

A quoted performance statistic answers a question asked somewhere else: how well did this score separate *that* sample, over *that* window, after *those* exclusions. It is a real answer to a real question, and it is not the question in front of a credit officer.

That question is narrower and harder. Among the applicants arriving at this bank, through these channels, in this market, this month — all of whom look current today — which will still be paying in twelve months, where should the boundary sit, and what is a decision at that boundary worth?

It can only be answered on the lender's own population, and the instruments are already in the building: the score, the file underneath it, the application data, and the outcomes the book has been generating all along. What is usually missing is not data or method. It is a written, dated statement of what is being measured, on whom, over what window — set down before the result is examined.

The bureau built a universal instrument. It is designed to do that job, and a lender should keep using it. But nobody lends to everyone, and the value of a decision is not a property of the score that informed it.

Measure it on your own book.

AUTHOR

NYK Advisory
Dubai & Riyadh

PUBLISHED

September 2026
Paper I of IV

PRACTICE AREA

02 — Underwriting & credit
risk
04 — Decision economics
& risk-based pricing

GET IN TOUCH

If a number in your book is moving and nobody can tell you why, that is usually the right moment for a conversation.

CONTACT

nykadvisory.io
hello@nykadvisory.io

PRACTICE

Credit risk & decisioning
Dubai & Riyadh

THE SERIES

Four papers through Q4
2026

REFERENCES

1. FICO, *FICO® Score Validation Guide* (2017).
2. VantageScore Solutions, *VantageScore 4.0 User Guide*, revised September 2022.
3. VantageScore Solutions, *Inaugural VantageScore 4.0 Trended Data Model Validation* (2018), Figure 10, "Performance Consistency Overall."
4. Federal Deposit Insurance Corporation, Division of Supervision and Consumer Protection (March 2007), *Risk Management Examination Manual for Credit Card Activities*, Chapter VIII: Scoring and Modeling.
5. Board of Governors of the Federal Reserve System / Office of the Comptroller of the Currency, SR 11-7, *Supervisory Guidance on Model Risk Management* (2011).
6. Central Bank of the UAE (11 February 2026), *Guidance on the Use of Artificial Intelligence and Machine Learning*.
7. Central Bank of the UAE (2022), *Model Management Standards*, Notice 5052/2022.