Strategic Whitepaper

November 2011

Normative Test Scores in a Performance-Oriented Personnel Selection Strategy

Issue:

When benchmarking (as in engagement, work-stress, safety, or performance-related applications), the use of normative scores can be a big mistake. Because if the norm group characteristics change in any way, the entire benchmarking process is rendered problematic. More importantly, if score magnitudes are considered related to performance, then it is the actual score which carries that relationship, not its "normed-percentile", sten, or T-score version.

Advanced Projects R&D Ltd.

- Psychological Test Design and Construction
- Predictive Analytics and Profile/Classifier Construction
- Independent Scientific Evaluation/Optimization of Psychological Test Validity

19 Carlton Road, Pukekohe, Auckland 2120, New Zealand

T	+64-9-238-6336
M	+64-21-415625
F	+64-9-280-6121
W	www.pbarrett.net
E	paul@pbarrett.net
S	pbar088

Executive Summary

Q1. When a test publisher/employer/recruiter begins using a psychometric test scale as part of a selection process, where a particular score on a scale is required to be used as a threshold for a "minimum likely performance/ literacy standard" or "filter" for applicants, the first question they face is "how do I choose the "threshold" score?

Answers

1. The response to Q1 requires a clear choice to be made, between setting a threshold subjectively or using an empirical evidence-based approach.

2. A subjectively determined threshold may be constructed by Subject Matter Experts or by some other "credentialed person/s" subjective judgments about a plausible threshold which will discriminate effectively between candidates who meet or do not meet an *indicative* performance standard (*indicative* here means that the test score is used as an indicator of likely job performance/training outcomes).

3. It is argued that the only conditions which might justify the setting of a subjective threshold is when:
   a) no empirical data are available to a test-user, in advance of the selection process, in order to determine an indicative performance threshold.
   b) such empirical data are physically impossible to acquire prior a specified date for setting that threshold.
   c) the psychometric test is so closely related to the specific job performance criteria that it is a simple matter to set a score threshold which possesses a near one-to-one correspondence with actual on-the-job performance.

4. However, if a subjective threshold is set, then it can only be considered temporary, with a requirement to collect objective performance data related to the job performance metrics prior and post the new selection process, such that an "evidence-based" threshold might eventually be attained.

5. In all other cases, an empirically determined threshold should be constructed, using test score data from "ideal" and "less than ideal" job incumbents, or from external data sources which can be argued to be representative of the particular employment situation. Such a threshold would thus possess an evidence-base for its use, along with likely error-rates and the means to later engage in a quantifiable predictive ROI analysis of the selection strategy.

Q2. Should raw or normatively-interpreted scores be used in selection settings? That is, should an employer use the raw scale score to represent a magnitude of some attribute for a candidate, or instead, re-express the score relative to a normative set of scores provided by an assumed homogenous group of individuals?

Answers

6. From some detailed "closely-matching-reality" simulation work, there is clearly no justification whatsoever for using transformed scaled scores such as stens, T-Scores, stanines etc. in a performance-oriented selection process, except where the norms are properly representative, substantive in constituent number, and remain static (i.e. are not cumulatively updated or "*bootstrapped*").

7. Where a user insists upon using transformed scores as part of a selection strategy, then detailed diagnostics must be implemented when norms are augmented, in order to guard against inadvertent degrading of performance outcomes.

The Detailed Reasoning/Logic concerning Question #1

Q1. When a test publisher/employer/recruiter begins using a psychometric test scale as part of a selection process, where a particular score on a scale is required to be used as a threshold for a "minimum likely performance/ literacy standard" or "filter" for applicants, the first question they face is "how do I choose the "threshold" score?

Here the typical scenario is an employer wishing to use a psychometric test as part of a new selection strategy, where previously they have been using interview data, performance data, biodata, and/or personal recommendation. They might also be considering replacing an existing test with a new one, where no data exists as to the functional relation between the "old" and "new" test scores. The primary use of the scale score/s for the test is as part of an applicant "filtering" process – where the test might deployed as an initial online screening device, or in-house "promotion-potential" indicator. A score-magnitude is required so as to "screen in/out" applicants who either exceed or score below this target threshold magnitude. The question is how to set that threshold or cut-score score so that an optimal screening process might take place.

The answer requires consideration of two options: setting a cut-score subjectively, or using an empirical evidence-based approach.

Subjective "Credentialed Opinion" Cut-Scores

Here, the test user will attempt to create/infer the link between threshold value and outcome. This link may be constructed by Subject Matter Experts (SMEs) or by using some "credentialed person/s" subjective judgments about a plausible threshold which will discriminate effectively between candidates who meet or do not meet an *indicative* performance standard (*indicative* here means that the test score is used as an indicator of likely job performance/training outcomes).

However, the use of SMEs is slightly peculiar, in that the decision to be made is not only about "which score is optimal", but also about the kind of function which relates a psychometric scale of responses (usually a cumulative sum-score) to job performance or training outcomes. It is the same conjoint judgment required if a "credentialed" person such as a psychologist, management consultant, or HR "expert" is called upon to offer an opinion.

For those who would insist that their judgment or opinion is "informed by experience, skills, competencies, or expertise", there is now a large body of evidence which exists as to the likely inaccuracy of such judgments (Grove and Meehl, 1996; Grove, Zald, Boyd, Snitz, & Nelson, 2000; and Ægisdóttir, White, Spengler, Maugherman, Anderson, Cook, Nichols, Lampropoulos, Walker, & Cohen (2006). Indeed, the paper by Meehl (1997) on "*credentialed persons vs credentialed knowledge*" highlights this very distinction between decisions made upon the basis of empirical facts versus those made by "experts" who rely upon "credentialed qualifications" for the validity of their judgments.

Evidence-Based Cut-Scores

Therefore, best practice in these conditions indicates that an optimum threshold score should be generated using an empirical evidence-base, which relates the psychometric scale score magnitudes directly to the desired criterion of job performance or training/promotion outcomes. To do this will require the testing of one or more samples of particular job incumbents, ranked or grouped in terms of magnitude or classes of desired criterion outcomes (or some function of these), such that their psychometric test scores might be functionally related to these, and an appropriate cut score chosen to maximize cross-validated (ideally) predictive accuracy whilst minimizing or balancing/optimizing false-positive and false negative rates. Obviously, such a process requires time,

sensitivity to employee concerns about other potential uses of such data, and some fairly sophisticated analysis of the data prior to setting what should be an optimum (from many perspectives) threshold. Alternatively, if the opportunity exists, the new test might be run alongside the existing selection processes, with a follow-up of job performance outcomes or training over time of those "selected" under the old system. However, this latter strategy whilst avoiding incumbent employee issues, is likely to result in up to a 6-month or even a year's delay in outcome evaluation and threshold setting.

Note that the choice of a threshold score using the methodology in the preceding paragraph is unlikely to be entirely objective; rather, the "objectivity" is bound up in the empirical information gathered to speak directly to the **consequences** of any threshold value chosen. Although the estimated costs and benefits of a particular threshold might rely upon factual outcomes, the balancing of those costs and benefits within the overall context of the selection process will remain a managerial-judgment decision, albeit now rather more informed than just credentialed or "expert" opinion.

One might be tempted to use just use published evidence of score ranges on a particular scale or test drawn from meta analyses or other published sources or, even other normative groups. But, such data may or may not generalize to a specific employer, in a specific market, in a specific country. *Validity generalization* is fine for just that, generalized "on-average" statements about validity. What if a target organization is not "average"?

Where it is impossible to collect in-house data specific to the criterion and employment/ applicant population at hand, then reliance on this 3rd-party data may be the only acceptable alternative between actually acquiring an in-house evidence-base, and making judgments based upon more subjective qualities of the variables being assessed. For those who question this advice, consider the situation where a company relies upon the normative "test manual" dataset where the correlation between the scale scores and rated job performance is 0.40. However, if the company tested its own employees on the scale, and used its own supervisor ratings of their job performance, the correlation might be substantively lower (or indeed higher) than the "test- manual" value – due to the fact that the constituent components of job performance in the test-manual dataset is not quite the same as in the company wanting to deploy the test. Further, supervisor ratings may be so unreliable in the current company that the expected criterion performance predictive validity is seriously attenuated from its expected value. In both instances, the company only finds out the worst when its new employees fail to perform, or train, as expected. That can be a very expensive mistake.

There are situations where the psychometric test content is virtually a work-sample. Under these conditions, it might be perfectly justifiable to define a threshold score "subjectively", as performance on the test relates almost identically to performance on the job, hence the functional relation between test score and outcome is direct and known *a priori*.

Of course, there are also situations where it is physically or financially impossible to acquire empirical evidence to support the choice of a threshold or cut-off score. However, such situations occur more due to the unwillingness of a prospective test user or consultancy to spend money (and time) in advance of the deployment of the test for selection purposes, than because of literal physical impossibility, or a seriously non-beneficial ROI.

I have to say in every case I have been involved with, where organizations have launched into using tests "straight out of the box", the excuses for not selecting an evidence-based threshold are primarily financial and "political". That is, the employer representative (HR director, manager) through to consultant and I'm afraid, even many commercially-oriented I/O psychologists, over-sell and over-simplify the strategy to executive management,

failing to understand the substantive consequences of taking such risks with selection. Invariably, both buyer and seller will have likely "moved on" just as the failures of the selection procedures start revealing themselves!

One final point, for companies with small numbers of employees, who wish to use psychometric tests, they have no option but to rely upon published 3rd party-evidence in support of the test they might wish to use. However, under these conditions, the ROI calculations associated with using a cut-score are dramatically different than those for a larger corporate; many other considerations are likely to entail. Further, they are likely to be entirely reliant upon the expertise of the seller of the psychometric test, and any training they undergo with a test publisher or the consulting company. Again, all I can recommend here is that such employers/prospective test users demand empirical evidence for any assertions made as to likely success in using the test in their company. But frankly, how many small employers would realise that a validity coefficient of 0.35, calculated over 32 individuals is useless for all practical decision-making purposes (the 95% confidence interval ranges between 0.0 and 0.62). Even over 100 individuals, the 95% confidence interval for that correlation of 0.35 varies between 0.16 and 0.51. Such simple calculations are sobering for any decision-maker about to invest perhaps thousand of dollars in psychometric test training and subsequent test usage. Perhaps the most simple rule is never buy from a test company or consultant who is unable to show predictive validity estimates on substantive sample sizes(greater than say 100 cases) for the test scales which are being considered for use in a selection setting.

Ultimately, setting an optimum cut-score is a complex matter. It is also one which should be approached from a robust evidence-based perspective where at all possible.

The Detailed Reasoning/Logic concerning Question #2

Q2. Should raw or normatively-interpreted scores be used in selection settings? That is, should an employer use the raw scale score to represent a magnitude of some attribute for a candidate, or instead, re-express the score relative to a normative set of scores provided by an assumed homogenous group of individuals?

The arguments below apply equally to ***I**tem **R**esponse **T**heory* (IRT) as to ***C**lassical **T**est **T**heory* (CTT: the familiar sum-of-item-response approach to creating scale scores). IRT is quite different from CTT in that it constructs a latent variable upon which questionnaire items are located in terms of the "amount" of this latent variable they "possess". Then, "measurement" of an individual's magnitude or "standing" on that latent variable is estimated by examining items to which they respond to "positively" (i.e. in the direction of the meaning of that latent variable). From the mathematics behind the construction of that latent variable, when using reasonably representative "calibration" samples, individuals may be "located" on the latent variable in the same way as the questionnaire items – in terms of the magnitude of the attribute they "possess". However, like CTT constructed scales, the "scaling" of the latent variable is "floating" in that there is no "true zero" as we might have for a quantitatively structured variable like length.

So. when it comes to the expression of a test score – both IRT and CTT protagonists can choose whether to re-express test scores in the original metric of the scale (with IRT scaling as logits or some derived equal-interval scale, and with CTT as simple summed item scores) – or – express an individual's IRT or CTT test scores relative to some subgroup of individuals who possess scores on the particular attribute of interest. Quite simply, the decision required is whether to treat the scores as **absolute** or **relative** indicators of magnitude.

If we treat the scores as absolute, then we use raw/logit scores, as the score is taken as indicative of the magnitude of attribute a person is said to possess at that point in time. This suits the use of scores as estimates of performance outcome, as score magnitude is to be scaled against category, probability, rank, or some continuous measure of outcome. Changes in scores are then indicative of likely changes in performance outcomes. In a sense, this use of scores might be better known as "*criterion referencing*" – where a test score is calibrated against a criterion such that the score may be used as a predictor for that criterion. But, the principle is that the score magnitude is treated as such – a *magnitude* of some attribute. Whether it is a predictor or otherwise of a criterion is just one of the ways the score might be used or interpreted.

However, it has become popular in I/O psychology to reference raw scale scores against groups of individuals who are homogenous with respect to one or more attributes e.g. female sales managers, call centre operatives, management graduates, bank business managers etc. The raw score is usually re-expressed as a standardized and normalized score, such that the value of the transformed score not only reflects *relative* magnitude but also the proportion of individuals in the homogenous group who possess that score. Indeed, that proportion, although based upon a single homogenous sample-group, is usually taken as some "estimated" population value. Naturally, this inference is conditional upon the satisfactory sampling of that hypothesised population. Angoff (1971) provides a detailed exposition of the various methodologies and concerns associated with constructing norms. Crocker and Algina (p, 432) also provide a nice step-by-step guide as to how to construct norms.

These *norm-referenced* scores are popular with practitioners because a raw score on a psychometric test conveys little about what the magnitude of an attribute might actually mean (the exception in some respects is an IRT score, but with no true-zero for its measurement, it also "floats" free until some "anchoring" is applied). But, when referenced against a homogenous group, an instant comparison may be made between the score and the "average" or "exceptional" score for a relevant group of individuals.

Are normative or absolute/criterion-reference scores interchangeable? No, is the short answer. With a norm referenced score, when you change the norm group, you likely change the transformed score. Not a problem in a situation where comparative interpretative analysis is being carried out on an individual's score, but a huge one where that score is interpreted in a selection situation using "normalised cut-scores" like stens or stanines to select or decline candidates. If those norms are augmented mid-selection (over a year or two), then unless the new norms match near-exactly the sample score distributional characteristics of the older norms, scores which might have resulted in selection might now result in rejection, and vice versa.

The Selection Scenario

At the moment, this is all "abstract". Let me now demonstrate just how badly things can go wrong in a selection scenario where an organisation constructs norms for itself during the selection process. This happens where an organisation wishes to use a test for which there are either no relevant norms, or a norm group with maybe just 100 or so "convenience-sampled" individuals from maybe a training course or two, who happen to possess occupational titles such as "sales manager" or "HR administrator" etc. Further, a screen-in/screen-out cut-score is required to be used.

The immediate issue is whether to use raw or norm-referenced scores for expressing the cut-score and applicant scores. Let's assume the organisation decides to ignore my advice above, and use norm-referenced scores, stens in fact – using a sten score as a cut-off. The test itself is an ability test – let's say a composite general mental ability (GMA) test.

Next, what norm group should be used? The test publisher (in line with most) will likely have "general population" norms. So the issue now is what constitutes a useful "cut-off" score in relation to the general population? What has to be considered here is the criterion being predicted, and how the general population norm is functionally related to the criterion outcome. Let's assume the criterion is a composite measure of training success. Without any data linking the test score to that outcome, then a population norm might only be useful in an abstract sense i.e. using 3rd-party published data to form inferences about what percentage of the population might be "suitable" for selection in terms of the attribute being assessed. This is a highly subjective and sub-optimal process and should only be undertaken as a "last resort" option, and not as a "strategic" initiative.

So, the organisation has little idea what the cut-score should be. It can initiate an objective investigative strategy as outlined above under the heading "Evidence-based cut-scores", it can make subjective, *credentialed opinion* kinds of inferences, it can try and draw inferences from published evidence-bases on samples of employees, or it might attempt to "*bootstrap*" its own norms.

Bootstrapped Norms

This latter path of action, on the face of it, sounds like a reasonable strategy. It is a way of constructing norms using data from applicants who apply for selection, whilst using a cut-score which is generated from some pilot data on actual performance outcomes. In our example here, the organisation would use the scores on the chosen test it administered to 100 or so more recent selected job applicants who had undergone part of all of its criterion process (training), create a norm group from these training incumbents, express the raw test score as a norm-referenced sten score, then examine these against a performance outcome and set an initial cut-score at say sten 4. That is, it might feel that if it used such a cut-score as a selection pre-screen, it would have been able to optimise the number of candidates who were successful in training versus those who were not, whilst minimising false positives and negatives (as well as taking into account the requirement to meet its recruitment targets).

It then proceeds to select staff using this norm group as the reference normative sample, and the cut-off score of 4 stens. Cohorts of new applicants (all those applying for the position and being administered the test as a screening device) are added to the norm base periodically. Sometimes these applicants are from recruitment fairs at universities, trade shows, colleges, sporting occasions, and other major events, and through media PR and internet advertising. The organisation is looking for talent with potential to be trained for specific tasks.

Sounds like a reasonable approach? It can be, but only as long as specific diagnostics are implemented when the normative sample is augmented with new data, and a constant vigilant process of outcome evaluation is

maintained. If not, well, the real-world simulation which now follows demonstrates only too clearly the cost of bootstrapping norms without implementing diagnostics such as those use below.

In the end, if the organisation had just worked with raw test scores from the outset, none of the problems associated with working with norm-referenced scores would have arisen. This issue arises both in benchmarking applications (organisational stress, employee engagement etc.) as well as selection applications. It is absolutely critical that organisations do not use norm-referenced scores for either. By all means use a scaled score (say between 0 and 100) for ease of interpretation and comparative analyses across cohorts and repeat samples etc. But they remain simple "magnitudes", and not transformed "relative" magnitudes.

The Simulation

The organisation is required to recruit several hundred new semi-and skilled employees each year. Each employee requires some form of training prior to being able to enter the workforce proper. A general mental ability test (GMA) is desired to be used as a pre-screen. That is, those candidates/applicants not achieving a minimum cut-score score on this test will be rejected, those meeting or exceeding the cut-score will be interviewed and likely accepted for training.

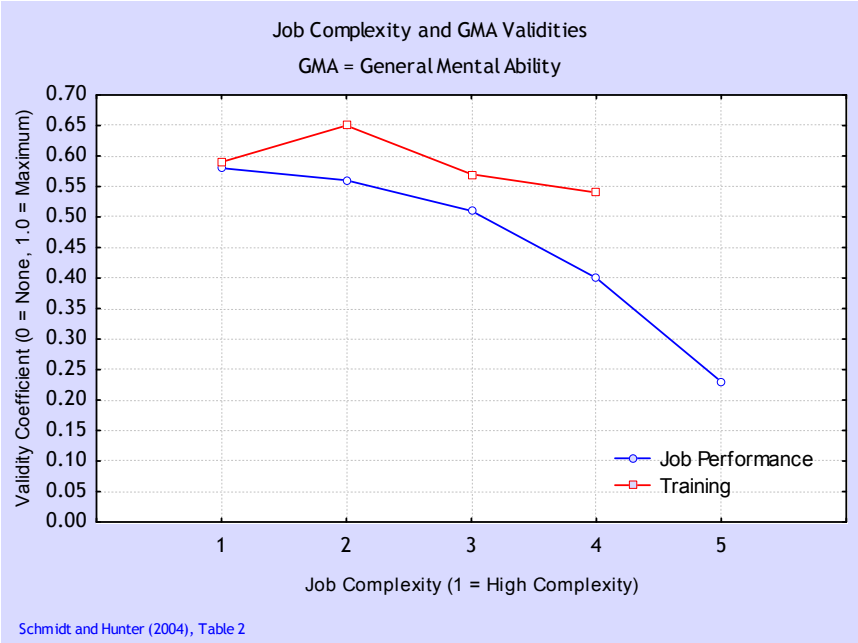
The GMA test is being used as a likely (from meta-analytic evidence) predictor of training performance. The test possesses general norms – but none which speak directly to the particular performance criteria at hand. The HR manager responsible for recruitment is minded to use norm-referenced transformed raw scores (stens) for candidates. HR assesses 210 already selected candidates using the new GMA test, and looks at an initial composite performance criterion (CPC) outcome defined as "training success". It creates a new norm-group from these 210 individuals, re-expresses an individual's raw score as a sten score, evaluates the relationship between the score and outcome, and decides that a sten score of 4 should be the cutoff. In future, applicants scoring 4 or more will be accepted for training and employment.

Details

- GMA norm-reference score range will be between 1 and 10 (stens), with a mean of **5.5** and SD of **2.0**.
- GMA general population raw score norm mean is **21**, with an SD of **6.3**.
- The population Composite Performance Criterion scores (CPC: calculated using a year's CPC scores in the organization) range between 0 and 100, with a mean of **70** and an SD of **7.3**.
- Initial normative group = 210 pre-selected incumbents in training. These incumbents have already been pre-selected via the previous selection process.

For the simulation, I will use the global "population" norm-group raw score mean and SD as representing the true population mean and SD on the test. Then, I will sample from this distribution by applying a cut-score to lower level GMA scores, mimicking the effect of the selection process which although it did not specifically use a cut-score, is likely to have selected candidates above some minimal literacy and ability threshold.

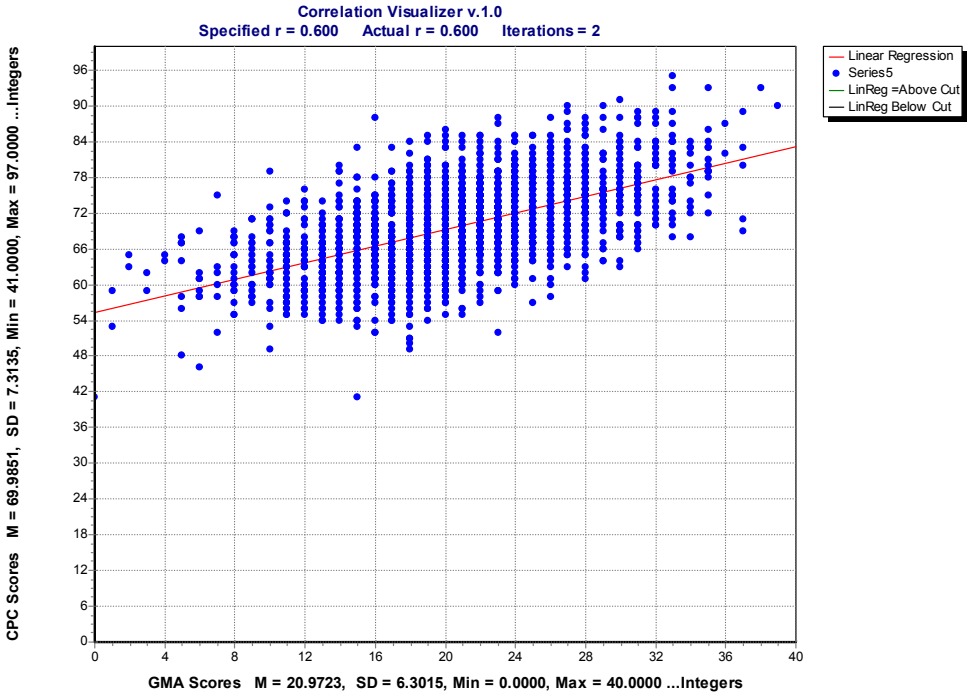
In order to generate a population dataset, I need to know what the expected correlation between GMA and CPC would likely be in the total population. A reasonable estimate might be taken from the literature, say Schmidt and Hunter (1998, 2004), or Salgado et al (2004). This puts the attenuation corrected meta-analytic correlation somewhere between, 0.4 and 0.6. A useful graph can be constructed from the meta-analytic data provided by Schmidt and Hunter 2004, Table 2. Let's be optimistic and set the population correlation at **0.60**.



Step 1

I need to create a simulation dataset which will closely approximate real-world data in this particular scenario. I'm using scores sampled from a bivariate normal distribution.

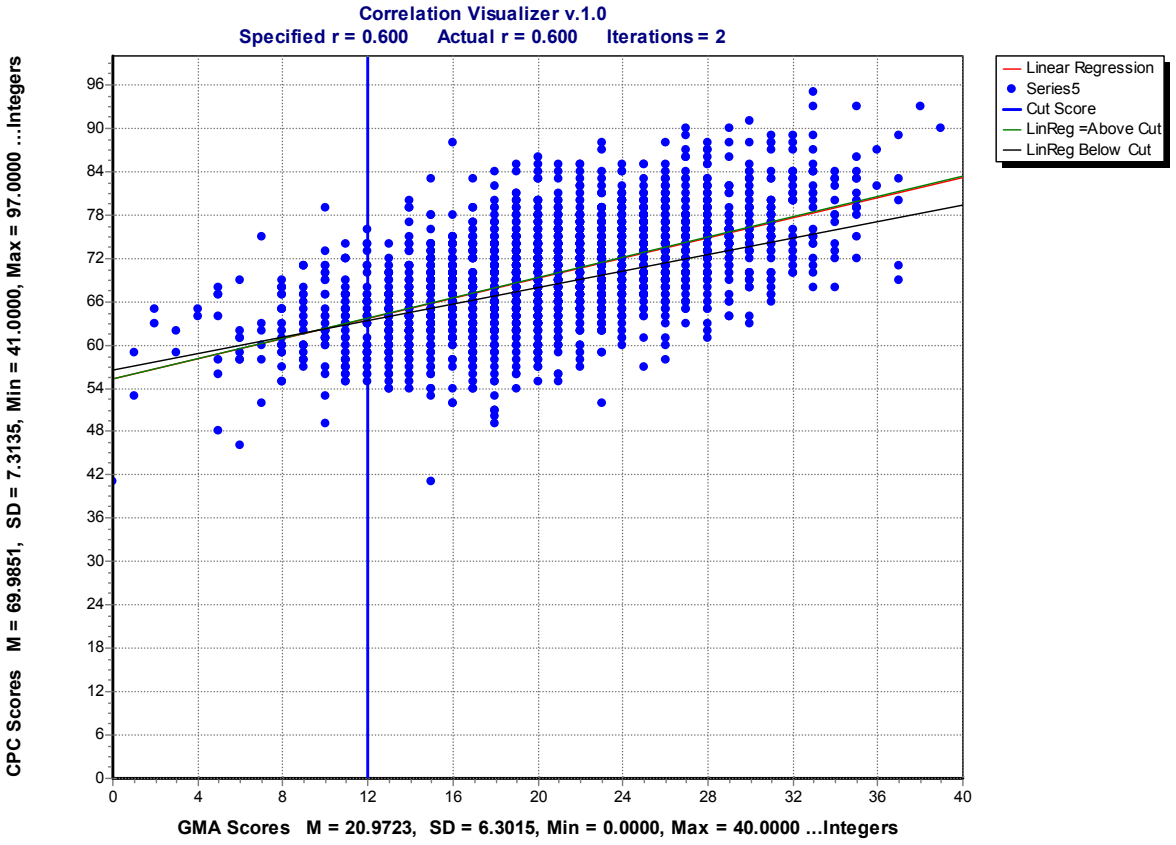
- First, I need to generate the population dataset where GMA and CPC scores correlate at 0.6. I create dataset #1 with 10,000 cases which satisfies the constraints of the known "population" means an SDs, and correlation.



- But we must take into account that individuals with particularly low GMA scores would never apply for a job such as this. The problem here is that unless the test publisher has attempted to equate the test with a gold-standard IQ test such as the Wechsler, Ravens, MAB, or some other IQ normed test, then we have no idea where say learning disabled or very low literacy individuals might score on the test (IQ < 70).

- So, I'll assume that as with an IQ-normed test, it is unlikely to see people with IQs less than 80 applying for the kind of semi-skilled and skilled "corporate-level work" that we are selecting for here. Likewise the test publisher is only likely to have tested individuals in work, or functioning at a work-capable intellectual level. This means the actual score range of the test (0-40) is likely to be offset against a widely normative sampled ability/IQ test.

- So, what we have to do (*solely in order to make the simulation more realistic*) is recalculate the expected correlation between the full GMA test range as given (0-40), and the CPC scores, given the GMA score range is actually an attenuated range over the whole ability range. This puts the likely applicant threshold score at about 1.34 SDs below the mean, equivalent to placing a cut-score of **12** into dataset #1. Doing this yields an expected correlation of 0.55.



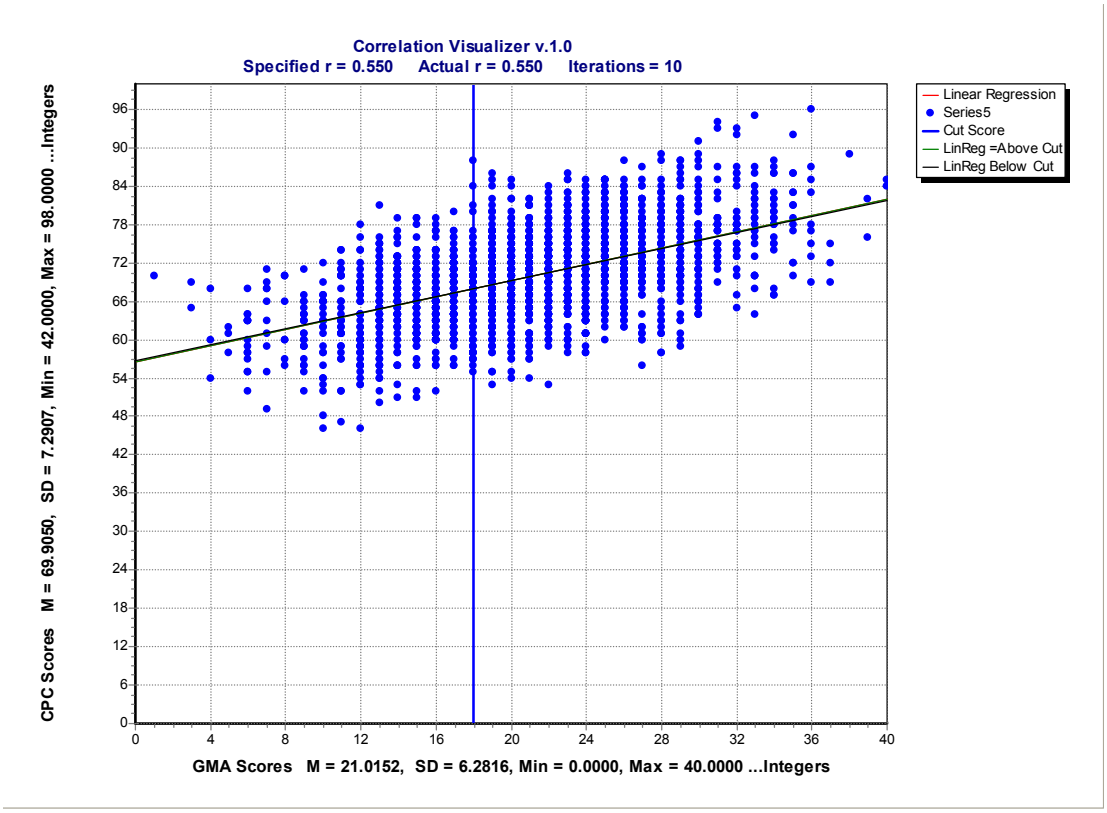
Step 2

Now, we have our more likely population estimate of GMA vs CPC correlation, I create **dataset #2** – which possesses the same score ranges for GMA and CPC, but with an expected correlation of **0.55.** This is the dataset we will now work with.

- Given we are also dealing with the fact that the incumbent sample of n=210 employees has been through a selection procedure, and given the semi and skilled nature of the positions we are selecting for, it is reasonable to assume the selection procedure might have been selecting individuals at a cut-score of **just below the average GMA test score**, in the population norm sample scores on the GMA test.

- A proposed cut-score of 18 would theoretically select-out approximately 31% of the likely applicant-standard general public. The organisation would be thus selecting about 70% of candidates.

- Using the population sample in dataset #2, and imposing a cut score of 18, the attenuated correlation between GMA and CPC scores in the restricted range (18-40) normative sample is **0.42** (instead of 0.55). This is calculated over 7077 cases. The "most-likely" score distribution we are now dealing with is that between the cut score of 18 and 40.



- We assume the n=210 incumbents have been selected from candidates, who, if they had been given the GMA test, would be seen to score 18 and above.

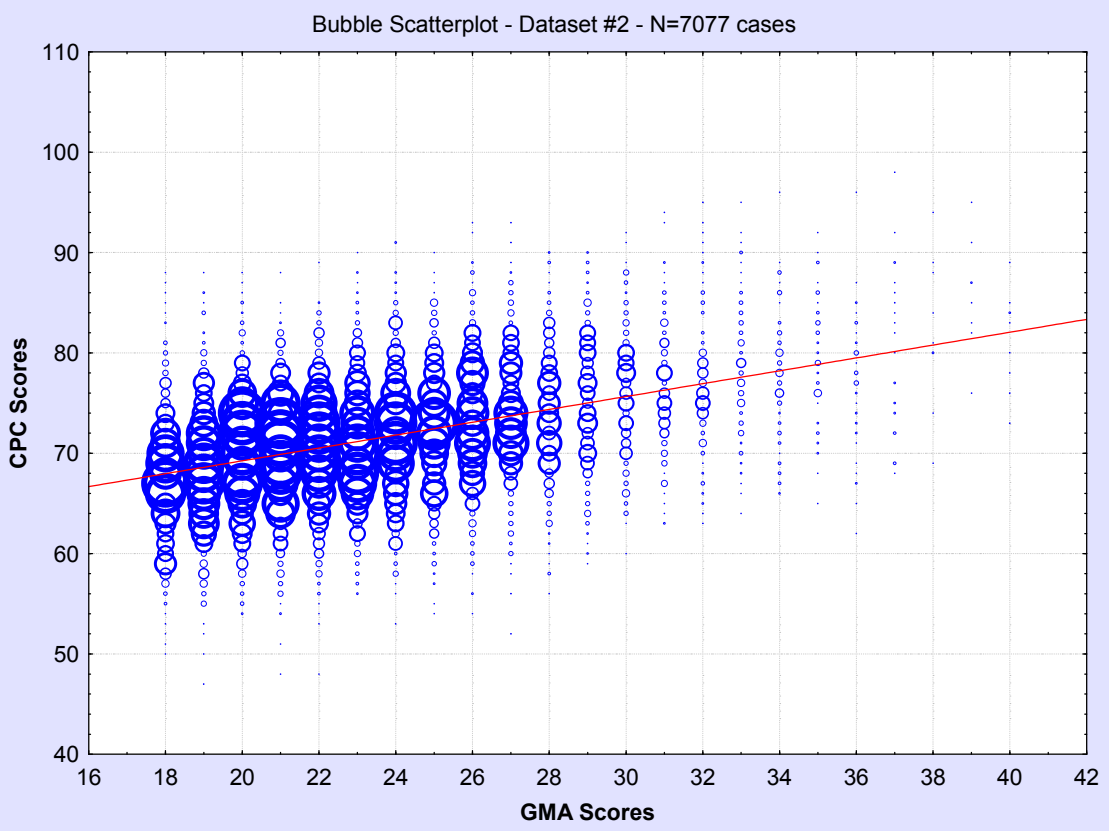
- The organisation now, in an effort to bootstrap norms, creates the selection-relevant raw-score to sten-score conversion table to be used for all future applicants.

Step 3

So far, we now have a population estimate of GMA vs CPC correlation where a likely GMA score is hypothesised to about 18 and above in pre-selected job incumbents. We now ask our 210 pre-selected incumbents to complete the actual GMA test.

- What we are likely to find is that the incumbent GMA test scores reflect a slightly skewed GMA distribution, in that a greater proportion will score nearer the cut-score of 18 and the overall mean of 21 than in the upper 20s. Remember, these 210 raw scores will be used to form the initial raw-score to sten-score conversion table as part of the "norm bootstrapping" procedure, so it's important that we create a realistic set of scores, not just a uniform randomly sampled set.

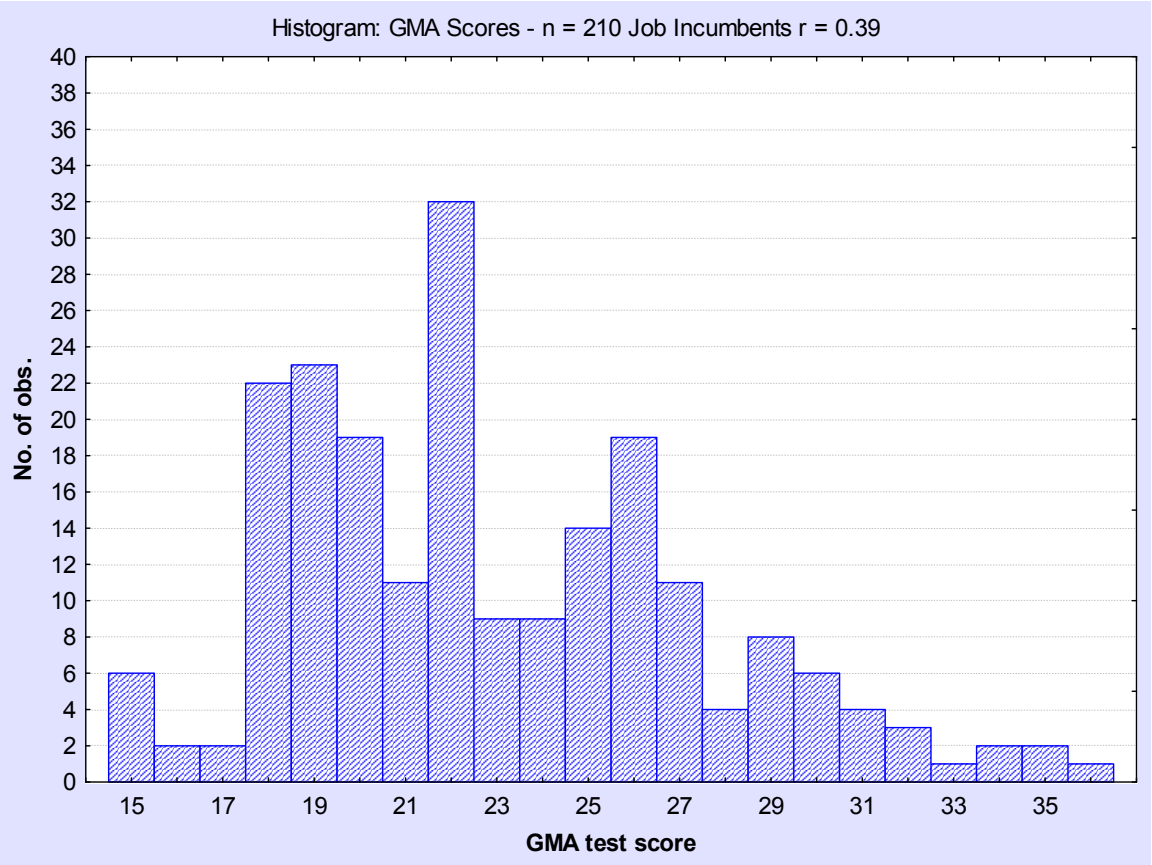
The scatterplot/bubbleplot below shows the effect I'm referring to. The larger the bubble, the more cases contained at that score-point. As this shows, the bulk of cases are around the lower GMA scores.



- So, in order to achieve a reasonable sampling, I'll acquire test scores in the range 18-40 of dataset #2 according to a truncated normal distribution sampling function rather than the usual "uniform" random number sampling. This will result in more randomly sampled cases being around the mean value of 21, but only selecting cases who score 15 and above, and limiting the number of cases who score below 18 to 10 or fewer (~5% of the total sample of n=210 cases). This mimics what might happen in practice, where a few individuals might have been selected using the older procedure, but in fact score lower than expected on the new test.

- Rather than sample from a hypothetical continuous bivariate normal distribution, I'll subsample from the 10,000 case dataset, as this is constructed using the kinds of integer valued variables we need to use, and these integer valued-variables do correlate at the required 0.55 in the "population". The sampling program was written in Delphi-2006 for speed and flexibility. It used uniform "sampling with replacement" from within the set of scores associated with a normally sampled GMA score.

The 210 cases so sampled show the following histogram of test scores, with an observed correlation between GMA and CPC scores of **0.39** (the expected correlation remember was 0.42 in this GMA score range attenuated dataset).



Variable	Descriptive Statistics (tester.sta)						
	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.	Skewness
GMA Scores	210	22.97	22	15	40	4.499	0.704

These data now form the normative reference sample against which all new candidates' scores will be referenced, and expressed as sten scores …

Step 4

Using the program Stanscore 4, (http://www.pbarrett.net/Stanscore/Stanscore4.html), we create the raw-score-to-sten lookup table for the 210 job incumbents …

Raw Score Range	Sten Equivalent
0 to 15	1
16 to 17	2
18 to 18	3
19 to 20	4
21 to 22	5
23 to 25	6
26 to 27	7
28 to 30	8
31 to 33	9
34 to 40	10

- So, on the basis of examining the CPC scores, the organisation decides to use **sten 4** and above as a selection threshold.

- This means raw–scores of **19 and above** will be used as the selection threshold for future applicants.

Of interest, if we had used the slightly constrained "applicant" general population norm of dataset #2 (10,000 cases), the dataset which contains the "general public" who might reasonably be expected to apply for an organizational position, the raw-score-to-sten lookup table is:

Raw Score Range	Sten Equivalent
0 to 8	1
9 to 11	2
12 to 14	3
15 to 17	4
18 to 21	5
22 to 24	6
25 to 27	7
28 to 30	8
31 to 33	9
34 to 40	10

Now what we see is that our "selection sten" of **4** encompasses raw scores **19-20** in our selection sample, while in the total expected applicant population it encompasses a range of **15-17**. There is not much we can say about this fact – but it does show that the same sten indicates different raw score magnitudes relative to which reference sample is being used for score transformation.

Let's now see what happens when a new sample of applicant data is added to our current N=210 norm dataset – bootstrapping the norm from 210 to 500 cases.

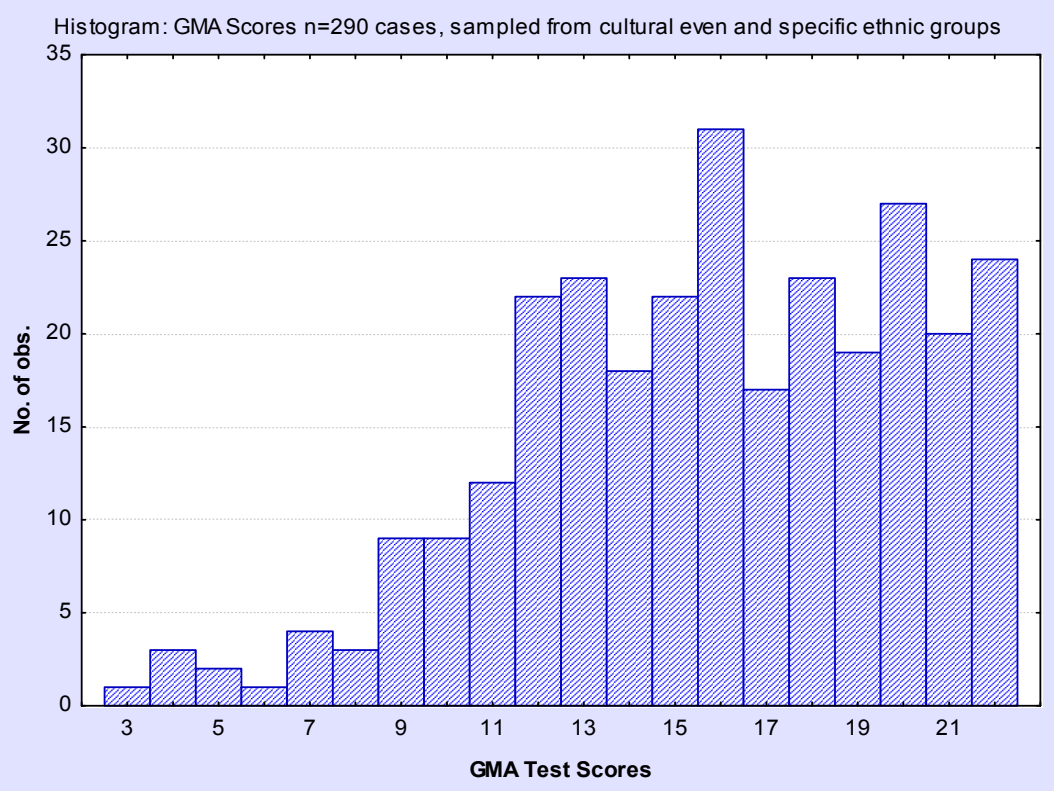
The additional 290 cases have come from two "recruitment" exercises:

- a local event which is known to attract many young people from a wide geographical area.
- a recruitment drive in unemployment job centres.

The pre-screen attracts 290 cases, who are scored against the n=210 norms. Then, the 290 are added to the existing 210 case norm to form the new n=500 case norm.

Unfortunately, no one actually looks carefully at the "before and after" raw-score distributions, or the raw-score-to-sten lookup tables. If they had, they would have seen the following:

The n=290 applicant raw score histogram:



Variable	Descriptive Statistics (tester1.sta)						
	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.	Skewness
GMA Scores	290	15.79	16	3	22	4.258	-0.487

The correlation between the GMA and CPC scores in this sample is: 0.37.

As can be seen from the above histogram, the majority of applicants just happened to show lower GMA scores. This is likely due to the specific sampling characteristics and demographics of this sample; it was not a "random sample" from the potential population of GMA scores (assuming even the constrained population of dataset #2.

The new raw-score-to-sten lookup table, **using an n=500 case norm** is:

The new "Sten 4" cut-score

Raw Score Range	Sten Equivalent
0 to 7	1
8 to 10	2
11 to 13	3
14 to 16	4
17 to 18	5
19 to 21	6
22 to 24	7
25 to 27	8
28 to 30	9
31 to 40	10

The **original n=210 norm table** was:

The old "Sten 4" cut-score

Raw Score Range	Sten Equivalent
0 to 15	1
16 to 17	2
18 to 18	3
19 to 20	4
21 to 22	5
23 to 25	6
26 to 27	7
28 to 30	8
31 to 33	9
34 to 40	10

Now you see the problem I hope. If the organisation fixates on using a sten of 4 as a selection threshold, whilst augmenting norms "blind", then the above can easily happen.

The new selection norms now mean that selecting candidates with sten 4 would result in selecting candidates who would have been rejected using the earlier raw-score-to-sten norm table.

We are now selecting “in” candidates who would have originally been assigned stens of just 1 or 2 using the initial normative dataset.

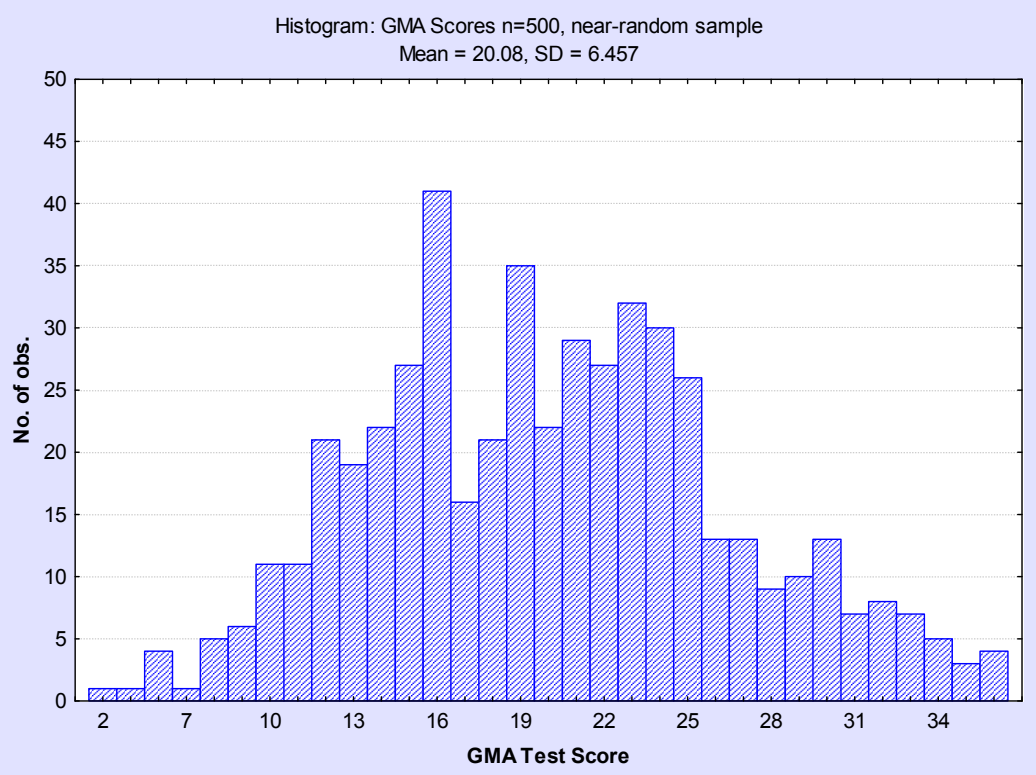
The result will be seen in a greater failure rate of training, and a lowered relationship between criterion outcome and sten score.

The opposite might happen if normative scores are augmented with non-randomly sampled applicants say from a university applicant recruitment drive. Augmenting norms with a group such as this would likely drive up the normative raw scores and cause “sten 4” to be associated with a much higher raw test score range. This means that in any future recruitment drive, fewer than expected “general public” applicants would attain sten 4.

It might be that new data which is used to supplant older norms matches the distributional characteristics of the older norm group. If this is the case, no changes will be observed in the new norm lookup table. But what if it isn’t? Let’s look at the consequences of using this new augmented norm of 500 cases on the next cohort of applicants …

Selection-Relevant Consequences

We now pre-screen 500 new applicants using the new raw-score-to-sten lookup table ... We'll assume this sample is more random than the last, being drawn from a wide range of recruitment sources. The GMA scores in this sample are distributed as:



Using the two norm lookup tables on these applicant data:

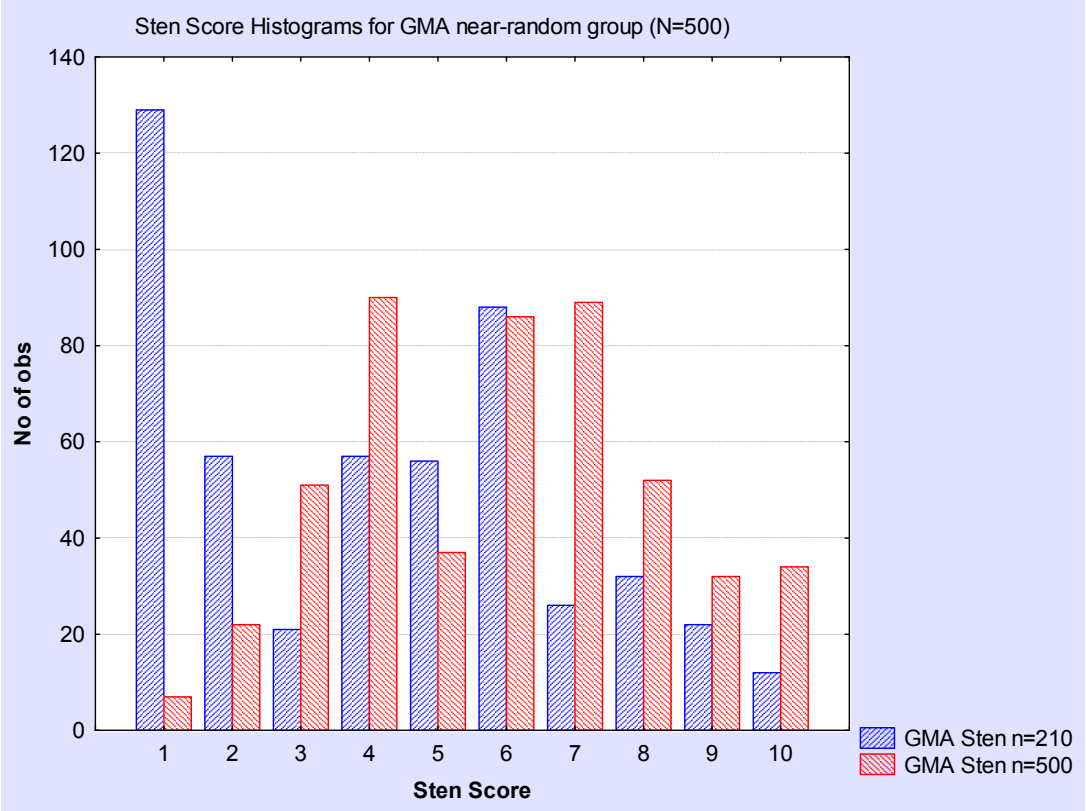
	Number of applicants selected with sten => 4	Mean CPC Performance Score	Number of selected applicants with greater than average (70) CPC performance scores.	Number of selected applicants with greater than 1 std.dev. (78) CPC performance scores.
N=210 Norm Table	293	71.3	163	36
N=500 Norm Table	420	69.8	193	38

The important feature of this table is that (420-293) = **127** extra candidates are selected for training who, using the original job incumbent norm of n=210, would not be considered as "satisfactory performance outcome candidates". i.e. a noticeable increase in the numbers of candidates expected to do well because they pass the sten 4 threshold, **but who now seem to fail more often in training**, would be the expected outcome.

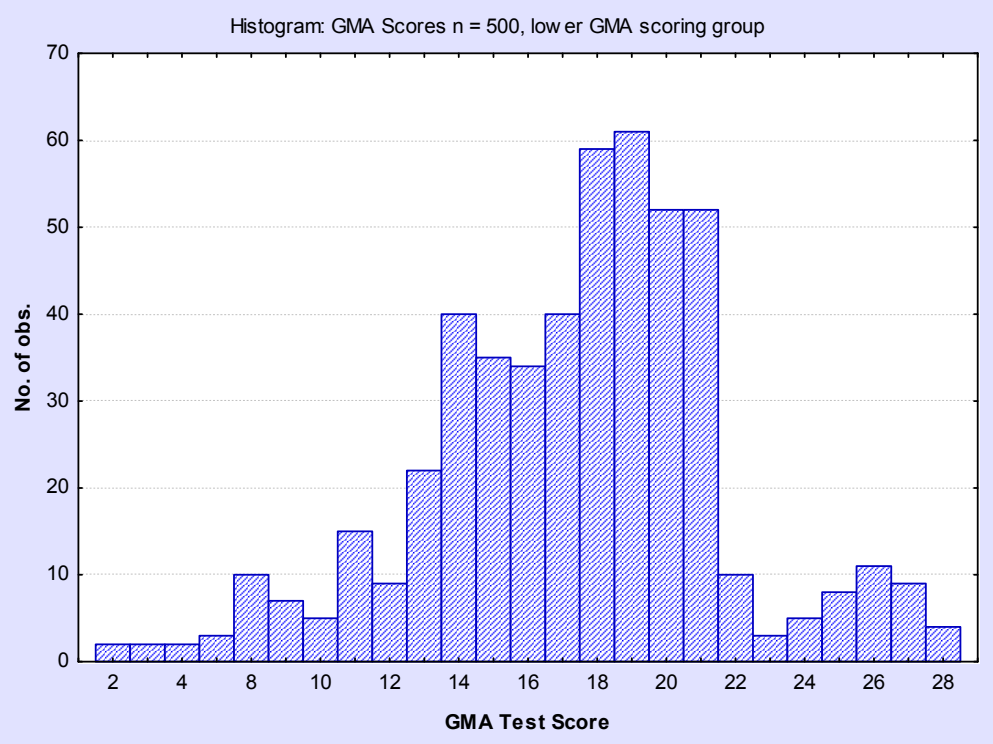
The worst case scenario is that few higher GMA candidates appear in any candidate sample, leading to a swamping of the "successful applicant" pool with those who are actually lower performance-outcome candidates. If the employer then adds these applicants to the norm group, things just go from bad to worse.

The critical calculation here is the **cost of failure of training**, in addition to the actual cost of training relative to the budget set aside for training. i.e. it may only be physically possible to "take on" 200 candidates at a time. If you choose the first 200 out of the 420 applicants (who just happen to contain more of the poorer GMA score candidates), then nearly all might subsequently fail training in that group.

The figure below shows the histograms of the two sets of sten scores for the 500 cases … with the effect of the lower GMA scorers in the n = 500 normative sample now plain to see ...



Let us now look at another dataset, which includes a greater proportion of low GMA scorers. Again, the sample size of 500 … and remember, the average GMA score in the population is 21 …



Using the two norm lookup tables on these lower scoring applicant data:

	Number of applicants selected with sten => 4	Mean CPC Performance Score	Number of selected applicants with greater than average (70) CPC performance scores.	Number of selected applicants with greater than 1 std.dev. (78) CPC performance scores.
N=210 Norm Table	215	69.4	85	17
N=500 Norm Table	423	68.4	137	25

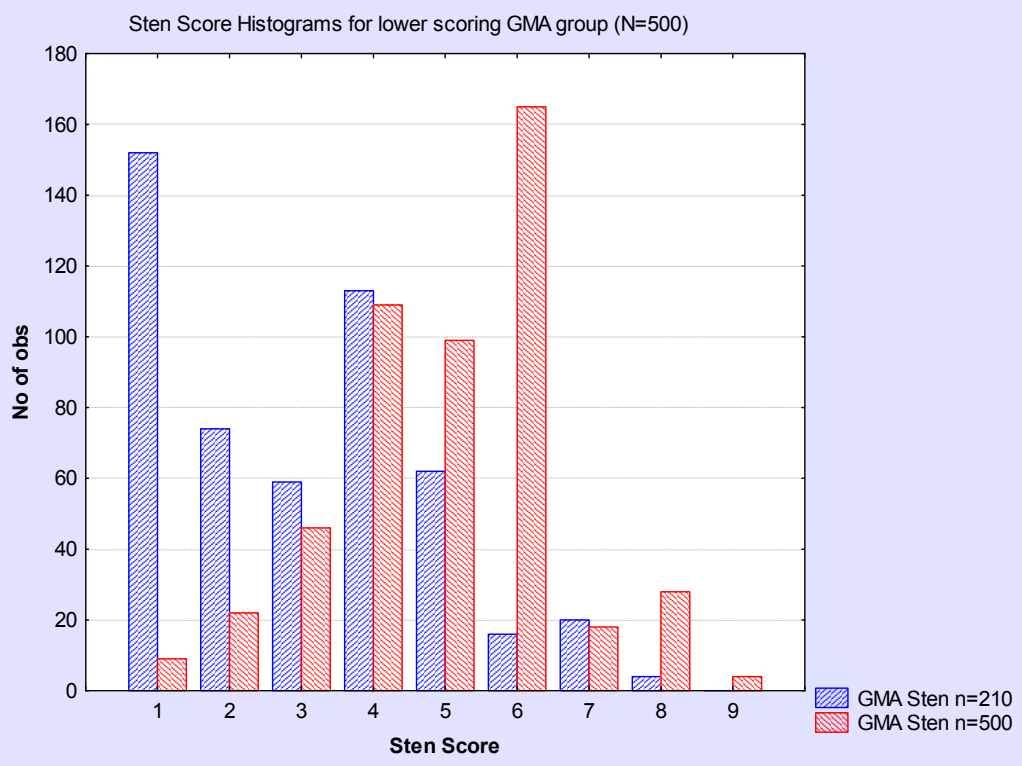
The important feature of this table is that (423-215) = **208** extra candidates are selected for training who, when using the original job incumbent norm of n=210, would not be considered as "satisfactory performance outcome" candidates.

Put another way:

- using the original n = 210 norm group, out of a total new candidate cohort of 500, we select 215 as "likely to succeed at training", with a 40% success rate.
- using the newer n = 500 norm group, out of a total new candidate cohort of 500, we select 423 as "likely to succeed at training", with a 32% success rate.

If we maintained the success rate of the group selected using the n=210 norm, and selected 423 applicants from a larger applicant pool, then we would expect to see 169 instead of 137 successful (above average) successes.

Again, to show the effect of the respective norm group on the resultant sten score distributions, I have transformed the dataset scores into stens using both normative lookup tables …



As can be seen, the number selected with sten scores of 4 and above are far higher using the augmented n=500 norm (which now contains many lower scoring GMA applicants) than when using the original norm group of n=210.

So, how do you guard against this potential problem?

1. Don't use norm-referenced scores; work with raw or convenient, linearly constructed, unified-metric scores.

2. If you insist on using bootstrapped norms, then look very carefully at each potential norm update sample for the effects of such an "augmentation" on the selection norms and resultant transformed scores. You may have to select a new sten cut-score for the duration you use that norm group.

3. Also, if you insist on using norm-reference scores, keep a constant check on the predictive accuracy of the selection test sten scores and the performance criterion. If this changes substantially from cohort-to-cohort, then this is also a "give-away" that something is happening at the raw-score level.

4. Never allow a test publisher/provider to hide the raw scores from you. In many cases the norms are updated by the test provider as a "service" to the client. Whilst convenient, it means you cannot investigate the potential consequences on selection outcomes each time the norms are updated.

5. In the end, I can only repeat, don't use norm-referenced scores in selection scenarios. What is the point of paying extra financial and analysis costs incurred in making sure you don't make the kind of mistakes which are impossible to make when using raw scores?

References

Ægisdóttir, S., White, M., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R.S., Nichols, C.N., Lampropoulos, G.K., Walker, B.S. & Cohen. G. (2006) The meta analysis of clinical judgment project: fifty-six years of accumulated research on clinical vs statistical prediction. *The Counseling Psychologist*, 34, 3, 341-382.

Grove, W.M., & Meehl, P. (1996) Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures. *Psychology, Public Policy, and Law*, 2, 2, 293-323.

Grove, W.G., Zald, D.H., Boyd, S.L., Snitz, B.E., & Nelson, C. (2000) Clinical vs mechanical prediction: a meta-analysis. *Psychological Assessment*, 12, 1, 19-30.

Salgado, J.F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J.P. (2003) A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology*, 88, 6, 1068-1081.

Schmidt, F.L., and Hunter, J.E. (1998) The Validity and Utility of Selection Methods in Personnel Psychology: practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 2, 262-274.

Schmidt, F.L., & Hunter, J. (2004) General mental ability in the world of work: occupational attainment and Job Performance. *Journal of Personality and Social Psychology*, 88, 6, 162-173.