http://www.cs.utep.edu/admissions/model.pdf

A Model of Computer Science Graduate Admissions Decisions

Version 1.3 September 21, 2004

Nigel Ward
University of Texas at El Paso
nigelward @ acm . org

0 Overview

This document presents a model of computer science graduate admissions decisions which enables the rough prediction of many accept/reject decisions. It explains how it works and why it was built this way.

It describes Model Version 1.0, Parameter Set Version 1.0 and Interface 1.6.

The motivation for the model is presented briefly in "The CS Graduate Admissions Estimator: About", at http://www.cs.utep.edu/admissions/about.html .

Contents:

1 Key Simplifying Assumptions

The first assumption is that the strength of any applicant can be represented by a single number, referred to below as the GQ score.

I also pretend that there is a single, universally shared way to evaluate CS graduate school applicants. Differences in the evaluation criteria that various departments present on their web pages are taken to be only a result of random omissions and other noise. This implies that the admissions policy of any school can be described with a single number, the GQ threshold.

I also assume that a model need not be perfect to be useful. Although the main features of the model are fairly well justified, there are a number of parameters whose values are somewhat arbitrary and should be re-examined for future versions.

2 Modeling Applicant Strength

2.1 Overall Model

The model here basically converts all the factors involved to one scale and combines them by a linear weighting, although the weighting is chosen differently in each case, to capture the actual importance or informativeness of each factor.

2.2 The Factors

This subsection considers each factor in turn, explaining how it is used in the model.

2.2.1 GRE Scores

To overview, the reduction of GRE scores to a single number involves four things. First, each score is normalized. Second, the contribution-level for each score is obtained by multiplying the height by ranking factor (depending on the profile of each applicant). Third, each contribution level is multiplied by the importance weight for that score. Finally, these values are summed and divided by the sum of the importance weights.

The units of the result can be considered to be "abstract GRE Q points above baseline", that is, the number of Q points above baseline that one would expect for an applicant with the same overall GRE strength and a perfectly balanced GRE profile.

Use in the Model First the GRE sub-scores are normalized to the same scale. This is done using a baseline value of V 500, Q 650, A 600, and AW 4.0. Scaling is equivalent for V, Q, and A, but for the AW the score is multiplied by 125.

Thus, the normalized value NV_i is given by:

$$NV_i = (RV_i - BV_i) \times SF_i \tag{1}$$

where RV_i is the raw value, BV_i is the "baseline value", and SF_i is the scaling factor.

For example, a GRE profile of V 600, Q 650, and AW 4.5 would normalize to NV_V 100, NV_Q 0 and NV_{AW} 62.

Second, the normalized subscores are ranked, that is ordered by size. and multiplied by a ranking factor to obtain the "contribution levels" (CL_i). 1 is the rank of the highest normalized value, 2 the rank of the second highest, and so on. Continuing the above example, $R_V = 1$, $R_{AW} = 2$, and $R_Q = 3$.

The ranking factors RF_i range from 1.33 for the lowest normalized value (representing the ability likely to be the limiting factor in the applicants success) to .67 for the highest. Specifically, the ranking factor $RF_{r,n}$ is given by

$$\frac{2}{3}(1 + \frac{r-1}{n-1}) \tag{2}$$

where r is the rank order of that normalized score, and n is the number of scores ranked. RF is used as a multiplicative factor (but note that this is only done for the numeric data, namely the GRE scores and the GPA):

$$CL_i = NV_i \times RF_{r,n} \tag{3}$$

Summarizing the example so far:

	raw RV	normalized value NV	rank R	ranking factor RF	contribution level CL
verbal	600	100	#1	.67	67
quantitative	650	0	#3	1.33	0
analytical writing	4.5	62	#2	1.0	62

Third, each contribution level is multiplied by an importance weight, namely .7 for V, 1.0 for Q, 1.0 for A and .7 for AW.

Finally these products are summed and then brought back to the scame scale by dividing by the sum of the importance weights.

$$CGRE = \frac{\sum_i CL_i \times IW_i}{\sum_i IW_i} \tag{4}$$

Continuing the example, (.7 * 67) + (1.0 * 0) + (.7 * 62) / (.7 + 1.0 + .7) = 38.

Rationale The problem of taking the three GRE scores and combining them into a single number is trickier than it may appear. Sampling the web, there appear to be two common methods: sum and minimum. Some schools publish a target total score for the GREs, reflecting the idea that the overall strength of a candidate is the average of his individual strengths. This suggests an additive method. More schools, however, publish a set of GRE scores. This is often presented as a baseline GRE profile, giving a desired or likely expected score-set for the sort of person likely to succeed in the graduate program. This suggests a minimum-based method, where the weakest score is what matters. This reflects the idea that a good applicant is one who is strong across the board; presuming that "balanced" individuals do better.

Sophisticated admissions committees probably use something in between these two methods, as will be discussed below.

In order to measuring "balanced" strength it is necessary to make comparisons across the scores. This is done by normalizing as described above, where the baseline values are chosen to represent the GRE profile of a well balanced candidate. The values for the baseline on each scale were chosen based on the observation that, for schools which publish minimum scores, the typical difference between Q and V is 150 or a little more, and the difference between Q and A is 50 or a little less. There is less information on the use of AW, but a 4.0 is occasionally listed in contexts which list Q, V, and A in these ranges. These baseline values happen to be the values formerly specified as "desired values" by UT El Paso. Interestingly these scores are all close to the 60th percentile values for each scale[1] except that the AW baseline is lower, which may be appropriate because engineers are known to do relatively poorly on it.

[1]ETS: Guide to the Use of Scores. `ftp://ftp.ets.org/pub/gre/994994.pdf`

While it might be better to use consensus (average) differences to establish the baseline, the model uses round numbers, since it is important that the code be easy to explain and verify, especially for Version 1. The absolute values for baseline do not matter, and were chosen without much thought, although I later discovered that people don't like to be given ratings that are negative numbers.

Following common practice, as reflected by the fact that many schools specify that they just add up V, Q, and A, points on these three sub-scores are given equal value, that is, the scaling factors are all 1. Regarding AW, to align the entire range would require treating 1 AW point as equivalent to 100 points of V, Q, or A; to match up a perfect V and a perfect AW would require a factor of 150; but the GRE score distributions suggest that 125 is an appropriate value, and that is what is used.

In order to compute something between the average and the minimum, and thus prefer more balanced applicants, the algorithm gives more weight to relatively weak scores. This means that a relatively low score is taken much more seriously than a relatively high one. Specifically, for a strength on one dimension to outweigh a weakness on another dimension, the strength has to be twice as large as the weakness. This is done by the ranking factors. The formula for ranking factors, Equation 2, is chosen to meet three needs: the sum of all the ranking factors is n (thus the total importance of all the numeric scores taken together is not altered), the contribution of the lowest normalized score is 2 times that of the highest, and scores of intermediate ranks are counted to intermediate degrees.

This method is in fact an "Ordered Weighting Averaging Operator"; the application to the problem of evaluating student applicants was proposed by Carlsson et. al[2].

Now there comes the question of how much importance to give the various scores: the "importance weight". Note that this is different from the scaling factor: even though a 600 V might have the same normalized value as a 750 Q, the Q is probably more important and reliable than the V as an indication of graduate success. The importance weights are based in part on frequency of mention: all graduate schools which give GRE numbers mention a Q score, so this is clearly very important. It is interesting to note that ETS reports that V, Q, and A are approximately equally correlated with graduate school first year grades. However accepting students who will earn good first year grades is less important to most admissions committees than accepting students who are likely to doing good thesis work and to graduate, and this, I guess, probably correlates more with Q and A than with V and AW.)

2.2.2 Grades

Use in the Model The undergraduate GPA is normalized into the same scale as the GRE sub-scores, and then ranked and weighted in in the same way.

For US GPAs the baseline value is 3.3, the scaling factor is 200, the importance weight is 2.5, and the ranking factor is computed as part of the computation of the ranking factors for the GRE sub-scores. The use of this scaling factor means, for example, that a 1 point higher GPA is taken to be on the same level as a 200 point higher GRE-Q.

Extending the above example, if add the additional information that the applicant had a GPA of 3.9, the normalized value is NV_{GPA} = (3.9 - 3.3) * 200 = 120.

[2]OWA Operators for doctoral student selection problem, Christer Carlsson, Robert Fuller and Szvetlana Fuller, in *The Ordered Weighting Averaging Operators: Theory and Applications*, Ronald R. Yager and Janusz Kacprzyk, eds., Kluwer 1997, pp 167-177

Now, since this is adding another numeric score, the ranking factors change, to .67 for NV_{GPA}, .89 for NV_V, 1.11 for NV_{AW}, and 1.33 for NV_Q. The contribution levels would therefore be 80 GPA, 89 verbal, 0 quantitative, and 70 analytical writing.

The total combined score would be
(2.5 * 80) + (.7 * 89) + (1.0 * 0) + (.7 * 70)/ (2.5 + .7 + 1.0 + .7) = 64.
If an "In Major or Recent GPA" is given, the value used for the above computation is the average of this and the overall GPA.

Rationale The GPA fits easily into the same computation used for the GREs. The algorithm uses 3.3 as the baseline GPA; this corresponds to the baseline GRE profile, meaning that perfectly balanced applicant with a 3.3 GPA would also be expected to have the baseline GRE scores.

A few schools specify that the minimum or desired GPA is computed not over all courses, but over "CS courses" or over "Courses over the past two years". Even schools which don't explicitly mention this also probably like to see a strong GPA in CS and in recent courses. The reason for providing only one data-entry box for these two numbers is that I think most admissions committees attempt to see each applicant in the best possible light, and will make their decision while looking at whichever of these is strongest.

A scaling factor of 200 Q points for one GPA point is used. This is done, as before, to determine whether a given GPA is consistent with, higher than, or lower than the GRE scores. Choice of the scaling factor here is tricky. One technique is line-fitting, based on those schools which report both GRE and GPA numbers in their admissions data; a glance at the data suggests that about 200 is appropriate. A second technique is to look at admissions websites which give explicit GPA-GRE scaling factors. Although no CS admissions websites give such information, other sites give values ranging from 200 to 400, with more clustered at the low end. A third technique is the simple observation that for CS applicants the GPA and the GRE-Q tend to max out together, and so a GPA of 4.0 can be taken to map to a Q of 800: given the choice of baseline above, with GPA of 3.3 and Q of 650, again a scaling factor of 200 seems about right.

The importance of the GPA relative to the GRE is another tricky issue. The ETS advises that undergraduate grades are the best single predictor of graduate success. It is also the case that undergraduate curricula in CS are fairly standardized, making undergraduate GPA even more likely to be a reliable predictor. Moreover, the undergraduate GPA reflects abilities such as programming, system integration, etc., which are not measured by the GRE at all. On the other hand, relatively few schools publish an expected GPA score, probably because grading standards are not consistent. This algorithm gives the GPA a weight of 2.5, making it roughly as important as all the GRE scores combined.

GPAs not on the US system can be handled one of two ways. One way is to convert to the equivalent US GPA, using for example the correspondences presented in Duke University's International Credentials Guide [3] and Colorado State's International Credential Guidebook[4]. The second way is to work directly from the raw GPA, using country-specific or school-specific baselines and scaling factors. While I only have enough data for a handful of schools, it's worth doing this anyway to illustrate the problems involved.

One might think that the needed parameters for each school could be estimated from an applicant database, taking the baseline from the average GPA for students near the baseline CGRE, and the scaling factor from the slope of the regression line of GREs on grades. However this doesn't

[3] `http://www.gradschool.duke.edu/Forms/Credentials.pdf`
[4] `http://graduateschool.colostate.edu/files/FormsPubs/CredentialGuidebook.pdf`

	baseline	scaling factor	importance weight
US	3.1	200	2.5
India			
default	.70	400	2.0
JNTU	.72	450	2.5
Madras	.70	450	2.5
Mumbai	.58	450	2.5
Nagpur	.60	300	2.3
Nagarjuna	.72	300	2.3
Osmania	.73	400	2.5
Viswesariya Tech	.72	300	2.3
Mexico	8.7	200	1.5
Other	3.3	300	1.0

Table 1: GPA Parameters

work, because applicants tend to self-select. For example, UT El Paso gets no applicants with both stellar GREs and GPAs, as they go elsewhere. It also gets no applicants whose GREs and GPAs are both weak; they also go elsewhere.

The method used here is to estimate the slope from two points, the cluster of typical UT El Paso applicants from that school and the ASU cutoff for applicants from that school (data courtesy of Chitta Baral).

For example, for Jawaharlal Nehru Tech. U. (JNTU), there seems to be a cluster of graduates with a GPA around .68 and a CGRE around -20: this is the first point. ASU uses a cutoff of .81; taken with a CGRE of around 35 (computed from their GRE minimum profile of 400V, 700Q, and 650AW as explained below in Section 3.3.2) this gives the second point. These two points give a slope of around 450 GRE points per GPA point, and a baseline intercept of .72. Parameters for the other Indian schools in the table were estimated the same way.

The Mexican GPA scaling was chosen based on published conversion factors, and the baseline was initially estimated the same way, with subsequent fine-tuning to produce predictions which better matched a handful of decisions made by a experienced admissions committee.

The importance weights also vary. This reflects two factors. The first is that undergraduate success in certain countries is probably less predictive of US graduate success. The second is that less importance is given where the parameters are known only approximately. These lower importance weights mean that for most foreign applicants the GREs contribute relatively more to the total score, as common sense would suggest. Table 1 summarizes.

2.2.3 Letters of Recommendation

Use in the Model The weight given to letters is simply the product of the believability score (entered under "recommender is a ... ") and the basis-for-judgment score (entered under "observing you as a ..."). Normalization of the value (the "warmth", entered under "describing you as ...") is done by treating 2 as the baseline and using a scaling factor of 100. No ranking factor is applied.

Thus for example a warmth of 2 corresponds to a GPA of 3.3, a warmth of 3 to a 3.8 GPA. An applicant whose letters were .7 on the believable scale, 2.5 on the basis scale, and 3.5 on the warmth scale, would therefore have and importance weight for letters of 1.75 and a normalized score of 150.

Rationale In general, the value of letters is mostly that they report on abilities not well measured by the GPA or the GRE.

(Of course, the role of letters in the admissions process is occasionally quite complex. In particular, there are two types of letters which are influential but very hard to quantify. The first type are insider letters from a faculty member of the department being applied to. These typically testify as to how the applicant can meet the department's specific needs, and include a promise to work with and support the applicant. The second type are letters which somehow put the hard facts (the GRE and GPA numbers) in a different light. I assume that students receiving these types of recommendations are close enough to the recommender to ask for specific guidance on where to apply, making use of this estimator unnecessary.)

Although many applicants have several letters of recommendation, the discussion here assumes that all the information can be unified and treated as a single letter. The content of this macro-letter is represented with three factors solicited from the user, as described in the instructions at the bottom of the Data Entry page[5].

Excluding unusually influential cases, I estimate that a letter can count for, at most, as much as the GPA, and for as much as the GREs, although not as much as both together. Thus letters are given a maximum importance weight of 3. However this weight is only achieved when the letter writer is someone whose words the committee will believe, and who knows the applicant well, which is why the believability and basis for judgment are multiplied. This is an important feature of this estimator: in giving a likely upper bound on how much a letter can help out it may help applicants plan more realistically.

The "recommender is a ..." menu selection is intended to predict the believability of the letter. This can be broken down into 3 sub-factors.

The first sub-factor is the perceived trustworthiness of the person who is writing the letter. This is an issue because recommenders may tend to exaggerate and help out the student. Trust is hard to model, but admissions committees probably have more confidence in letters written by people who they know personally. They probably also have confidence in famous people, who have a reputation to lose by writing uncritically warm letters. They also probably trust more people who move in the same circles. Somehow it seems that someone you might meet at a conference, or who might share a mutual acquaintence, is more trustworthy. Conversely, people in very different fields or in distant countries are probably harder to trust.

Incidentally, I think the trustworthiness issue is the main reason why some schools provide checkboxes to indicate whether the applicant is in the top 20%, 10%, 5%, etc. I doubt that such numbers are suitable for directly feeding into a formula, at least not without statistics on the distribution of ability levels in various populations. However the presence of such checkboxes probably does help remind the recommender not to immoderately extol the virtues of the applicant.

The second sub-factor is the specificity of the letter. While committees don't expect the recommender to provide evidence to support every comment, specifics do help demonstrate that he knows what he's talking about and is thinking rationally.

The third sub-factor within believability is whether the recommender knows what it takes to succeed in graduate school. Your manager at work may be a perceptive, wise, and trustworthy person, but still not know much about what it takes to do well in graduate school in CS.

[5] `http://www.cs.utep.edu/admissions/entry.html#instructions`

To keep things simple, these three sub-factors are conflated into a single believability index ranging from 0.0 to 1.0. This is presented in the data entry screen as a menu with a scale anchored by some common cases.

Turning now to the second factor determining the weight given to a letter, it is clearly important that the recommender actually knows the applicant. More specifically, a letter is more weighty if the writer has observed the applicant when doing more of the things that graduate students have to do. (Multiple letters are often useful as providing information on more of the applicant's activities.) This factor, the "basis for judgment", is in the range from 0 to 3.0. Since this feeds into the importance weight, it means that applicants whose recommenders know little about them, or who have no research experience, are not penalized. Rather letters with a weak basis simply have little effect.

Although letters have an importance weight they do not have a ranking weight. This is because, unlike GREs and the GPA, I don't believe most admissions committees have clear ideas about the minimal level of letter that they want to see. Although the undergraduate experience is uniform in many respects, students vary greatly in their interactions with faculty, and thus it is not possible to have strong expectations for what should be in the letters.

Thus for letters, and also for the statement of purpose as discussed below, the contribution level is just the normalized value:

$$CL_{letters} = NV_{letters} \tag{5}$$

$$CL_{statement} = NV_{statement} \tag{6}$$

Finally comes the question of what the letter actually says. I normalize "letter warmth" to the same scale as the GRE and the GPA, using a baseline of 2 and a scale factor of .4 GPA points or, equivalently 80 GRE points. Thus a "future CS hero" recommendation counts at the level of a putative 4.2 GPA.

The reason why letter warmth must be normalized to the same scale before it is added in is so that the final score will have a consistent interpretation whether or not letter values are specified.

2.2.4 Statement of Purpose

Use in the Model Here normalization is done using a baseline of 2 and a scaling factor of 100. Thus a strength of 2 corresponds to a NQ of 0 and a GPA of 3.3, a strength of 1 to a GPA of 2.8, etc. The weight is a fixed .5.

Rationale Statements of purpose are weighted relatively lightly, because most applicants find it relatively easy to figure out what the admissions committee wants to hear and to get help writing something that says that. (Of course, once in a while a statement of purpose is decisive. A truly bad statement can sink an otherwise strong application. A truly good statement can make the committee interpret grades, GREs, and letters in a new light. These rare cases are not handled by this model.)

I think there are two important things which are often documented only in the statement. The first is motivation, which of course is predictive of success in graduate school. The second is

realism, especially realism about what can be learned and accomplished in the graduate program at the target school.

For a letter to convincingly display motivation and realism, it helps if the narrative relates them to the applicant's own personality, abilities, and experiences. Most undergraduates don't have superior self-knowledge, nor strongly distinguishing experiences, but for older applicants these can be a real plus.

Although I think those are the main things which admissions committees look for in letters, there are also other factors. For example, some schools like letters to show that the student has carefully examined the target school's web page. Others like to see a match to faculty research interests. Others use letters as a way to examine writing skills, although these are also measured by the GRE-AW.

In the model all these things are conflated into one simple "strength" scale.

It's worth noting that the model could express warmths and strengths directly on the NV_Q scale and do without a baseline; or that it could use a scale with the importance-weights factored in. Either method would save one free parameter, but I think that the model is clearer presented this way.

2.2.5 Major

Non-CS majors are handled by modifying the baseline and importance weight used for the GPA. For engineering, science, and math majors the baseline is raised 60 GRE points (.3 US GPA points), and for other majors 120 points. The importance weight is reduced by 10% for engineering, science, and math majors, and 20% for other majors; this is because non-CS grades are proably less predictive of CS graduate performance.

Of course, non-CS majors may also have strengths, typically a clearer motivation, however these should be fully reflected in the letters and statement of purpose.

2.2.6 Fellowships

Use in the Model A full long-term fellowship, for example a 3-year NSF graduate fellowship, is counted as a straight bonus of 20 GQ points; a partial or one-year fellowship as 10.

Rationale Admissions committees usually have access to the same information about applicants that scholarship committees do, so having a fellowship is probably not an independent predictor of aptitude. The reason this factor is included in the model is that having a source of funding contributes to the likelihood of succeeding in graduate school. In reality this factor may be more important at more expensive schools.

Although it was possible to map letter warmth and statement strength to the basic GRE-GPA scale, this is clearly impossible for fellowships. Lacking a fellowship should not count as negative evidence for anyone, and having a fellowship should boost the GQ for anyone. Thus this is a straight bonus.

2.2.7 Target-Group Membership

Use in the Model These are also straight bonuses, of 20 or 50 GQ points, directly selected by the user.

Rationale Benefactors, legislatures, funders, and the general public all want universities to improve society, in various ways, and sometimes this affects admissions decisions. For example, at UT El Paso the university's mission includes strengthening the economy of the region, which implies that the Computer Science department must support local industry, which suggests that the admissions committee look favorably on applicants who are working in local companies or who are likely to stay in the area after graduation.

Another factor that may affect the admissions decision is the desire to assemble a student body which is not just a collection of people with individual merit, but is also strong as a group, with diversity in terms of experiences, skills, and demographics.

Modeling these group preferences is tricky. To keep things simple, I'll assume that there are two levels of preference involved: "encourage" and "give priority to". "Encouragement" can be hidden in the noise, so that it only applies if there are two "equally qualified" applicants, or, operationally, two applicants whose strengths are so similar, or so incomparable, that the admissions committee can claim to be unsure which really is better. Operationally, I guess this amounts to about a 20 point preference. The next level of preference is more active and obvious. There is probably a constraint here, namely that this not lead to an unmanageably bifurcated student body. I guess that for most schools this amounts to up to about a 50 point preference. This would be the value for a department with a clear mission reviewing an applicant whose group membership clearly is aligned with that mission, and which already has a high variation in ability levels.

Some of the preferences are probably additive — for example female in-state applicants probably get a double bonus — although this is limited by the fact that of course no school is going to accept unqualified applicants or those unlikely to succeed in the program.

2.2.8 Factors Omitted from the Model

There are a number of other factors which could be included in the model but are not, because they are less frequently important and/or hard to quantify.

Undergraduate School Prestige There are three reasons why this might be relevant. First, prestigious schools may tend to grade harder. Due to a lack of data this is omitted in this model. Second, prestigious schools may prepare students better for graduate school, although this is probably not a major consideration, as the CS curriculum is fairly uniform, at least at well known or accredited US programs[6]. Of course, prestigious schools may have more enrichment-type activities available, but participation in these will typically be reflected in the letters of recommendation. Third, prestigious schools are more selective on the input side. This is not a big consideration for US schools, where the typical undergraduate selection process considers no factors that usefully complement the information available to the graduate admissions committee, but it may be relevant for applicants from abroad. Lists of prestigious foreign schools may be found in Credentials Guides, as cited in Section 2.2.2.

[6] for a list, see `http://www.abet.org/accredited_programs/CACWebsite.html`

GRE Subject Test in Computer Science Although this is a very nice test, it seems not to be much used by admissions committees, especially for applicants with an undergraduate CS major.

TOEFL Score This seems to be treated only as a requirement by most admissions committees. (Common requirements are 550/213, 570/230, and 600/250, with a tendency for more selective schools to have higher requirements.) A weaker TOEFL score is generally fatal, but a stronger TOEFL probably does not contribute significantly to a student's evaluation.

While the TOEFL has the theoretical advantage that it does not presuppose a US education, in practice it appears to be partly redundant to the GRE Verbal, especially at the ability level expected for graduate school. Incidentally the Verbal-TOEFL correlation is reported as .64 for non-native speakers [7].

Nationality and Culture Americans generally belittle standardized tests, whereas students from Asian countries traditionally study seriously for them. Such differences may affect the correlation between GRE score and actual ability.

Native Language Some schools expect lower GRE verbal scores from international applicants.

Coursework Many schools expect specific preparation, especially for applicants without an undergraduate CS degree.

Same School Some schools disfavor their own undergraduates, as students may grow more if they do graduate work at a different school.

Research Match Some schools prefer applicants who will match the needs of the faculty, especially those faculty who are not overloaded. However this is probably not generally important for newly entering graduate students, who are expected to be adaptable.

Publications These can be a huge plus, of course, but the contribution is hard to quantify.

Other Factors Other factors, such as a non-traditional career path, an advanced degree in some other field, etc. may also influence some admissions committees.

2.2.9 Summary

$$GQ = \frac{\sum_i CL_i \times IW_i}{\sum_i IW_i} + FV + GP \tag{7}$$

Since all of the weighted inputs are converted to the same scale, it doesn't matter if values for some factors are omitted. However, to help minimize the number of misleading estimates, the implementation refuses to give an estimate until the user enters at least 3 numeric values, i.e. 3 out of the GPA and the GREs.

The result is dubbed "GQ", which stands for "Generalized Quantitive", or maybe "Guesstimated Quality". The units are GRE-Q points above baseline. Again, this is the number of Q points above baseline that one would expect for an applicant of equivalent attractiveness but a perfectly balanced profile in terms of GREs, GPA, and the other factors. In other words, if an applicant has a score of X quant points over baseline, and everything else is at the same level, then his GQ will also be X.

Of course the GQ score is not particularly useful by itself, but it permits comparison to the acceptance criteria at various schools.

[7] ETS Research Report 69 (`http://www.ets.org/research/dload/RR-02-16.pdf`)

2.3 Verification

The implemented model was run on a test set of 55 UT El Paso applicants from 2003-2004. This uncovered a few bugs in the parameters. After these were fixed, proper testing began. The first test was whether the system actually predicted the admissions decisions. In most cases it did, but I discovered that the UTEP admissions committee gives very low weight to Mexican grades, but considering the outcomes, I'm not sure that this was wise, so I've left the weight unchanged. The other prediction failure was due to a factor "omitted from the model" Section 2.2.8. Otherwise the model did well.

The second test was to compare the actual GQ scores to those produced by an earlier algorithm, Longpre's simple linear model. Deviations in rankings were identified, and the source of each was examined. No undesirable deviations were found; rather each was explicable in terms of a feature deliberately designed into the new algorithm. (Incidentally, the largest differences between the ratings by this model and by the simple linear model arose, in practice, for applicants with weak verbal scores and those with undergraduate degrees other than CS.)

3 Estimating Admissions Thresholds

Now, having seen how to compute a single number representing the applicant's strength, namely the GQ score, the next task is to see how to use that to predict likelihood of acceptance at a number of schools.

Recalling the assumption that the admissions policy of any school can be described with a single number, the GQ threshold, the aim in this section is to explain how to estimate that number for each school.

However this is complicated because schools report admissions criteria using a variety of methods: some report the average GRE of acceptees, some give a set of minimum GRE scores, others report a minimum sum of GRE scores, and so on.

3.1 Reconciling Incommensurate Descriptions

3.1.1 Preliminaries

Schools typically report values on a number of dimensions, such as GRE-V, GRE-Q, and GPA. In order to get anywhere these must be converted to a single value. Fortunately this can be done conveniently using the model already developed. That is, I assume that the selectivity of a department can be measured in the same way as the quality of an applicant.

Specifically, this means plugging in the values each department reports into the model and reading off the result. Since the parameters regarding GPAs and letters are less reliable than those for GREs, this was done only using GRE values: thus the results discussed in this section are all CGRE values.

There is, however, one big problem: no school, except for UT El Paso, reports an actual GQ or CGRE threshold number.

So it is necessary to estimate the GQ threshold from other numbers that are published. This subsection discusses two touchstone reporting methods: "minimum GRE scores and "average GRE

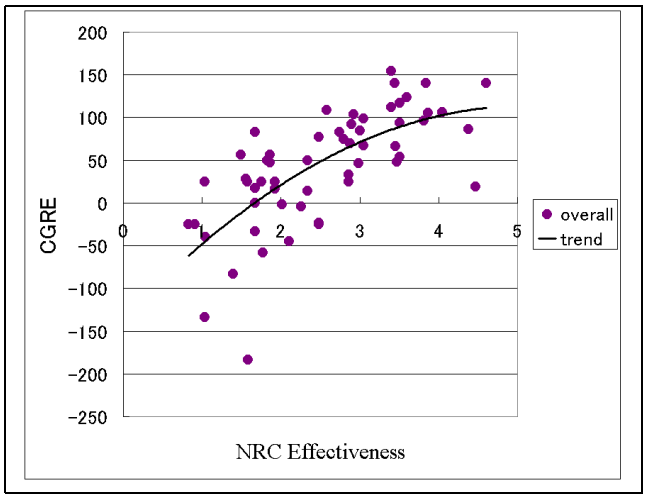

Figure 1: Overview of Published GRE Scores as a Function of School Desirability

scores". Logically, the minimum reported score should be some number of points (TM) below the threshold, and average score for acceptees should be some number of points (AT) above the threshold. The rest of this section explains how TM and AT were estimated.

Of course, in real life, TM and AT will vary from school to school. AT is largely a function of the distribution of the applicant population, and TM is largely a function of how schools chose to represent their decision-making. However, to keep things tractable, this model treats TM and AT as constants.

3.1.2 Estimating AT and TM from Individual Schools

The first way to estimate these parameters is to examine those four helpful schools which report both minimum and average GRE values for acceptees, namely BYU, Rice, UF, and UNC.

For schools which specify percentiles, rather than actual values, conversion is done using the 2002-2003 GRE Guide[8].

Then, plugging each set of GRE values into the calculator, the average difference between the min-CGRE and the avg-CGRE is 72: this gives one estimate for AT+TM.

3.1.3 Estimating AT and TM from Comparable Schools

The second method is to examine schools in the same broad quality bands. If school A and school B are similar in desirability, it's likely that their applicant pools are similar and their acceptance thresholds are also similar. Although the actual desirabilities are unknown, a reasonable proxy is the old NRC effectiveness rating[9].

[8]Reasoning that the scores being reported were mostly from that period. As it happens, the percentiles in 2003-2004 are not that different, as seen in the Guide `ftp://ftp.ets.org/pub/gre/994994.pdf`.

[9]The National Research Council Study of Ph.D. Programs in Computer Science, as found on the CRA web site (`http://www.cra.org/statistics/nrcstudy2/home.html`)

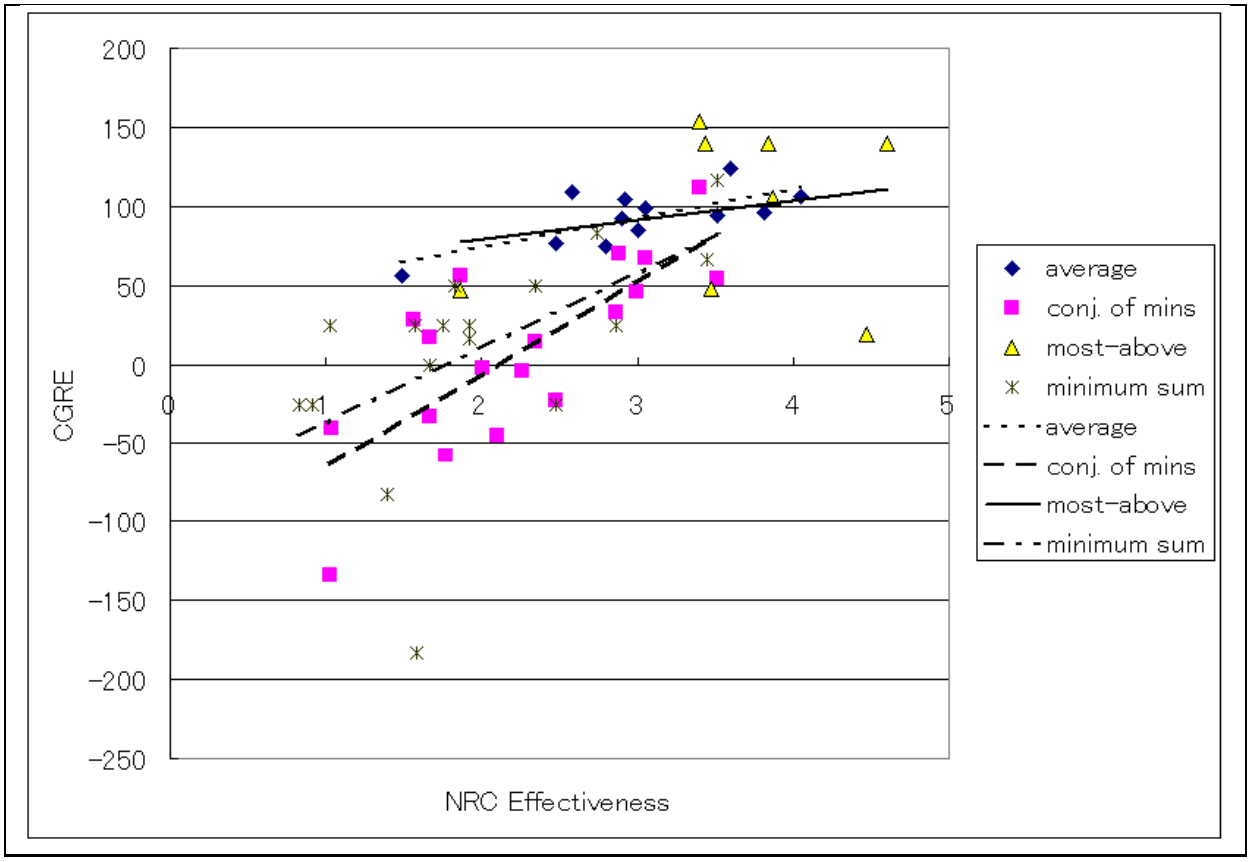

Figure 2: Published GRE Scores of Various Types as a Function of Desirability

Figure 1 shows CGRE versus effectiveness for all GRE value sets reported by all schools on the NRC list, regardless of whether they are referred to as a minimum, an average or something else. The curve is a second-order approximation to the data. This shows that, as expected, there is an overall tendency for better schools to be more demanding of their applicants.

Figure 2 shows the same data with the points split into categories based on how the GRE scores are reported.

First note that, as expected, the "avg" line is above the "min" line, meaning that, for two comparable schools, the one giving the average GRE will be reporting numbers higher than the one reporting a minimum GRE.

Second, note that the way GREs are reported seems to depend on the desirability of the school. Specifically, there is a strong tendency for top-ranked schools to report averages and for mid-ranked schools to report minimums. (Incidentally, there is also a tendency for even weaker schools to report only GPA requirements.)

The only region of data directly helping estimate the value for AT+TM is that of schools with effectiveness ratings from 2.5 to 3.5 (ignoring one outlier). In this region the difference between min-based CGREs and average-based CGREs is typically somewhere around 40: this gives another estimate for AT+TM.

Incidentally, the fact that the slopes of the regression lines are clearly different may suggest that the value of AT+TM is not constant but varies as a function of school desirability. However there is another interpretation, that the graph really involves two parallel curves, both with the shape of the curve in Figure 1, with the avg curve above the min curve. On this interpretation it is reasonable to treat AT+TM as constant.

3.1.4 Estimating AT and TM from a Sample Distribution

The third way to estimate AT and TM is to examine the distribution of acceptees at a specific school. At UT El Paso, AT is around 40; that is, the average acceptee CGRE is about 40 points above the threshold. However there is a complication, in that the average CGRE is lower than the CGRE of the average. The difference arises because few applicants will have GRE subscores as well balanced as those seen in the average. Based on 55 applicant datapoints, this difference is assumed to be 10 points, thus a better estimate for AT is 40 - 10 = 30.

TM can be estimated as the sum of a "balancing factor" and a "discretion factor".

The balancing factor arises because, if the GRE profile is expressed as the conjunction of hard minimums, most applicants will be caught out by the one GRE score which has the lowest normalized value. This means that an adjustment is necessary. For example, a student with unbalanced GREs may be a minimal applicant by School X's standards, but would have a CGRE higher than that given by simply plugging the X values into the model. Based on 55 datapoints, I estimate the average difference as 75 points. Thus, if School X's CGRE is directly computed as 14, the typical applicant would have to get a CGRE score of 89 to pass the conjoined 3 conditions. Of course, the balancing factor will be lower to the extent that applicants are actually looking at the GRE profiles specified and targeting their applications to schools where they can squeak in.

The discretion factor is a measure of whether the published minimum means "exceed this and we'll consider you" rather than "exceed this and you're in". Based on the discussion in Section 3.3.2, it seems reasonable to assume this is about 50 points on average, meaning that the CGRE threshold is estimated to be about 50 points above the effective minimum CGRE.

Adding these two factors gives an estimate for TM of 75 + 50 = 120.

For AT+TM the estimate is therefore 30 + 125 = 155.

3.1.5 Final Estimates for AT and TM

Thus the estimates for AT+TM range are 72, 40, and 155.

The highest value is clearly anomalous, probably reflecting an unusually diverse applicant pool and an unusually thoughtful admissions committee.

The value of 72 probably reflects the fact that schools which report both average and minimum can afford to report a lower minimum without giving false hope to too many people. Hence AT+TM for these schools is probably higher than usual.

Overall, therefore, a value near 40 is probably most generally valid. Giving some credence to the other estimates, I round up and use a value of 50 for AT+TM.

The next question is, of course, what fraction of this is AT and what fraction is TM? Based on the logic in Section 3.1.4 and also on the fact that the CGRE is closer to an average than to a minimum, TM should be larger than AT.

Thus AT is estimated as 20 and TM as 30.

To me it's surprising that AT comes out so small. Of course, the CGRE score represents a combination of information about the applicant, and so the variance is much lower than the variance on any single GRE score. But the small value still implies that there is not much variety in the talent level (GQ score) of the attendees at most schools. I guess this means that the market is fairly efficient.

3.2 Modeling Uncertainty

Before specifying the details of interpreting published scores, this section digresses to discuss sources of error. This is important since it would be a disservice to report values without indicating the degree of uncertainty, especially where the uncertainty regarding one school is higher than for another.

Probably what most potential applicants would like is to have schools categorized into three classes: where they probably will be accepted, where they probably have no chance, and where they may have a chance. This is the information that best supports the very common strategy of applying to one or two safe schools and maybe also a few take-a-chance schools.

The specific points we'd like to provide, therefore are the GQ point above which chances are very good (say $> 90\%$), and the GQ point below which chances are very poor ($< 10\%$).

However, given the limitations of the modeling, doing so would mean that the region of uncertainty would often be so wide as to be completely uninformative. Instead, therefore, I report the 50% points, namely the GQ score above which I estimate the applicant has at least a 50% chance of acceptance, and the GQ below which I estimate the applicant has at least a 50% chance or rejection. Thus, for example if these points are at a GQ of 25 and 65, a person with a GQ of 24 would be estimated to have at least a 50% chance of rejection, and a person with a 66 at least a 50% chance of acceptance. Of course, applicants further above the 65 are expected to have substantially higher chances of acceptance, and symmetrically. Regarding applicants with scores between 25 and 65, I would decline to bet either way.

Rather than getting too technical, these two points are presented to users only implicitly in the course of lising "schools where you would be likely to be accepted" and "schools where you may have a chance".

While it would be nice to estimate these two points separately, for convenience I directly estimate only the GQ threshold itself and a margin of error. Pretending that uncertainty is symmetric, for each school I'll just report two points: GQthrehold - margin, and GQthreshold + margin.

The major likely sources of error include:

- User Interface Errors: It could be that users have problems understanding what the data entry screen is asking for, trouble accurately entering the data, trouble navigating to the results screen, etc. On the output side, users might have trouble understanding the output format or trouble interpreting what the results mean for them. Although these are very real possibilities, following tradition I'll just call these user errors, and not worry about them for purposes of estimating the margin of error.

- Lack of Information: The estimator gives GQ scores even when as few as 3 numeric values are entered. It also gives scores even when there is little informativeness, as for GPAs for unknown countries. It would be possible to use the sum of the importance weights (Section 2.2.9) as a measure of the amount of information available to the admissions committee, and consider this when estimating the margin of error. However, except for risk-adverse departments which want to minimize the downside potential, I think that the quantity of information is probably not a direct factor in admissions decisions, so this is not considered when estimating the margin of error.

- Incorrect Fundamental Assumptions: The assumption that a single number can represent applicant strength is doubtless incorrect, but is probably not a major contribution to error.

- Incorrect Input-Side Parameters: Certainly the model uses inaccurate values for many parameters relating to applicant evaluation, including such central ones as the GRE-Verbal baseline and such specific ones as the JNTU scaling factor. For this reason the minimal margin is estimated as 10 points.

- Incorrect Assumption of Uniform Decision-Making: The assumption that all schools make admissions decisions the same way is also clearly incorrect. The next subsection considers, for each school, how similar its decision-making is likely to be to that of the model, and estimates the margin based in part on this.

- Inadequate Information Regarding Decision-Making: Schools vary greatly in how much information they provide regarding admissions decision-making. I simply refuse to consider schools which do not publish at least two GRE scores. For other schools I add up to 40 points to the margin of error, depending on the quantity, clarity, currentness, believability, and utility of the information on the web.

  Schools also vary in how they specify the decision making. Some ways are inherently harder to map to the model, and the next subsection considers this also when assigning margins for each method.

  Even schools which use essentially the same algorithm for ranking applicants may vary in the parameters used. The only one of these which is sometimes inferable is the GRE baseline scores. Up to 30 points are added to the margin if these are unbalanced relative to those used in this model.

3.3 Interpreting Specific Reporting Methods

Having got all that out of the way, this section explains, for each of the common ways in which admissions criteria are expressed, how the published data is used to estimate the GQ threshold.

3.3.1 Average GRE for Acceptees

Since AT=20, the threshold GQ is estimated as acceptees' average CGRE - 20. The margin of error is estimated as 20.

Some schools report averages for enrollees, rather than acceptees, but these are assumed to be the same.

Some schools report average percentiles rather than average scores. Since percentiles appear to be an approximately linear function of scores, at least in the typical ranges, these are directly converted using the Guide (*ibid*).

3.3.2 Minimum GRE for Acceptees

Since TM=30, the threshold GQ is estimated as the CGRE of the minimum scores + 30. The basic margin of error is estimated to be 30.

There are three sub-cases to consider.

First, some schools clearly have a very small discretion factor (Section 3.1.4); that is, any applicant who exceeds the minimum is judged "qualified" and almost certain to be accepted. Schools like these tend to couch the GREs as "requirements", "criteria" or "conditions for admissions" (rather than conditions for application). These schools are handled poorly by the GQ model, because they ignore most of the factors it incorporates. Only 10 points are added to get the threshold, and the margin is 40 points.

Second, on the other hand, some schools appear to have large discretion factors. Such schools use phrases such as "average scores are much higher", "we consider all factors", "meeting the minimums does not guarantee acceptance", "decision-making is a holistic process", "we look for (skill set) as evidenced by (all factors)", "attempt to predict the likelihood of success", and so on. In essence, the message behind the minimums for such schools is "we look at everything, but if your GREs are below X, then it's really unlikely your GPA and other factors are going to be strong enough to pull up your GQ, so save your money and don't apply here".

To put an upper bound on the discretion factor, it's worth noting that publishing a GRE minimum represents a trade-off. Set too high, it runs the risk of losing too many potential applicants who might have been above the GQ threshold. At UT El Paso, which very aggressively identifies "diamond in the rough" applicants, the greatest recorded difference between GQ and CGRE is 158 points (yes, she was accepted, and she's doing very well). However, if the published minimum is set low, it runs the risk of giving false hope to too many who really should have been discouraged from applying. I personally would think a CGRE minimum of 100 below the GQ threshold would be a good choice, but Figure 2 suggests that this is unrealistic. For schools which appear to be selective, 40 points are added to get the threshold. The margin is unchanged at 30.

Third, there are, unfortunately are a number of schools which publish minimums but do not indicate whether they are to be interpreted as requirements for acceptance or as requirements for consideration. For such schools I just add 30 and use 50 as the margin.

Orthogonally, some schools require a certain value for only two GRE scores. As this is less restrictive, 10 fewer points are added to get the threshold, and the margin is increased by 5 points.

3.3.3 Soft Minimum

Although stating a conjunction of minimum scores, as above, a number of schools clarify that the minimum has some leeway. Typical expressions are "normally required minimum", "these are not hard cutoffs", "as guidelines", and "nominal minimum". I quantify this leeway as 20 points relative to a hard minimum; thus the threshold GQ is estimated as 10 points above such a soft minimum CGRE.

Regarding the margin of error, the fact that this decision-making style corresponds well to that in the model argues for a small margin, but on the other hand it's not clear exactly how to interpret vague adjectives like these. A margin of 30 is used.

3.3.4 Most Above

Another common way to describe admissions criteria is to specify something like "most acceptees had scores over x, y, and z".

This resembles a specification of the average scores, given three assumptions. First, although the word "most" is ambiguous, it's not unreasonable to assume it means 51%. Second, although the word "above" is also ambiguous, it's not unreasonable to assume it means only 1 point higher than the threshold. Third, although it refers to a median, for lack of knowledge about the asymmetry, the simplest assumption is that the median equals the average.

There is, however, one clear difference: "most above" implies that each applicant is above the stated number on all the subscores. This is exactly the 10 point "complication" mentioned in the first paragraph of Section 3.1.4. Comfortingly, Figure 2 may also be showing that the most-above line is higher than the average line. Therefore, to compute the threshold 10+AT=30 points are subtracted from the most-above CGRE.

Due to the ambiguities the estimated margin of error is 40 points.

3.3.5 Minimum Sum of Scores

Another common way to describe acceptance criteria is to specify that the sum of the GRE scores should be above some minimum. For example, school B specifies that the sum of V, Q, and A be above 2050.

The obvious way to convert to a CGRE is to subtract the model's baseline values (which sum to 1750) and divide by 3. These are the CGRE values for min-sum plotted in Figure 2.

$$CGRE_{minsum} = (published - 1750)/3 \tag{8}$$

One would expect that scores reported with min-sum would be higher than scores reported with min, because sum is permissive in allowing an applicant to trade off a strength in one score for a weakness in another, point-for-point.

Looking at the graph, this seems as if this may be true. Although logically I would expect the difference to be greater, from the graph it looks as if the min-sum curve is about 20 above the min curve. Accordingly, to compute the GQ threshold from the min-sum CGRE, I add AT - 20 = 10. The margin is estimated as 30.

Orthogonally, as before (Section 3.3.2), 10 pints are subtracted if the sum of only two GRE scores is specified.

3.3.6 Average Sum of Scores

This is logically weaker than average for the same reason that min-sum is weaker than min. After converting to a CGRE using Equation 8, to get the GC threshold I subtract 30 points. Since there are only a few schools which use this, it is hard to estimate whether this appropriate, so the margin of error is 30.

3.3.7 The GQ Score Critical Point

One school reports acceptance using the GQ score itself, namely the University of Texas at El Paso. Here also there is no fixed threhold, but –25 is a useful point to report In fact, MS applicants with a GQ over –25 had over 90% chance of acceptance, and those below had about a 10% chance of acceptance, in 2003-2004 decisions.

One caveat does need to be mentioned here. At UT El Paso the GQ score is typically somewhat above the CGRE, that is, the average acceptee's GQ is higher than his CGRE. While grades, letter warmth, and statement strength are likely to give an average contribution of 0; that is, their normalized scores are likely to be symmetrically distributed around the CGREs (since the normalization parameters for these were chosen to make this probably true), the other two factors, namely the bonuses for group membership and fellowships, are only ever positive, so those create a systematic bias. In the interest of allowing applicants to accurately predict their chances, for UT El Paso I am reporting the GQ score rather than the score based only on CGRE, at the risk of weaker comparability to the practice at other schools.

The margin used is the minimum, namely 10 points.

3.3.8 Hybrids

Some schools use hybrid descriptions of their admissions decision-making. In such cases the threshold is interpolated. The margin of error is smaller to the extent that these are compatible, larger if there are apparent internal contradictions.

3.3.9 Summary

Table 2 summarizes how the threshold is estimated from published data.

given a specification in terms of a	convert using	then add	basic margin of error
average sum	Equation 8	-30	30
most above	the model	-30	40
averages	the model	-20	20
GQ critical point	the model	0	10
minimum sum	Equation 8	10	20
hard minimums for acceptance	the model	10	40
soft minimums	the model	10	30
hard minimums (unclear)	the model	30	50
hard minimums to apply	the model	40	30

Table 2: Estimating GC Thresholds from Data Presented in Various Ways

As a sanity check on these values, it's useful to take the perspective of a CS faculty committee deciding how to report an acceptance policy which involves a secret threshold X. If the analysis above is correct, then they could reasonably report an average sum of X + 30, or a hard minimum to apply of X - 40, or anything in between, as seen in the table. Fortunately this does seem plausible.

3.4 Verification

To check the plausiblity of these conversions, all schools which reported GRE averages were sorted by infered threshold and suspicious values examined. This was repeated for each of the reporting methods, and then finally for the entire list of 72 schools. After slips were corrected, the ordering corresponded fairly well with my naive beliefs about which schools would have higher thresholds.

3.5 Likely Utility

Given all these complications, one might reasonably ask whether using the GQ score to predict admissions decisions is better than the clear alternative, namely implementing a little function for each school to express the decision criteria that the school actually uses.

The answer is not yet known. One the one hand, it's clear that such an approach would be better for modeling some schools. On the other hand, there are also many schools where the current model probably is a more accurate account of what they really do than is the account given on their web site. (To give just one example, many schools describe their use of Analytical scores but not Analytical Writing scores, although the switch-over happened over a year ago.)

4 Interface

The main properties of the interface should be self-evident.

Possibly non-obvious design considerations included perceived trustability, speed of access across low-bandwidth lines, usability (at the expense of total accuracy), and accessibilty to users without color monitors or color printers.

5 Implementation

The only significant implementation decision was the choice to do everything in client-side Javascript. This was done to make the system more survivable, more portable, more modifiable, and easy for others to inspect.

There are only two big disadvantages to Javascript. First, two hunks of code, namely the applicant evaluation functions and school display functions, had to be manually copied from file to file in order to reuse them for testing etc. Second, compatibility is weak: making it work for both Explorer and Netscape was more interesting than I'd have liked, and I have no confidence it will work for those I haven't tested.

The code itself `http://www.cs.utep.edu/admissions/display.html` includes comments interleaved in the Javascript code.

6 Possible Improvements

Of the many possible improvements, I believe the most important would be to tweak the GPA baseline and scaling factor, and the second most important to tweak the baselines, scaling factors, and importance weights for the various GRE scores.