Paper: jc*-**-**-****

Towards a Model of Computer Science Graduate Admissions Decisions

Nigel Ward

Computer Science, University of Texas at El Paso
500 West University Avenue, El Paso, Texas 79968, USA
nigelward@acm.org
[Received 00/00/00; accepted 00/00/00]

Abstract

Abstract. Potential applicants to graduate school find it difficult to predict, even approximately, which schools will accept them. We have created a predictive model of admissions decision-making, packaged in the form of a web page that allows students to enter their information and see a list of schools where they are likely to be accepted. This paper explains the rationale for the model's design and parameter values. Interesting issues include the way that evidence is combined, the estimation of parameters, and the modeling of uncertainty.

Index Terms

student assessment, acceptance criteria, application evaluation, decision-making, combination of evidence, ordered weighted average, uncertainty, GRE scores, GPA, letters of recommendation

1. Introduction

The process by which potential graduate students and graduate schools come together is time-consuming for the applicants, for the people who write letters of recommendation, and for admissions committees [1, 2]. In an attempt to make the process more efficient, we have created a web tool that can help inform potential graduate students in Computer Science (CS) about how applications are evaluated. This was developed at the Department of Computer Science at UT El Paso as a service to the computer science community. Using this tool may enable applicants to better target their efforts, to the benefit of all.

At the heart of the tool is a model of the behavior of computer science graduate admission committees. Previous work has shown that a department's rejection decisions can be predicted by a linear model derived by multiple regression [1]. Our model goes beyond this to produce accept/reject predictions for a large number of departments, for rather heterogeneous applicants, and for applicants for whom only partial information is available. Since the data available was inadequate for automatic model-building, our model uses computations which closely follow the considerations and reasoning used by admissions committees, thereby enabling the parameter values to be inferred logically and justified intuitively.

This paper focuses on those aspects of the model which are interesting in terms of combination of evidence, parameter estimation, and reasoning under uncertainty. Less interesting details are documented elsewhere [3]. Although the paper only addresses one domain, the issues faced may also arise in other problems involving complex decision-making and the inference of hidden decision-making behavior from incomplete information.

This paper states the key assumptions in Section 2, presents the model for computing applicant strength in Section 3, explains how various admissions policies are modeled in Section 4, and discusses some directions for future work in Section 5.

2. Key Simplifying Assumptions

The first assumption behind the model is that the strength of any applicant can be represented by a single number, referred to below as the GQ score (GQ stands for "Generalized Quantitative", or maybe "Guesstimated Quality").

The second assumption is that there is a single, universally shared way to evaluate CS graduate school applicants. Although clearly a simplification, this is not an unreasonable one, at least judging from the admissions committee discussions witnessed by the author at three very different institutions: the University of California at Berkeley, the University of Tokyo, and the University of Texas at El Paso. Differences in the evaluation criteria that various departments present on their web pages are taken to be only the result of random omissions and other noise. This implies that the admissions policy of any department can also be described with a single number, the threshold GQ.

The third assumption is that a model need not be perfect to be useful.

3. Modeling Applicant Strength

In essence, the model normalizes all the factors involved to one scale and combines them using a weighting which is chosen differently in each case to capture the actual importance or informativeness of each factor. This section considers each factor in turn.

	raw RV	normalized value NV	rank R	ranking factor RF	contribution level CL	importance weight IW	product	CGRE
verbal	600	100	#1	.67	67	.7	47	
quantitative	650	0	#3	1.33	0	1.0	0	
analy. writing	4.5	62	#2	1.0	62	.7	43	
						sum: 2.4	sum: 90	38

Fig. 1. Example of GRE Normalization, Ranking Factor Application, and Weighting

3.1. GRE Scores

The Graduate Record Examination (GRE) [4] is a standardized test produced by the Educational Testing Service (ETS) and required by many United States graduate schools of their applicants. This test is designed to measure aptitude for graduate study. Results on its three parts, Verbal Reasoning, Quantitative Reasoning, and Analytical Writing, are reported independently. Verbal (V) and Quantitative (Q) scores range from 200 to 800, and Analytical Writing (AW) scores range from 0 to 6.

3.1.1. Normalizing

In order to measure balanced strength, as discussed below, it is necessary to make comparisons across the scores. This is done by first normalizing them to the same scale. For this purpose it is convenient to define a baseline set of values, representing a typical well-balanced applicant. The absolute values chosen for this baseline do not affect any results since, as will be seen, the baseline is used to define a metric, not a criterion.

The values for the baseline on each scale were chosen based on the observation that, for departments which publish minimum scores, the typical difference between Q and V is 150 or a little more, which is what one might expect given the intellectual profile of the typical Computer Science student. There is less information on the use of AW, but a 4.0 is occasionally listed in contexts which list Q around 650 and V around 500. The model uses a baseline with round numbers to make it easier to explain and verify: specifically, the GRE scores are normalized using a baseline value of V 500, Q 650, and AW 4.0.

Following common practice, as reflected by the fact that many departments specify that they just add up V and Q points, these two scores are given equal value: the scaling factors are 1. The scaling factor for AW can be estimated in several ways: to align the entire range would require treating 1 AW point as equivalent to 100 points of V or Q; to match up a perfect V and a perfect AW would require a factor of 150, and the GRE score distributions suggest that 125 is an appropriate value. The model uses 125.

Thus the normalized value NV_i is given by:

$$NV_i = (RV_i - BV_i) \times SF_i \quad (1)$$

where RV_i is the raw value, BV_i is the baseline value, and SF_i is the scaling factor. For example, consider an applicant with a GRE profile of V 600, Q 650, and AW 4.5. These values would normalize to NV_V 100, NV_Q 0 and NV_{AW} 62.

3.1.2. Combining

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 departments 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 departments, 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 acceptance, reflecting the idea that a good applicant is one who is strong across the board, that is, that "balanced" individuals do better. This suggests a minimum-based method, where the weakest score is what matters.

Sophisticated admissions committees probably use something in between these two methods. In order to do so the model gives more weight to relatively weak scores. This method is in fact an "Ordered Weighting Averaging (OWA) Operator" [5]; an application in evaluating students was proposed by Carlsson et. al [6].

Thus, as the second step, the normalized subscores are ranked, that is ordered by size, where 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 $R_Q = 3$. Based on this ranking factors are applied. The formula for ranking factors is

$$RF_r = \frac{2}{3}(1 + \frac{r-1}{n-1}) \quad \ldots \quad (2)$$

where r is the rank and n is the number of scores. This formula is chosen to meet three needs: the contribution of the lowest normalized score is 2 times that of the highest (meaning that for a strength on one dimension to outweigh a weakness on another dimension, the strength has to be twice as large as the weakness), the sum of all the ranking factors is n (thus the total importance of all the numeric scores taken together is not altered), and scores of intermediate ranks are counted to intermediate degrees. Formula 2 is the simplest (i.e. linear) formula which satisfies these needs. Thus 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 applicant's success) to .67 for the highest. This is the model's combining operator.

Ranking factors are only applied to the numeric data, namely the GRE scores and the Grade Point Average (GPA). Thus the "contribution levels" are given by:

$$CL_i = NV_i \times RF_{r,n} \quad \ldots \quad (3)$$

3.1.3. Weighting

Now 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. This is also different from the ranking weight, which is only responsible for putting more emphasis on weaker scores.

Choice of the importance weights can be based on various considerations. One fact to consider is that ETS reports that V and Q are approximately equally correlated with graduate school first year grades. However true student success is measured less by grades than by good thesis work and timely completion, and these probably correlate more with Q than with V and AW. Perhaps the best way to chose the importance weights is to base them on frequency of mention: for example, all departments which give GRE numbers mention a Q score, so this is likely most important. Based roughly on frequency of mention on web sites and on frequency of mention in admissions committee meetings, the importance weights of the model are .7 for V, 1.0 for Q, and .7 for AW.

3.1.4. Summing

Finally these products are summed and brought back to the same scale by dividing by the sum of the importance weights, giving the "Composite GRE" (CGRE).

$$CGRE = \frac{\sum_i CL_i \times IW_i}{\sum_i IW_i} \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \quad (4)$$

To summarize, 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 normalized value by a ranking factor (depending on the score profile of the 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. Figure 1 illustrates.

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

3.2. Undergraduate Grades

The GPA fits easily into the same computation as the GREs: it is normalized relative to a baseline; the normalized GPA is then compared with the normalized GRE scores to determine whether a given GPA is consistent with, higher than, or lower than the GRE scores; based on this the ranking factors are applied; and finally an importance weight is applied.

3.2.1. US Grades

For US grades the model 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 scaling factor of 200 Q points for one GPA point is used. Both these values are based on roughly fitting a line to a scatterplot of data from departments which report both GRE and GPA numbers in their admissions data.

The importance of the GPA relative to the GRE is a tricky issue. ETS advises that undergraduate grades are the best single predictor of graduate success. It is also the case that undergraduate curricula in CS are generally similar across schools, making undergraduate GPA even more likely to be a reliable predictor. Moreover, the undergraduate GPA reflects important abilities such as programming and system integration, which are not measured by the GRE at all. On the other hand, relatively few departments publish an expected GPA score, probably because grading standards are not consistent across schools. For this reason the model gives the GPA an importance weight lower than it might otherwise be, namely 2.5, which makes it roughly as important as all the GRE scores combined.

Extending the above example with the additional information that the applicant had a GPA of 3.9, this gives the normalized value NV_{GPA} = (3.9 – 3.3) * 200 = 120. 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 the "In Major GPA" or "Recent GPA" is known, the value used for the above computation is the average of this and the overall GPA.

3.2.2. Foreign Grades

GPAs not on the US system can be handled by using country-specific or school-specific baselines and scaling factors. Estimating these is, however, not easy. One might try to estimate

the needed parameters from a department's applicant database by doing a regression of undergraduate GPA on the Composite GRE values. However correlations can be elusive, 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 because they also go elsewhere.

Thus a wider data set is needed. The model's parameters are currently estimated using two datapoints for each school: the cluster of typical UT El Paso applicants from that school and the Arizona State University cutoff for applicants from that school. For example, for graduates of Jawaharlal Nehru Technological University (JNTU), UT El Paso sees a cluster of applicants with a GPA around .68 and a CGRE around -20: this is the first datapoint. For JNTU applicants Arizona State University uses a GPA cutoff of .81 with a corresponding CGRE threshold of around 35 (computed from their GRE minimum profile of 400V, 700Q, and 650AW, as explained below in Section 4.5.2): this gives the second datapoint. These two datapoints give a slope of around 450 GRE points per GPA point, and a baseline intercept of .72.

For schools or countries where the GPA parameters have not yet been estimated, the vast majority, users are required to convert their average to the 4.0 US GPA scale before input.

The importance weights for foreign grades are lower than those for US grades, for two reasons. The first reason is that undergraduate success in certain countries is probably less predictive of US graduate success. The second reason is an "uncertainty penalty" to ensure that factors affected by parameters which are estimated only roughly should have less influence. While this uncertainty could be dealt with directly, the use of lower importance weights for foreign GPAs serves as a simple way to ensure that for foreign applicants the GREs contribute relatively more to the total score, as common sense would suggest.

3.2.3. Non-CS Majors

Although non-CS majors often bring special strengths, such as clearer motivation, these should be fully reflected in the letters and statement of purpose. Non-CS majors are, however, generally less strong as candidates: the model handles this by applying different baselines and importance weights for such GPAs.

For engineering, science, and math majors the baseline is raised 60 GRE points (.3 US GPA points), and for other majors 120 points. This adjustment is done for two reasons. The first is the likelihood of weak preparation. Of course a non-CS person coming into a CS Masters will take a few extra courses to fill in the missing background, but gaps may still remain, making it possible that that a non-CS person will do less well, especially on project and thesis work. The second reason is weaker grading scales. For example an A- student from an Information Technology (IT) program probably has demonstrated less performance than an A- student in CS. The baseline adjustment means, for example, that an IT major would need a 4.0 GPA to be as desirable as a CS major with a 3.4.

In addition, the importance weight is reduced by 10% for engineering, science, and math majors, and 20% for other majors, based on the assumption that non-CS grades are less predictive of CS graduate performance.

3.3. Letters of Recommendation

In general, the value of letters of recommendation lies mostly in the information they provide on abilities not well measured by the GPA or the GRE.

3.3.1. Normalization

Letter warmth is entered by a pull-down menu offering a number of positions along a scale, from "person not suited for graduate school" through "future good graduate student" (typical)

to "future CS hero". The associated values are normalized to the same scale as the GREs and GPA and then added in.

3.3.2. Ranking

Letters are not given a ranking weight. This is because, unlike GREs and the GPA, it seems unlikely that admissions committees have clear ideas about the minimal level of letter required. Although many aspects of the undergraduate experience are fairly uniform, 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 the contribution level is just the normalized value.

3.3.3. Importance

Letters are given a maximum importance weight of 3, which is about as much as the GPA, and about as much as the GREs. However this weight is only achieved when the letter writer is someone whose words the committee will believe and who knows the applicant well. This is an important feature of the model: by providing a likely upper bound on how much a letter can help out, it may help applicants plan more realistically.

Specifically, the importance weight for a letter is the product of the believability and the basis for judgment.

The first of these, the believability score, is entered under using a pull-down menu labeled "recommender is a ..." and offering choices such as "nobody special" and "researcher known to the admissions committee", as seen in Figure 2. This simple presentation is done to make the user's job easier, hiding a great deal of potential complexity. For example, there are probably at least two sub-factors involved in believability. The first sub-factor is the perceived trustworthiness of the person who is writing the letter. (Incidentally, the trustworthiness issue may be why some departments' recommendation forms provide checkboxes to indicate whether the applicant is in the top 20%, 10%, 5%, etc. It is unlikely that such numbers are suitable for directly feeding into a formula, unless one has 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 within believability is whether the recommender knows what it takes to succeed in graduate school. However to keep things simple, these sub-factors are conflated into a single believability index.

Regarding 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 informative if the writer has observed the applicant 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 and abilities.) Thus the basis for judgment score is entered using a pull-down menu labeled "observing you as a ...", with options such as "student in the classroom" and "apprentice researcher". 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. Instead letters with a weak basis simply have little effect.

3.4. Other Factors

The model also includes factors for the statement of purpose, for fellowships, and for membership in targeted groups, as detailed in [3]. Factors omitted from the model include undergraduate school reputation and standards (except as reflected in the GPA-related parameters), GRE Subject Test in Computer Science scores, TOEFL score, nationality, culture, native language, specific coursework, area of research interest, and publications.

3.5. Summary of the Model

Since all of the inputs are normalized to the same scale it does not matter if some values are unknown: the same computation applies and the result has the same meaning. This is useful, for example, when the AW score is not yet known or when the user is not willing to estimate the likely warmth of his letters of recommendation. However, to help minimize the number of misleading estimates, the implementation refuses to give an estimate without at least three numeric values, i.e. three out of the GPA and the GREs.

In sum:

$$GQ = \frac{\sum_i CL_i \times IW_i}{\sum_i IW_i} + \text{Other Factors} \quad (5)$$

where i ranges over {V, Q, AW, GPA, Recommendation, Statement}. 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 does permit comparison to the acceptance criteria at various departments. Thus, as mentioned before, the absolute value of the baseline does not matter, since both departmental criteria and applicant strength are normalized using the same baseline. (However the absolute value may matter psychologically; indeed, since the current baseline is set fairly high, the GQ score comes out negative for some students, which may be impolitic.)

3.6. Modeling Uncertainty

As the discussion above suggests, the GQ metric handles uncertainty only indirectly, in the form of low importance weights assigned for foreign grades, non-CS grades, and uninformative letters of recommendation. It would be interesting instead to explicitly model uncertainty. For example, since the model gives GQ scores even when incomplete information is entered, one could use the sum of the importance weights for the factors given as a measure of the amount of information available to the admissions committee. More sophisticated ways to represent and combine the uncertainty associated with each of the factors are also easy to imagine.

Alternatively one could create a model where each factor is treated in terms of the information it provides about the probability of various GQ scores (although it might be tricky to determine the conditional probabilities necessary to approximate the model's combining operator, that is Equations 2 and 3).

Such techniques could quantify the uncertainty associated with the GQ score for each applicant. However this would be of value only to the extent that departments are risk-averse. If, on the other hand, departments are willing to accept the chance of downside risk as the price for the chance of an upside reward, then they will make decisions based only on the most likely interpretation of an applicant's strength. Assuming that this is generally the case, it makes sense to report GQ scores without reporting the associated uncertainties.

3.7. Verification

The parameters of the model were cross-checked by inferring them in multiple ways, as suggested above.

The model as a whole was checked against a test set of 55 UT El Paso applicants from 2003-2004. This uncovered a few typographical errors in the parameters. After these were fixed, proper testing began. The first test was whether the system actually predicted the admissions decisions. For this the threshold GQ for acceptance was set to -25. In all but four cases this

	NV_V	NV_Q	NV_{AW}	CGRE
Joe	20	20	100	39
Bill	100	20	20	39
Kate	60	80	60	65
average CGRE				48
CGRE threshold ($CGRE_t$)				35
averages	60	40	60	
CGRE of the averages ($CGRE_a$)				52
minimums	20	20	20	
CGRE of the minimums ($CGRE_m$)				20

Table 1. Acceptee Data for Hypothetical Department X

gave correct predictions. One class of failures was clearly due to the (previously unknown) fact that the UTEP admissions committee had been giving very low importance to Mexican grades. However, considering the outcomes for the three students who were accepted as a result, this must be considered a bug in the committee's decision-making, rather than something worth modeling. (The ability of a model to improve in this way on the collective seasoned wisdom of a committee came as a surprise to us, although perhaps it should not have [7, 1, 8].) The other prediction failure was due to a factor omitted from the model about which the committee had special information for one student.

The second test was to compare the GQ-based ranking of applicants to the ranking produced by an earlier model, a simple spreadsheet formula developed by Luc Longpre [9]. (This formula was developed to help the admissions committee maintain consistent decision-making across semesters.) Deviations in rankings were identified and the source of each was examined. No undesirable deviations were found; rather each was due to a feature deliberately designed into the new model.

4. Estimating Admissions Thresholds

Having seen how to compute a single number representing the applicant's strength, namely the GQ score, this section explains how to use that to predict the likelihood of acceptance at various departments.

As noted earlier, the working assumption is that all departments evaluate applicants in the same way. Accordingly it is reasonable to describe the admissions policy of each department with a single number, the GQ threshold, that is, the number such that an applicant is accepted iff his GQ is higher than this threshold. This section explains how these thresholds are estimated.

4.1. Preliminaries

Departments typically report values on a number of dimensions, such as GRE-V, GRE-Q, and GPA. Fortunately these can be reduced to a single value using the model already developed, using the assumption that the selectivity of a department can be measured in the same way as the quality of an applicant.

Since the parameters regarding GPAs and letters are less reliable than those for GREs, the threshold estimates are based only on GRE information: thus the values discussed in this section are all CGRE values, computed using Equation 4.

4.2. Averages, Minimums, and Thresholds

Since no school, except for UT El Paso, reports an actual GQ or CGRE threshold number, it is necessary to estimate the GQ thresholds from whatever numbers are published. This is complicated because departments report admissions criteria using a variety of methods.

This subsection discusses two touchstone reporting methods: "minimum GRE score set (conjunction of minimums)" and "average GRE scores". To make the discussion easier to follow, Table 1 shows the set of acceptees and some summary statistics for hypothetical department X. Here CGRE_t is the actual (secret) threshold. Department X reports the average scores of acceptees for each of the three GRE scores, and CGRE_a is the value obtained from these three values using Equation 4. Department X also reports a conjunction of minimums, such that only applicants with at least the specified values on each of the three GRE scores are accepted; CGRE_m is the value obtained from these using Equation 4.

The remainder of this subsection uses three types of evidence to estimate the typical relations between CGRE_t, CGRE_a and CGRE_m.

4.2.1. CGRE_a – CGRE_m from Individual Departments

The first approach is to examine the departments which report both minimum and average GRE values for acceptees, namely BYU, Rice, UF, and UNC. (For departments which specify percentiles, rather than actual values, conversion is done using the GRE Guide [4].) For these four the average difference between CGRE_a and CGRE_m is 72.

4.2.2. CGRE_a – CGRE_m from Comparable Departments

The second approach is to examine departments in the same broad quality bands. If department A and department B are similar in desirability, it is 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 NRC effectiveness rating [10], which is dated but still useful.

Figure 3 shows CGRE versus effectiveness for all GRE value sets reported by all departments 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 departments to be more demanding of their applicants.

Figure 4 shows the same data with the points split into categories based on how the GRE scores are reported. As expected, the "average" line (based on the CGRE_a values) is above the "conj. of mins" line (based on the CGRE_m values), since, for two comparable departments, the one giving the average GRE will be reporting numbers higher than the one reporting a minimum GRE. Incidentally, the way GREs are reported seems to depend on the desirability of the department: there is a tendency for top-ranked departments to report GRE averages, for mid-ranked departments to report GRE minimums, and for weaker departments to report only GPA requirements.

The only region of data allowing direct comparisons between CGRE_a and CGRE_m is that of effectiveness ratings from 2.5 to 3.5. In this region CGRE_a – CGRE_m is typically somewhere around 40. It is not unreasonable to treat this as constant, on the hypothesis that the underlying tendency is such that Figure 4, properly plotted, would show parallel curves, all with the shape of the curve in Figure 3, with the avg curve above the min curve.

4.2.3. CGRE_a – CGRE_t from a Sample Distribution

The third approach is to examine the distribution of acceptees at a specific department, here 55 recent applicants to UT El Paso. For this population the average acceptee CGRE is about

40 points above the threshold. However there is a complication, namely that the average CGRE is lower than the CGRE of the average, this is seen in Table 1. This difference arises because the model's combining operator is stricter than averaging and few applicants will have GRE subscores as well balanced as those seen in the average. Based on the same 55 datapoints, this difference is estimated as 10 points, giving an overall estimate of 40+10 = 50 for CGRE_a – CGRE_t.

4.2.4. CGRE_t – CGRE_m from a Sample Distribution

The distribution data can further be used to relate CGRE_t to CGRE_m. One part of the difference between the two is due to the leniency of the model's combining operator relative to min. The difference arises because, if the acceptance criterion 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. For example, in Table 1 Joe is a minimal applicant by Department X's standards, because of the low NV_V, but his CGRE is higher than CGRE_m and CGRE_t. Estimating the relation between average CGRE and CGRE_m from the 55 UT El Paso datapoints gives a difference of 75 points. Thus, for example, if Department X's CGRE_m is 20, an applicant with typically unbalanced scores would need a CGRE of 95 to pass the conjoined 3 conditions. Of course this difference will be less to the extent that applicants are actually looking at the GRE profiles which departments specify and targeting their applications to places where they can squeak in.

4.2.5. Final Estimates

Thus the three estimates for CGRE_a – CGRE_m are 72, 40, and 50+75=125. The highest value, 125, probably reflects an unusually diverse applicant pool and an unusually thoughtful admissions committee. The middle value, 72, probably reflects the fact that departments which report both average and minimum can afford to report a relatively lower minimum without giving false hope to too many people. Overall, therefore, a value near 40 is probably most generally valid. Giving some credence to the other estimates, the model uses a value of 50 for CGRE_a – CGRE_m.

Further, based on the analysis from the sample distribution above, it seems that CGRE_t is probably closer to CGRE_a than to CGRE_m. In Figure 2, this means that the inferred threshold values form a curve lying between the avg line and the conj. of mins line, somewhat above the middle. Specifically the CGRE_t for each department is taken to be 20 GQ points below CGRE_a for that department. It is somewhat surprising that this appears to be so small, since this implies that the average acceptee is not much better that the worst acceptee, that is, there is not much variety in the talent level, in terms of GQ score, of the graduate students at most departments. Perhaps this means that "the market" for graduate student admissions is fairly efficient.

4.3. Sources of Error and Uncertainty

Before continuing with the details of how published scores are interpreted, this subsection digresses to consider the sources of error and uncertainty. This is important since it would be a disservice to give accept/reject predictions without indicating the degree of uncertainty, especially when the uncertainty regarding one department is greater than for another. The major likely sources include:

1. User Interface Problems: 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.

2. Lack of Information about the Applicant: As noted earlier, the uncertainty associated with applicants can vary, but this is probably not a major factor in prediction uncertainty.
3. 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.
4. Incorrect Input-Side Parameters: Certainly the model uses inaccurate values for many parameters relating to applicant evaluation. However it is almost linear, so small errors in parameter values will not have disproportionate effects.
5. Incorrect Assumption of Uniform Decision-Making: The assumption that all departments make admissions decisions the same way is clearly incorrect. As explained below, for each department the associated uncertainty is greater to the extent that its decision-making seems to diverge from that of the model.
6. Inadequate Information Regarding Decision-Making: Departments vary greatly in how much information they provide regarding admissions decision-making. Departments must publish at least two GRE scores to even be included in the listing provided by the tool. In general, the quantity, clarity, currentness, believability, and utility of the information on the web are factors reducing the uncertainty regarding a department's threshold.

The tool only models the sources of uncertainty that differ among departments, that is sources 5 and 6. Although this uncertainty could perhaps be computed in a principled way, it is currently handled in a very simple manner, as seen below.

4.4. Communicating Uncertainty Judgments to Users

Although in general it can be difficulty to present information about uncertainty in a way easy to understand, here this is relatively easy, since potential applicants' needs are relatively simple; they probably just want departments categorized into three classes: 1. where they probably will be accepted, 2. where they probably have no chance, and 3. where the outcome is effectively unpredictable. This is the information that best supports the common strategy of applying to one or two safe departments and maybe also a few take-a-chance departments.

To make such a categorization it is necessary to infer for each department the GQ point above which chances are good and the GQ point below which chances are poor. Choosing these points involves deciding how much certainty to require before offering a prediction; this is tricky because there is a trade-off between producing many predictions and producing only highly reliable predictions. Considering likely user preferences, the two numbers currently used are the GQ score at which I estimate the applicant has a 80% chance of acceptance, and the GQ at which I estimate the applicant has a 80% chance of rejection. These estimates are subjective, but they are at least made consistently, as detailed below. Thus, for example, if these number are at 25 and 65, a person with a GQ of 24 would be estimated to have at least a 80% chance of rejection, and a person with a 66 at least a 80% chance of acceptance. Of course, applicants further above the 65 are expected to have higher chances of acceptance, and similarly for lower scores.

To make these numbers easy to understand, they are presented to users graphically, as seen in Figure 5.

While it might be better to estimate these two points separately, for convenience the model uses just a simple margin of error which is assumed to be symmetric. Thus it just reports two points: GQthreshold - margin and GQthreshold + margin.

4.5. Interpreting Specific Reporting Methods

This subsection explains how published data is used to estimate the GQ threshold and margin, for each of the common ways in which admissions criteria are expressed.

4.5.1. Average GRE for Acceptees

In this case, as explained in Section 4.2.5, the threshold is estimated as the average acceptee's CGRE minus 20. The margin of error is estimated as 20.

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

Some departments report average percentiles rather than average scores. In the typical ranges, percentiles appear to be an approximately linear function of scores, so these are directly converted using the Guide [4].

4.5.2. Minimum GRE for Acceptees

In this case, as explained above, the threshold is estimated as 30 points above the CGRE computed from the specified minimum values. This generic value is used if there is no additional information, however most departments give more information about how the minimums are used.

On the one hand, some departments clearly accept almost all applicants who meet the minimums. Departments like these tend to couch the reported GRE scores as "requirements", "criteria" or "conditions for admission". These departments are handled poorly by the GQ model because they ignore most of the factors it incorporates. In this case only 10 points are added to get the threshold, and the margin is 40 points.

On the other hand, some departments appear to be more selective. Such departments use phrases such as "average scores are much higher", "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 departments is "we look at everything, but if your GREs are below these minimums, then it's so unlikely that your GPA and other factors are strong enough to pull up your GQ that you probably shouldn't even apply here".

The studied ambiguity in these phrases reflects the fact that a department which chooses to publish a GRE minimum faces a trade-off. If too high, it risks scaring away too many potential applicants who might have been above the GQ threshold. At UT El Paso, which aggressively identifies "diamond in the rough" applicants, the greatest recorded difference between GQ and CGRE is 158 points. (The applicant was accepted and did well.) However, if the published minimum is too low, it risks giving false hope to those who really should not be encouraged to apply. Thinking idealistically, a CGRE minimum of 100 below the GQ threshold might be a good choice, but the analysis of Section 4.2 suggests that most departments are not so willing to consider applicants with weak scores. For departments which appear to be selective, 40 points are added to obtain the threshold. The margin is estimated as 30.

One minor point to note is that some departments specify a minimum 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. Another minor point is that some departments specify a set of minimum scores which are unbalanced relative to the baseline used in the model. Up to 30 points are added to the margin depending on the degree of imbalance.

4.5.3. Soft Minimum

Of the departments which state a conjunction of minimum scores, some specify that there is some leeway. Typical expressions are "normally required minimum", "these are not hard cutoffs", "as guidelines", and "nominal minimum". This leeway is quantified as 20 points relative to a generic hard minimum; thus the threshold GQ is estimated as 10 points above such a soft minimum CGRE.

Regarding the margin of error, this style of decision-making corresponds well to that in the model, which argues for a small margin, but on the other hand the vague adjectives make the interpretation uncertain. A margin of 30 is used.

4.5.4. 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, one can assume it means 51%. Second, although the word "above" is also ambiguous, one can assume it means just 1 point higher than the threshold. Third, although this references the score median, in the absence of information about the distribution it is simplest to assume that the median equals the average.

There is, however, one clear difference: this wording implies that each applicant is above the stated number on all the subscores. This is exactly the 10 point "complication" mentioned in Section 4.2.3. Therefore, to compute the threshold 10+20 =30 points are subtracted from the CGRE computed from x, y, and z.

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

4.5.5. 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. This is more common for departments using the old format GRE, which had three equally scaled scores: V, Q, and A. For example, department B specifies that the sum of V, Q, and A be above 2050.

The obvious way to convert this to a CGRE is to subtract the model's baseline values (which, including A, sum to 1750) and divide by 3. These are the CGRE values for min-sum which were plotted in Figure 4.

$$CGRE_{minsum} = (published - 1750)/3 \quad . . . . . (6)$$

One would expect that scores reported with min-sum would be higher than scores reported with a conjunction of mins, because sum is permissive in allowing an applicant to trade off a strength in one score for a weakness in another, point-for-point. Figure 4 suggests that this is true. Although logically one might 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, 10 is added (30 – 20 = 10). The margin is estimated as 30.

Orthogonally, as before, 10 points are subtracted if the sum of only two GRE scores is specified.

4.5.6. Summary

Table 2 summarizes how the thresholds are estimated from published data.

As a sanity check on these values, imagine a CS graduate admissions committee deciding how to report an acceptance policy which involves a secret threshold T. If the analysis above is correct, they could reasonably report an average sum of T + 30, or a hard minimum to apply of T – 40, or various values in between, as seen in the table. Fortunately this does seem reasonable.

4.6. Verification

To roughly check the plausibility of these conversions, all departments which reported GRE averages were sorted by inferred threshold and suspicious values examined. This was repeated

given a description in terms of	convert using	then add	margin of error
average sum	Equation 6	-30	30
most above	Equation 4	-30	40
averages	Equation 4	-20	20
GQ threshold	Equation 4	0	10
minimum sum	Equation 6	10	30
hard minimums for acceptance	Equation 4	10	40
soft minimums	Equation 4	10	30
hard minimums (unclear)	Equation 4	30	50
hard minimums to apply	Equation 4	40	30

Table 2. Estimating GQ Thresholds from GRE Data Presented in Various Ways

for each of the reporting methods, and then finally for the entire list of 72 departments. After typographical slips were corrected, the ordering corresponded fairly well with various information about relative selectivity.

5. Future Work

Future work includes examining the accuracy of the predictions for various departments and improving them, comparing this model to other possible models, modeling uncertainty better, and evaluating the actual utility of the tool to students. This section discusses each topic briefly.

A priority for future work is evaluating the ability of the model to predict the admissions decisions of various departments. Anticipating that it will not perform perfectly, improving the model will be another priority. It seems likely that the weakest part of the model will be its failure to account for different standards of grading across undergraduate school. This could be handled, as above, by assigning different importance weights or different baselines to GPAs from different schools, depending on the quality of their undergraduate CS programs, although obtaining data on this could be difficult. More generally, all the parameters could benefit from better tuning using more data.

Another topic for future work is comparing this model to other models. One alternative model could be based on a family of specific little functions, one for each department literally expressing the published criteria for that department. This would likely be better for some departments, but on the other hand there are also many departments where the model presented here probably is a more accurate account of actual committee behavior than the simplified account publicly given. (To give just one example, many departments still describe their use of Analytical scores but not Analytical Writing scores, although the format change happened over a year ago.) Another alternative model would include more parameters per department. Currently each department is modeled with just two parameters, one representing the inferred acceptance threshold and one representing the estimated margin of error. Probably 22 parameters would allow highly accurate modeling of most departments, with all but 3 or 4 being the same almost universally. On the other hand, it would also be interesting to compare the current GQ model with simpler ones, to see whether adequate predictions could be obtained with fewer parameters.

All of these activities will require the acquisition of more data on applicant pool statistics and admissions decisions at various schools. To obtain such data may be a challenge, due to privacy

concerns and feelings that admissions decision-making methods are "trade secrets". However it seems likely that, for most Computer Science departments, handling graduate admissions is not seen as a core business activity nor a source of competitive advantage, and so a standard resource that alleviates problems may be welcomed, as long as it introduces no unfairness. It remains to be determined what degree of standardization or sharing would be welcomed, and at what level: reporting, metrics, models, tools, websites, etc.

Another topic for future work would be a more principled modeling of uncertainty.

Finally, the use of the model needs to be explored, to ensure that the system meets the true needs of students. While there is anecdotal evidence that it does (in the form of emails of thanks), it would be helpful to know more about why students use or do not use it, how they use it, how they use it in conjunction with other sources of information [11], and how the information they find affects their decisions about whether and where to apply.

6. Conclusion

This paper has presented a model of Computer Science graduate admissions decisions. While the model is too complex for casual users to work through, it lends itself to implementation in the form of a calculator on the web, the "Acceptance Estimator for Computer Science Graduate Admissions", currently located at http://www.cs.utep.edu/admissions/. Although clearly not a final solution, this tool does illustrate how admissions decisions can be modeled and where the issues lie. As such, it forms a case study in the modeling of complex decision-making and the inference of hidden decision-making behavior from incomplete information.

Acknowledgements

The author would like to thank Luc Longpre for inspiration, Vladik Kreinovich for generous advice and encouragement, Chitta Baral for the ASU data, and Salamah Salamah and an anonymous reviewer for comments.

References:

[1] A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making. Robyn M. Dawes. American Psychologist, vol. 26, pp 180-188. 1971.
[2] The (Un)Predictability of Computer Science Graduate School Admissions. Nigel Ward. Communications of the CACM, 49, to appear, 2006.
[3] A Model of Computer Science Graduate Admissions Decisions. Nigel Ward. University of Texas at El Paso Computer Science Technical Report UTEP-CS-04-30, (also at http://www.cs.utep.edu/admissions/model.pdf), 2004.
[4] ETS: Guide to the Use of Scores. ftp://ftp.ets.org/pub/gre/994994.pdf, 2004.
[5] On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decisionmaking. Ronald R. Yager. IEEE Transactions on Systems, Man, and Cybernetics, vol. 18, pp 183-190. 1988.
[6] 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.
[7] Epistemology and the Psychology of Human Judgment. Micheal A. Bishop and J. D. Trout. Oxford University Press, 2005.
[8] Psychological Science can Improve Diagnostic Decisions. John A Swets, Robyn M. Dawes, and John Monahan. Psychological Science in the Public Interest, vol. 1, pp 1–26, 2000.
[9] MSApplicants.xls. Luc Longpre. spreadsheet used at the University of Texas at El Paso, 2003.
[10] The National Research Council Study of Ph.D. Programs in Computer Science, 1993/1995, as found on the CRA web site at http://www.cra.org/statistics/nrcstudy2/home.html.
[11] The Information Needs of Prospective Graduate Students. Rodney T. Hartnett. 1979. GRE Board Technical Report 77-8R.

Name:
Nigel Ward

Affiliation:
Department of Computer Science
University of Texas at El Paso

Address:
500 West University Ave, El Paso, Texas 79968, USA

Brief Biographical History:
Nigel Ward received a Ph.D. in Computer Science from the University of California at Berkeley in 1991. From 1991 to 2002 he was a member of the Engineering faculty at the University of Tokyo, where he co-directed the Human-Computer Interaction Laboratory. His primary research area is real-time responsiveness into spoken dialog systems, with a focus on the timing, phonetics, and meaning of non-lexical utterances such as uh-huh and um.

Membership in Learned Societies:
• ACM, ISCA, ACL, AAAI, IEEE, LSA, IPA, AMTA, IPSJ, JSAI

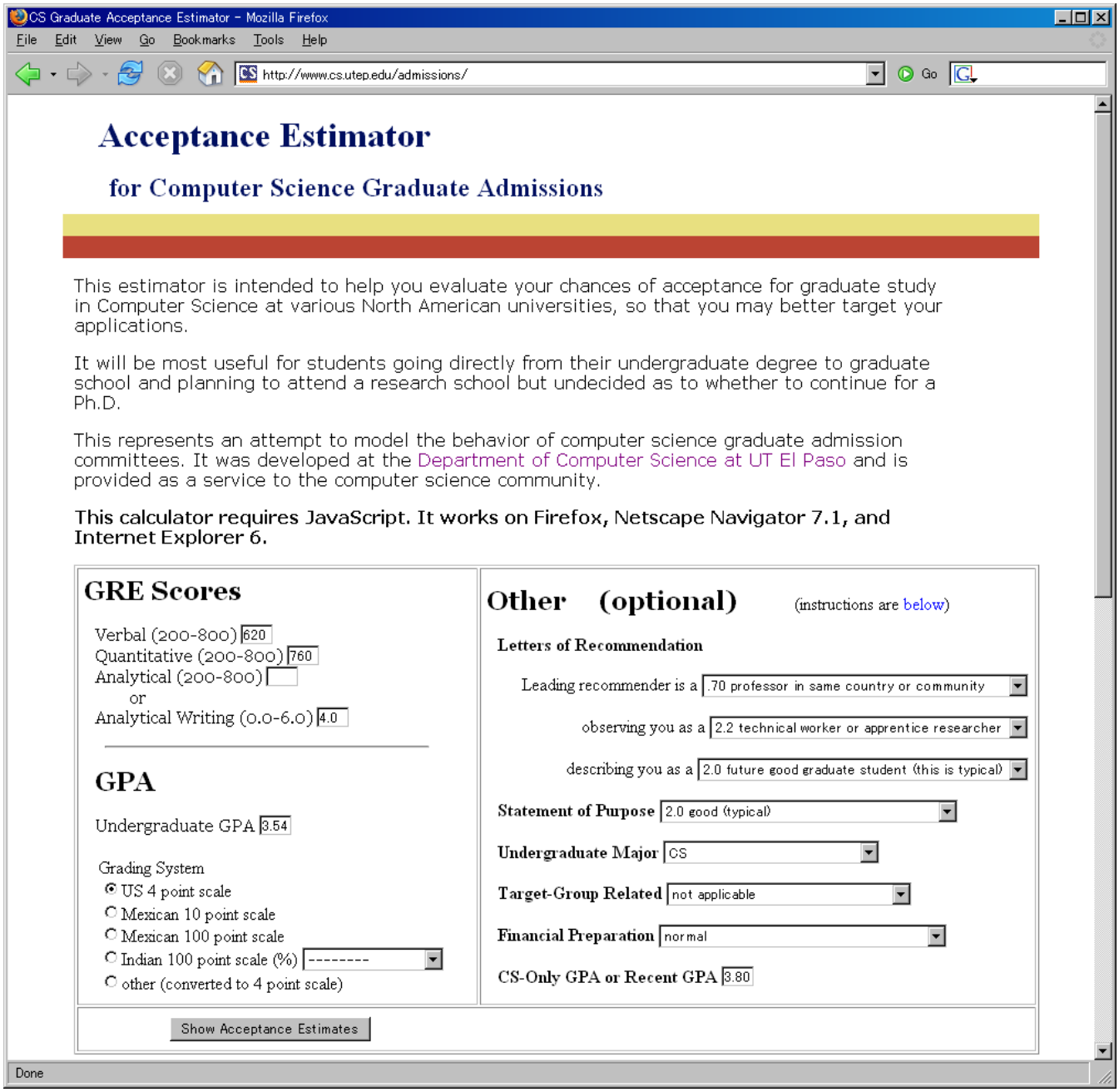

Fig. 2. Data Entry Screen

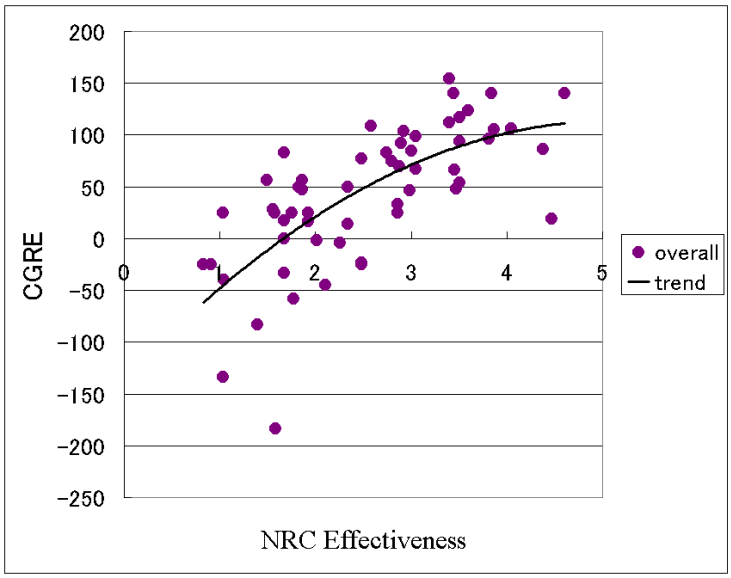

Fig. 3. Overview of Published GRE Scores as a Function of Department Desirability

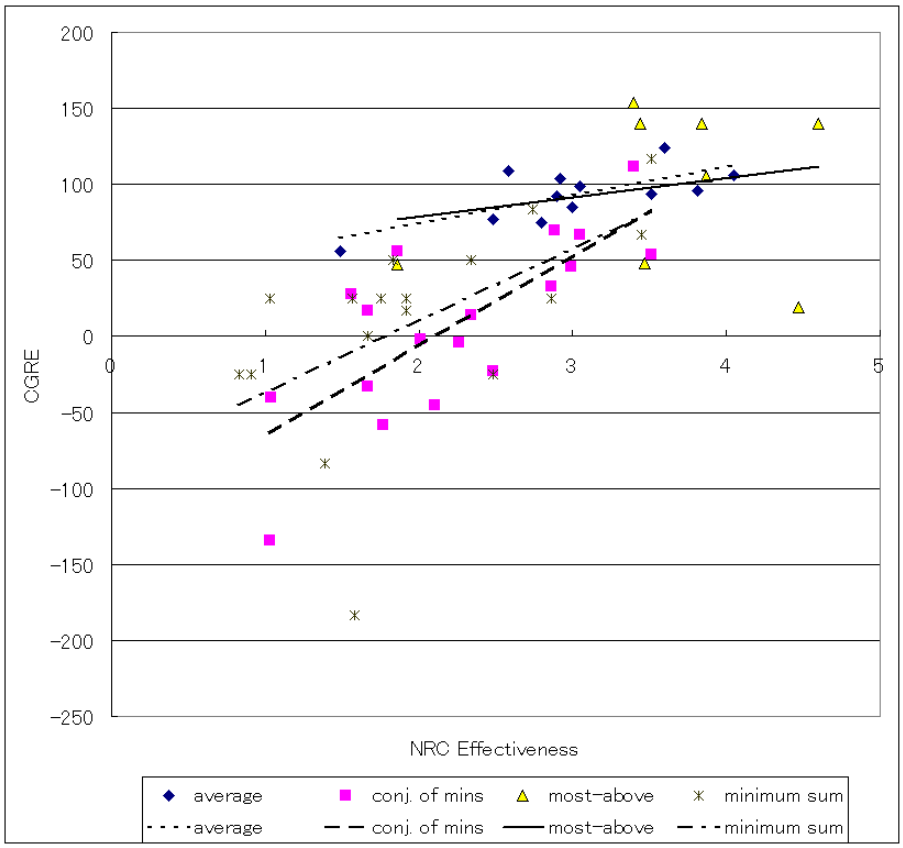

Fig. 4. Published GRE Scores of Various Types as a Function of Desirability

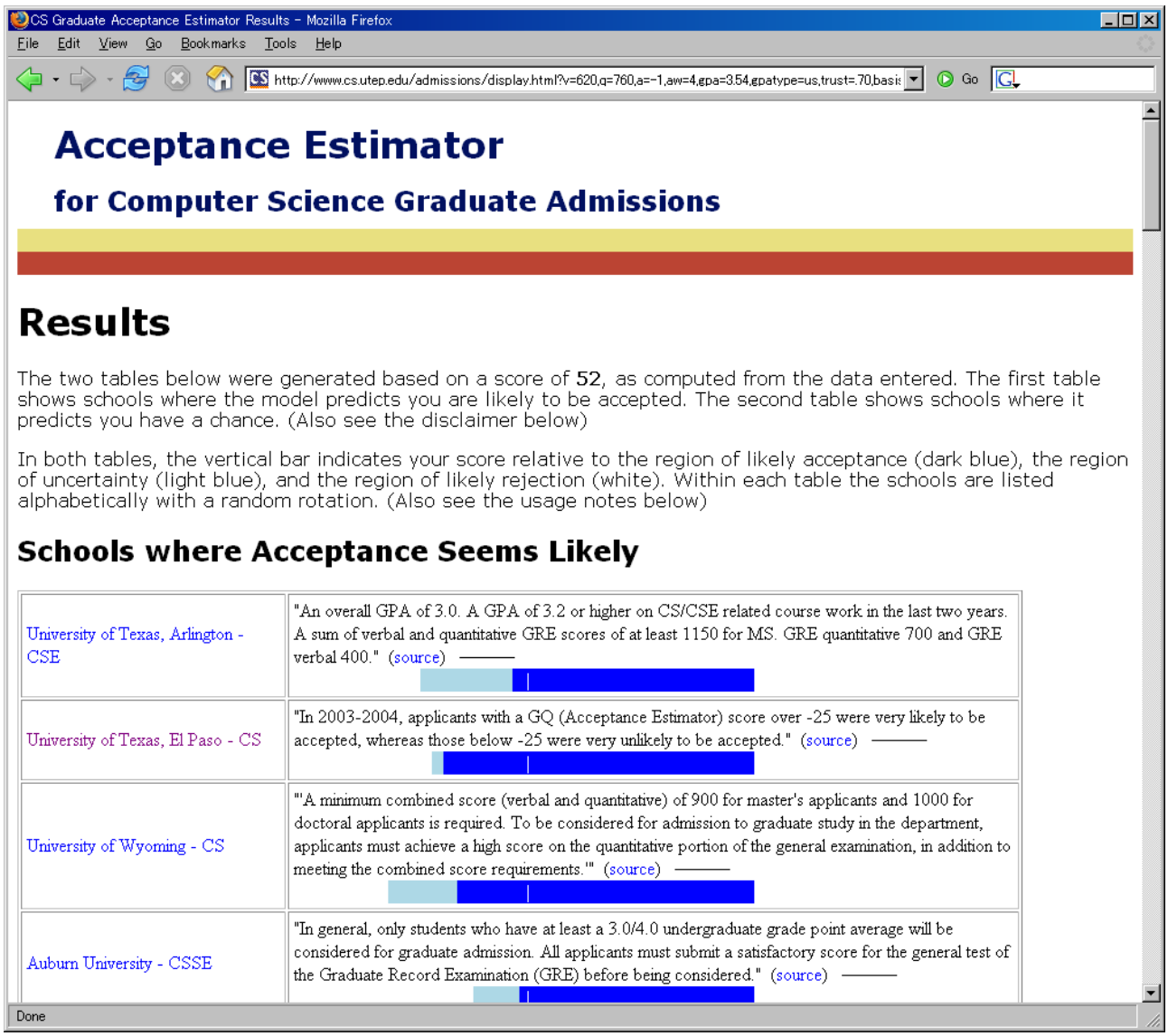

Fig. 5. Results Display