A Probabilistic Rasch Analysis of Question Answering Evaluations

Rense Lange
Integrated Knowledge Systems
renselange@earthlink.net

Warren R. Greiff
The MITRE Corporation
greiff@mitre.org

Juan Moran
University of Illinois, Urbana-Champaign
jmoran@ncsa.uiuc.edu

Lisa Ferro
The MITRE Corporation
lferro@mitre.org

Abstract

The field of Psychometrics routinely grapples with the question of what it means to measure the inherent ability of an organism to perform a given task, and for the last forty years, the field has increasingly relied on probabilistic methods such as the Rasch model for test construction and the analysis of test results. Because the underlying issues of measuring ability apply to human language technologies as well, such probabilistic methods can be advantageously applied to the evaluation of those technologies. To test this claim, Rasch measurement was applied to the results of 67 systems participating in the Question Answering track of the 2002 Text REtrieval Conference (TREC) competition. Satisfactory model fit was obtained, and the paper illustrates the theoretical and practical strengths of Rasch scaling for evaluating systems as well as questions. Most important, simulations indicate that a test invariant metric can be defined by carrying forward 20 to 50 equating questions, thus placing the yearly results on a common scale.

1 Introduction

For a number of years, objective evaluation of state-of-the-art computational systems on realistic language processing tasks has been a driving force in the advance of Human Language Technology (HLT). Often, such evaluations are based on the use of simple sum-scores (i.e., the number of correct answers) and derivatives thereof (e.g., percentages), or on ad-hoc ways to rank or order system responses according to their correctness. Unfortunately, research in other areas indicates that such approaches rarely yield a cumulative body of knowledge, thereby complicating theory formation and practical decision making alike. In fact, although it is often taken for granted that sums or percentages adequately reflect systems' performance, this assumption does not agree with many models currently used in educational testing (cf., Hambleton and Swaminathan, 1985; Stout, 2002). To address this situation, we present the use of Rasch (1960/1980) measurement to the HLT research community, in general, and to the Question Answering (QA) research community, in particular.

Rasch measurement has evolved over the last forty years to rigorously quantify performance aspects in such diverse areas as educational testing, cognitive development, moral judgment, eating disorders (see e.g., Bond and Fox, 2001), as well as olfactory screening for Alzheimer's disease (Lange et al., 2002) and model glider competitions (Lange, 2003). In each case, the major contribution of Rasch measurement is to decompose performance into two additive sources: the difficulty of the task and the ability of the person or system performing this task. While Rasch measurement is new to the evaluation of the performance of HLT systems, we intend to demonstrate that this approach applies here as well, and that it potentially provides significant advantages over traditional evaluation approaches.

Our principal theoretical argument in favor of Rasch modeling is that the decomposition of performance into task difficulty and system ability creates the potential for formulating detailed and testable hypotheses in other areas of language technology. For QA, the existence of a well-defined, precise, mathematical formulation of question difficulty and system ability can provide the basis for the study of the dimensions inherent in the answering task, the formal characterization of questions, and the methodical analysis of the strengths and weaknesses of competing algorithmic approaches. As Bond and Fox (2001, p. 3) explain: "The goal is to create abstractions that transcend the raw data, just as in the physical sciences, so that inferences can be made about constructs rather than mere descriptions about raw data." Researchers are then in a position to formulate

initial theories, validate the consequences of theories on real data, refine theories in light of empirical data, and follow up with revised experimentation in a dialectic process that forms the essence of scientific discovery.

Rasch modeling offers a number of direct practical advantages as well. Among these are:

- Quantification of question difficulty and system ability on a single scale with a common metric.
- Support for the creation of tailor-made questions and the compilation of questions that suit well-defined evaluation objectives.
- Equating (calibration) of distinct question corpora so that systems participating in distinct evaluation cycles can be directly compared.
- Assessment of the degree to which independent evaluations assess the same system abilities.
- Availability of rigorous statistical techniques for the following:
  - analysis of fit of the data produced from systems' performance to the Rasch modeling assumptions;
  - identification of individual systems whose performance behavior does not conform to the performance patterns of the population as a whole;
  - identification of individual test questions that appear to be testing facets distinct from those evaluated by the test as a whole;
  - assessment of the reliability of the test – that is, the degree to which we can expect estimates of systems' abilities to be replicated if these systems are given another test of equivalent questions;
  - identification of unmodeled sources of variation in the data through a variety of methods, including bias tests and analysis of residual terms.

The remainder of the paper is organized as follows. First, we present in section 2 the basic concepts of Rasch modeling. We continue in section 3 with an application of Rasch modeling to the data resulting from the QA track of the 2002 Text REtrieval Conference (TREC) competition. We fit the model to the data, analyze the resulting fit, and demonstrate some of the benefits that can be derived from this approach. In section 4 we present simulation results on test equating. Finally, we conclude with a summary of our findings and present ideas for continuing research into the application of Rasch models to technology development and scientific theory formation in the various fields of human language processing.

2 The Rasch Model for Binary Data

For binary results, Rasch's (1960/1980) measurement requires that the outcome of an encounter between computer systems ($1, \ldots, s, \ldots, n_s$) and questions ($1, \ldots, q, \ldots, n_q$) should depend solely on the differences between these systems' abilities (S_s) and the questions' difficulties (Q_q). Together with mild and standard scaling assumptions, the preceding implies that:

$$P_{sq} = (1 + e^{Q_q - S_s})^{-1} \quad (1)$$

In a QA context, P_{sq} is the probability that a system with the ability S_s will answer a question with difficulty Q_q correctly. For a rigorous derivation of Equation 1 and an overview of the assumptions involved, we refer the reader to work by Fisher (1995). Fisher also proves that sum-scores (and hence *percentages* of correct answers) are sufficient performance statistics *if and only if* the assumptions of the Rasch model are satisfied.

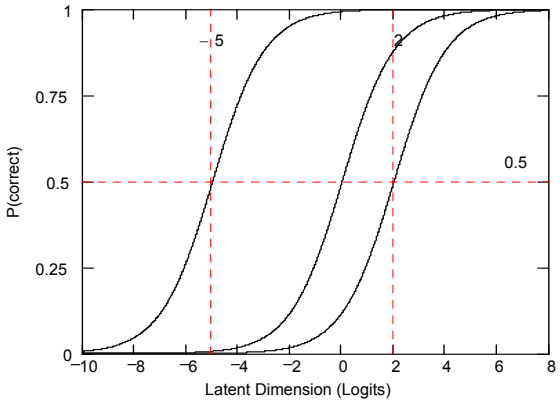

Figure 1. Three Sample Rasch Response Curves

The binary Rasch model has several interesting properties. First, as is illustrated by the three solid lines in Figure 1, Equation 1 defines a set of non-intersecting logistic response-curves such that $P_{sq} = 0.5$ whenever $S_s = Q_q$. In the following, such points are also referred to as question's *locations*. For instance, the locations of the three questions depicted in Figure 1 are -5, 0, and 2. Second, for each pair of systems i and j with $S_i > S_j$, for any question q, system i has a better chance of responding correctly than system j, i.e., $P_{iq} > P_{jq}$. Thus, the questions that are answered correctly by less capable systems always form a probabilistic subset of those answered correctly by more capable systems. Third, restating Equation 1 as is shown below highlights that the question and system parameters are additive and expressed in a common metric:

$$\ln\left(\frac{P_{sq}}{1-P_{sq}}\right) = S_s - Q_q \quad (2)$$

Given the left-hand side of Equation 2, this metric's units are called *Logits*. Note that this equation further implies that S_s and Q_q are determined up to an additive constant only (i.e., their common origin is arbitrary).

Finally, efficient maximum-likelihood procedures exist to estimate S_s and Q_q independently, together with their respective standard errors SE_s and SE_q (see e.g., Wright and Stone, 1979). These procedures do not require any assumptions about the magnitudes or the distribution of the S_s in order to estimate the Q_q, and vice-

versa. Accordingly, systems' abilities can be determined in a "question free" fashion, as different sets of questions from the same pool will yield statistically equivalent S_s estimates. Analogously, questions' locations can be estimated in a "system free" fashion, as similarly spaced Q_q estimates should be obtained across different samples of systems. In this paper, the model parameters, together with their errors of estimate, will be computed via the Winsteps Rasch scaling software (Linacre, 2003).

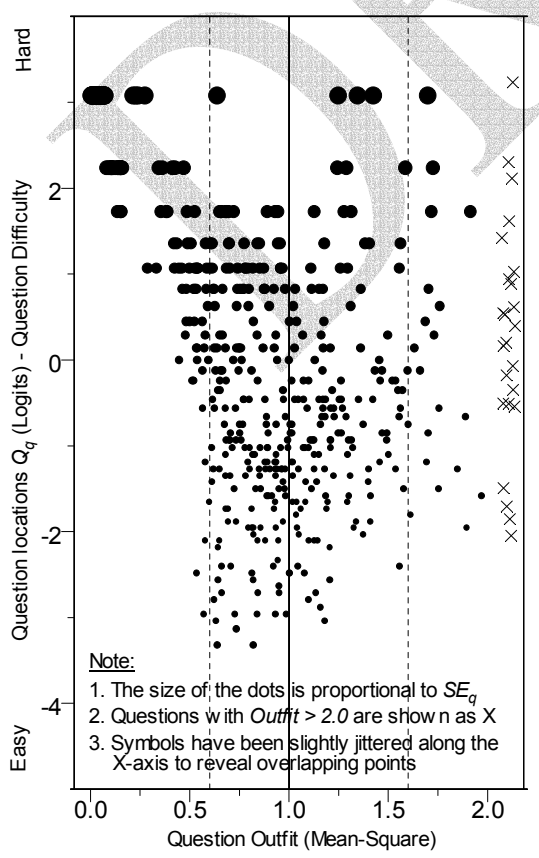

Figure 2. Questions by Q_q, $Outfit_q$, and SE_q

3 Analysis of TREC Evaluation Data

We used the results from the Question Answering track of the 2002 TREC competition to test the feasibility of applying Rasch modeling to QA evaluation. Sixty-seven systems participated, and answered 500 questions by returning a single precise response extracted from a 3-gigabyte corpus of texts. While the NIST judges assessed each answer as *correct*, *incorrect*, *non-exact*, or *unsupported*, we created binary responses by treating each of these last three assessments as *incorrect*. Ten questions were excluded from all analyses, as these were not answered correctly by any system.[1] The final data set thus consisted of 67 systems' responses to 490 questions.

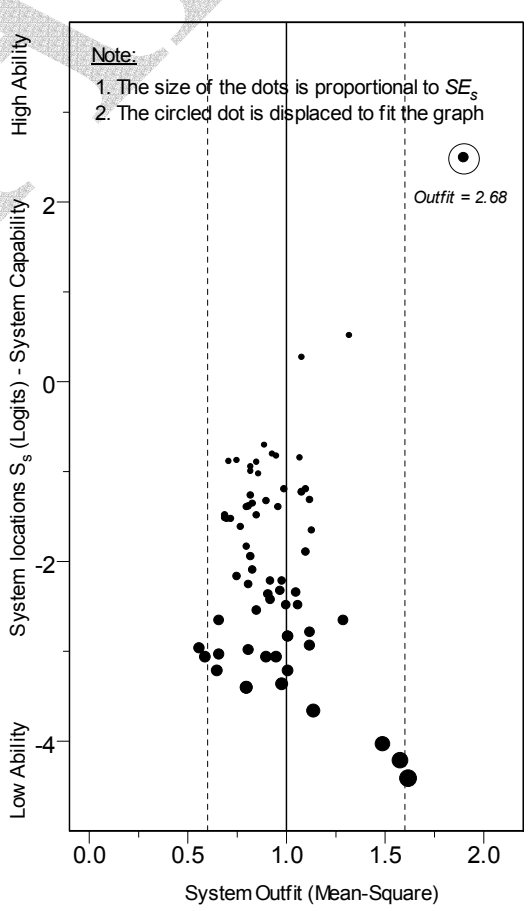

Figure 3. Systems by S_s, $Outfit_s$, and SE_s

3.1 Question Difficulty and System Ability

Maximum-likelihood estimates of the questions' difficulty and the systems' abilities were computed via Winsteps. Figure 2 displays the results associated with the questions, whereas Figure 3 addresses the systems. Each dot in these displays corresponds to one question or one system. For questions, the Y-value gives the estimate of the questions' difficulty (i.e., Q_q); for systems, the Y-value reflects the estimate of systems' ability (S_s). For questions, lower values correspond to easier questions, while higher values to difficult questions. For systems, higher values correspond to greater ability. As is customary, the mean difficulty of the questions is set at zero, thereby implicitly fixing the origin of the estimated system abilities at *-1.98*. As was noted earlier, a

[1] When *all* respondents answer some question q correctly (or incorrectly), the parameter Q_q cannot be estimated. Note that raw-score approaches implicitly ignore such questions as well since including them does not affect the order of systems' "*number right*." Of course, by changing the denominator, *percentages* of right or wrong questions *are* affected.

constant value can be added to each Q_q and S_s without thereby affecting the validity of the results. The X-axes of Figures 2 and 3 refer to the quality of fit, as described in section 3.3 below.

As an example, consider a question with difficulty level -2. This means that a system whose ability level is -2 has a probability of .5 (odds=1) of getting this question correct. The odds of a system with ability of -1 getting this same question correct would increase by a factor[2] of 2.72, thus increasing the probability of a correct answer to $P_{sq} = 2.72/(1+2.72) = .73$. For a system at ability level 0, the odds increase by another factor of 2.72 to 7.39, giving a probability of .88. On the other hand, a system with an ability of -3, would have the even odds decrease by a factor of 2.72 to .369, yielding $P_{sq} = .27$. In other words, increasing (decreasing) questions' difficulties or decreasing (increasing) systems' abilities by the same amounts affects the log-odds in the same fashion. The preceding thus illustrates that question difficulty and system ability have additive properties on the log-odds, or, *Logit,* scale.[3]

The smoothed densities in Figure 4 summarize the locations of the 490 questions (dotted distribution) and the 67 systems (solid). It can be seen that question difficulties range approximately from -3 to +3, and that most questions fall in a region about -1. Systems' abilities mostly cover a lower range such that the questions' locations ($Mean_Q = 0$ *Logits*) far exceed those of the systems ($Mean_S = -1.98$ *Logits*). In other words, most questions are *very* hard for these systems. The vast majority of systems (those located near -1 or below) have only a small chance (below 15%) of answering a significant portion of the questions (those located above 1), and an even smaller chance (below 5%) on a non-negligible number of questions (those above 2). Of those systems, a large portion (those at -2 or below) will have even less of a chance on these questions.

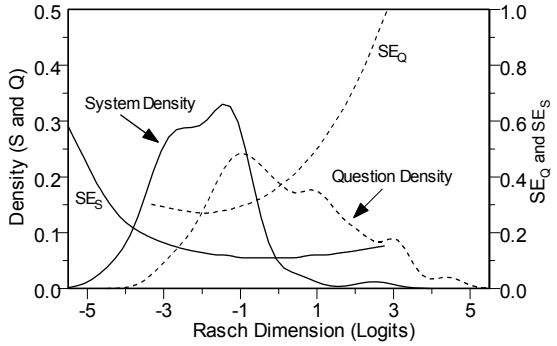

Figure 4. Smoothed System and Question Location Densities

[2] The value of *e,* since we are working with natural logarithms.

[3] Note that a number of measures used in the physical sciences likewise achieve additivity by adopting a log scale.

3.2 Standard Error of Estimate

The two U-shaped curves in Figure 4 reflect the estimates of error, SE_q for questions (dotted curve) and SE_s for systems (solid curve), as a function of their estimated locations (X-axis). As is also reflected by the size of the dots in Figure 3, it can be seen that SE_s is smaller for intermediate and high performing systems (i.e., S_s between -3 and 1 *Logits*) than for low performing systems (S_s < -3 *Logits*). This pattern suits the "horse-race" nature of the TREC evaluation well since the top performing systems are assessed with nearly optimal precision. While the most capable system shows somewhat greater SE_s, calculation shows its performance is still significantly higher than that of the runner up ($z = 10.64$, $p < .001$).

Figure 4 further shows that SE_q increases dramatically beyond -1 *Logits* (this is also reflected in the size of the dots in Figure 2). For instance, the standard error of estimate SE_q exceeds *0.5 Logits* for questions with $Q_q > 1$ *Logit*. Thus, the locations of the hardest questions are known with very poor precision only.

3.3 Question and System Fit

According to the Rasch model, a system, *A,* with middling performance is expected to perform well on the easier questions and poorly on the harder questions. However, it is possible that some system, *B,* achieved the same score on the test by doing poorly on the easy questions and well on the harder questions. While the behavior of system *A* agrees with the model, system *B* does not. Accordingly, the *fit* of system *B* is said to be poor as this system contradicts the knowledge embodied in the data as a whole. Analogous comments can be made with respect to questions. Rasch modeling formalizes the preceding account in a statistical fashion. In particular, for each response to Question *q* by System *s,* Equation 1 allows the computation of a standardized residual z_{sq}, which is the difference between an observed datum (i.e., *0* or *1*) and the probability estimate P_{sq} after division by its standard deviation. Since the z_{sq} follow an approximately normal distribution, unexpected results (e.g., $|z_{sq}|>3$) are easily identified. The overall fit for systems (across questions) and for questions (across systems) is quantified by their *Outfit*.[4] For instance, for System *s*:

$$Outfit_s = \sum_q z_{sq}^2 / (n_q - 1) \qquad (3)$$

Since the summed z^2_{sq} in Equation 3 define a χ^2 statistic with expected value $n_q - 1$, the *Outfit* statistic ranges

[4] Additionally, systems' (or questions') "*Infit*" statistic is defined by weighting the z^2_{sq} contributions inversely to their distance to the contributing questions (or systems). As such, *Infit* statistics are less sensitive to outlying observations. Since this paper focuses on overall model fit, *Infit* statistics are not reported.

from 0 to ∞, with an expected value of 1. As a rule of thumb, for rather small samples such as the present, *Outfit* values in the range 0.6 to 1.6 are considered as being within the expected range of variation.

Figure 2 shows 46 questions whose *Outfit* exceeds *1.6* (those to the right of the rightmost dashed vertical line) and the *Outfit* values of 24 of these exceed 2.0 (shown in the graph by *Xs*, plotted at the right with horizontal jitter). These are questions on which low performing systems performed surprisingly well, and/or high performing systems performed unexpectedly poorly. Thus, there is a clear indication that these questions have characteristics that differentiate them from typical questions. Such questions are worthy of individual attention by the system developers.

Questions and systems with uncharacteristically small *Outfit* values are of interest as well. For instance, in the present context it seems plausible that some questions simply cannot be answered by systems lacking certain capabilities (e.g., pronominal anaphora resolution, acronym expansion, temporal phrase recognition), while such questions are easily answered by systems that possess such capabilities. We might find that these questions would be answered by the vast majority, if not all, of the high performing systems and very few if any of the low ability systems. The estimated fit would be much better (small *Outfit*) than expected by chance. Again, the identification and analysis of such overfitting questions and, similiarly, overfitting systems may greatly enhance our understanding of both.

3.4 Example of System with large *Outfit*

Note that Figure 3 above shows that the best performing system also exhibits the largest *Outfit (2.68)*, and we investigated this system's residuals in detail. Table 1 indicates that this system failed (*Datum = 0*) on eight questions (*q*) where its modeled probability of success was very high ($P_{sq} > 0.98$). Thus, the misfit results from this system's failure to answer correctly questions that proved quite easy for most other systems.

q	Q_q	*Datum*	P_{sq}	*Residual*	*z*
1411	-1.51	0	0.98	-0.98	-7.39
1418	-1.96	0	0.99	-0.99	-9.26
1465	-1.74	0	0.99	-0.99	-8.28
1672	-1.59	0	0.98	-0.98	-7.67
1671	-1.51	0	0.98	-0.98	-7.39
1686	-2.11	0	0.99	-0.99	-9.97
1697	-1.89	0	0.99	-0.99	-8.92
1841	-1.66	0	0.98	-0.98	-7.97

Table 1. Misfit Diagnosis of Best Performing System

These are the eight questions listed in Table 1:

1411 What Spanish explorer discovered the Mississippi River?
1418 When was the Rosenberg trial?
1465 What company makes Bentley cars?
1642 What do you call a baby sloth?
1671 Where is Big Ben?
1686 Who defeated the Spanish armada?
1697 Where is the Statue of Liberty?
1841 What's the final line in the Edgar Allen Poe poem "The Raven?"

This situation should be highly relevant to the system's developers. Informally speaking, the best system studied here "should have gotten these questions right," and it might thus prove fruitful to determine exactly why the system failed. Even if no obvious mistakes can be identified, doing so could reveal strategies for system improvement by focusing on seemingly "easy" issues first. Alternatively, it might turn out that precisely those aspects of the system that enable it do so well overall also cause it to falter on the easier questions. Ascertaining this might or might not be of help to the system's designers, but it would certainly foster the development of a scientific theory of automatic question answering.

3.5 Impact of Misfit

The existence of misfitting entities raises the possibility that the estimated Rasch system abilities are distorted by the question and system misfit. We therefore recomputed systems' locations by iteratively removing the worst fitting questions until 372 questions with $Outfit_q < 1.6$ remained. In support of the robustness of the Rasch model, Figure 5 shows that the correlation between the two sets of estimates is nearly perfect ($r = 0.99$), indicating that the original and the "purified" questions produce essentially equivalent system evaluations. Thus, the observed misfit had negligible effect on the system parameter estimates.

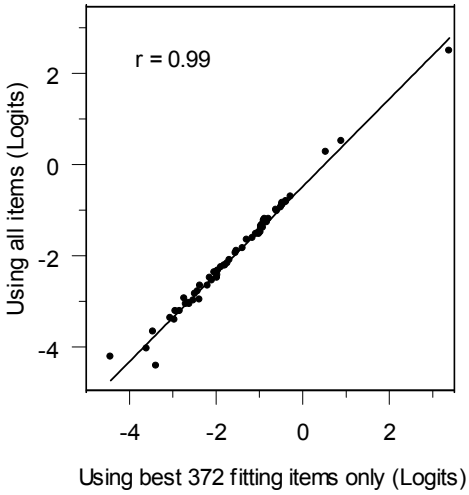

Figure 5. Effect of Removing Misfitting Questions on the Estimated System Abilities S_s

4 Test Equating Simulation

A major motivation for introducing Rasch models in educational assessment was that this allows for the calibration, or *equating,* of different tests based on a limited set of common (i.e., repeated) questions. The purpose of equating is to achieve equivalent test scores across different test sets. Thus, equating opens the door to calibrating the difficulty of questions and the performance of systems across the test sets used in different years.

Since appropriate data from different years are lacking, a simulation study was performed based on different subsets of the 490 available questions. We show how system abilities can be expressed in the same metric, even though systems are evaluated with a completely different set of questions. To rule out the possibility that such a correspondence might come about by chance (e.g., equally difficult sets of questions might accidentally be produced), a worst-case scenario is used. The simulation also provides a powerful means to demonstrate the superiority of the Rasch *Logit* metric compared to raw scores as indices of system performance.

To this end, we divide the available questions from TREC 2002 into two sets of equal size. The *Easy* set contains the easiest questions (lowest Q_q) as identified in earlier sections. For the simulation, this will be the test set for one year's evaluation. A second, *Hard* set, serves as the test set for a subsequent evaluation, and it contains the remaining 50% of the questions (highest Q_q). By design, the difference in questions' difficulties is far more extreme than is likely to be encountered in practice. The Rasch model is then fitted to the responses to the *Easy* set of questions. Next, based on questions' difficulties and their fit to the Rasch model, a subset of the *Easy* questions is selected for inclusion in the second test in conjunction with the *Hard* question set. These questions are said to comprise the *Equating* set, as they serve to fix the overall locations of the questions in the *Hard* set.

Normally, this second test would be administered to a new set of systems (some completely new, others improved versions of systems evaluated previously). However, for the purposes of this simulation, we "administer" the second test to the same systems. The Rasch model is then applied to the *Hard* and *Equating* questions combined, while fixing the locations of the *Equating* questions to those derived while scaling the *Easy* set. The Winsteps software achieves this by shifting the locations in the *Hard* set to be consistent with the *Equating* set – but without adjusting the *spacing* of the questions in the *Hard* or *Easy* sets. If the assumptions of the Rasch model hold, then the *Easy* and *Hard* question sets will now behave as if their levels had been estimated simultaneously. Since the same systems are used in the simulation, and the questions have been calibrated to be on the same scale, the estimated system abilities S_s as derived from the *Easy* and *Hard* question sets should be statistically identical. That is, these two estimates should show a high linear correlation and they should have very similar means and standard deviations (see e.g., Wright and Stone, 1979, p. 83-126).

Common wisdom in the Rasch scaling community holds that relatively few questions are needed to achieve satisfactory equating results. For this reason, the simulation study was performed three times, using *Equating* sets with 20, 30, and 50 questions, respectively (i.e., about 4, 6, and 10% of the total number of questions in the present study).

4.1 Findings

The simulation results are summarized in Table 2, whose rows reflect the sizes of the respective *Equating* sets (i.e., 20, 30, and 50). Each first sub-row reports the Rasch scaling results, while the second sub-row reports the raw-score (i.e., number correct) findings. The columns report a number of basic statistics, including the mean (*M*) and standard deviations (*SD*) of the *Logit* and raw-score values in the *Easy* and *Hard* sets, and the correlation (r_{linear}) between systems' estimated abilities based on the *Easy* and *Hard* sets.

Size of Equating Set	Index	M_{easy}	SD_{easy}	M_{hard}	SD_{hard}	r_{linear}
20	*Rasch*	-0.66	1.10	-0.65	1.23	0.90
	# Correct	92.40	47.53	27.39	30.70	0.77
30	*Rasch*	-0.68	1.10	-0.66	1.21	0.92
	# Correct	92.88	47.92	29.82	31.80	0.80
50	*Rasch*	-0.78	1.11	-0.77	1.18	0.94
	# Correct	94.76	49.62	31.01	32.29	0.82

Table 2. Results of the Simulation Study

The major findings are as follows. First, inspection of the r_{linear} columns indicates that Rasch equating consistently produced higher correlations between systems' estimated abilities as estimated via the *Easy* and *Hard* question sets than did the raw scores for each set. Second, for obvious reasons the raw-score estimates based on the *Easy* sets are considerably higher than those based on the *Hard* sets. However, Table 2 also shows that the standard deviations of the number correct estimates obtained for the *Easy* sets exceed those of the *Hard* sets as well (sometimes by over 100%). In other words, when raw scores (or percentages) are used, the "spacing" of the systems is affected by the difficulty of the questions.

Third, the Rasch approach by contrast produces very similar means and standard deviations for the capability estimates based on the *Easy* and *Hard* question sets. This holds regardless of the size of the *Equating* sets. For instance, when 50 equating questions are used, the estimated abilities based on the *Easy* and *Hard* sets have nearly identical SD (i.e., 1.11 and 1.18 *Logits*, respectively). The corresponding averages for this case are -0.78 and -0.77 *Logits*, i.e., a standardized difference (effect size) of less than 0.01 *SD*. Similarly small effects sizes are obtained for the other cases. Further, given the superior values of r_{linear}, it thus appears that Rasch equating provides an acceptable equating mechanism even when maximally different question sets are used. In fact, already for *Equating* sets of size 20 a correlation of 0.90 is produced.

Fourth, as a check on the results, scatter plots of the various cases summarized in Table 2 were produced. The left panel of Figure 6 shows the Rasch capability estimates obtained for the *Hard* question set plotted against those for the *Easy* set, and it can be seen that these estimates are highly correlated ($r_{linear} = 0.94$). The corresponding raw scores are plotted in the right panel of Figure 6. In addition to showing a lower correlation ($r_{linear} = 0.82$), raw scores also clearly posses a non-linear component, and in fact the quadratic trend is highly significant ($t_{quad} = 13.10,\ p < .001$). Thus, in addition to being question-dependent, raw score and percentage based comparisons suffer from pronounced non-linearity.

Despite the favorable results, we remind the reader that the above simulations represented a worse-case scenario. Indeed, more realistic simulations not reported here indicate that Rasch equating can further be improved by omitting misfitting questions and by using less extreme question sets.

5 Conclusions

In this paper we have described the Rasch model for binary data and applied it to the 2002 TREC QA results. We addressed the estimation of question difficulty and system ability, the estimation of standard errors for these parameters, and how to assess the fit of individual questions and systems. Finally, we presented a simulation which demonstrated the advantage of using Rasch modeling for calibration of question sets.

Based on our findings, we recommend that test equating be introduced in formal evaluations of HLT. In particular, for the QA track of the TREC competition, we propose that NIST include a set of questions to be reused in the following year for calibration purposes. For instance, after evaluating the systems' performance in the 2004 competition, NIST would select a set of questions consistent with the criteria outlined above. Using twenty to fifty questions from a set of 500 will probably be sufficient, especially when misfitting questions are eliminated. When the results are released to the participants, they would be asked not to look at these equating questions, and not to use them to train their systems in the future. These equating questions would then be included in the 2005 question set so as to place the 2004 and 2005 results on the same *Logit* scale. The process would continue in each consecutive year.

The approach outlined above serves several purposes. For instance, the availability of equated tests would increase the confidence that the testing indeed measures progress, and not simply the unavoidable

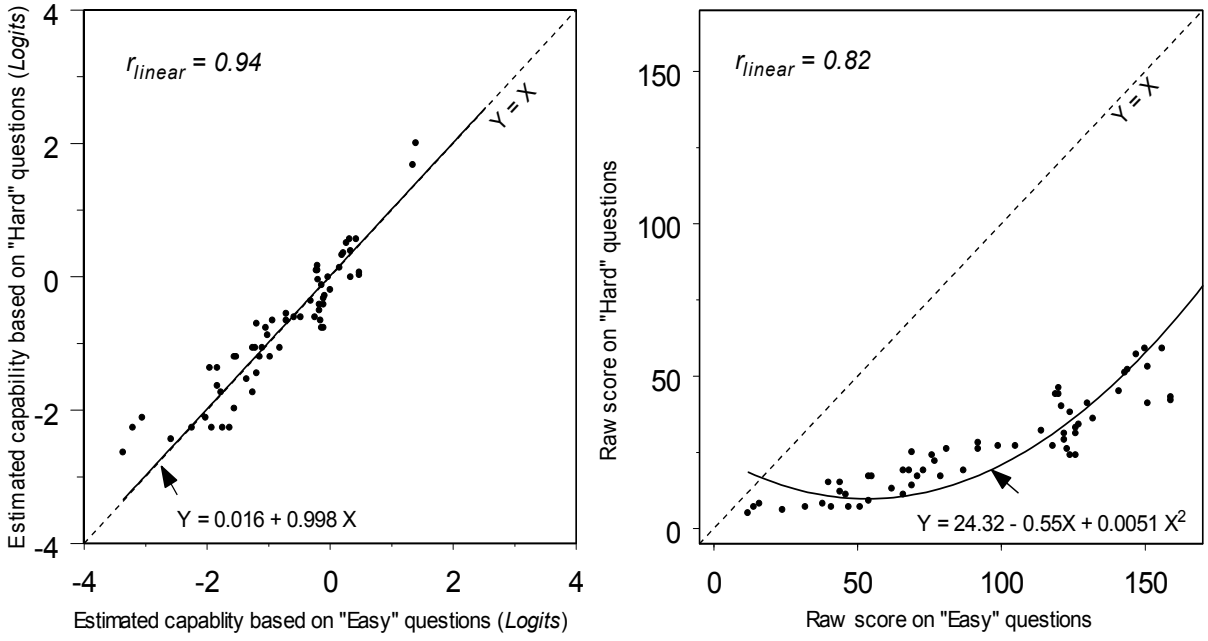

Figure 6. Systems' Performance on *Easy* vs. *Hard* Questions Based on Rasch Scaling (left) and Raw Scores (right)

variations in difficulty across each year's question set. Additionally, it would support the goal of making each competition increasingly more challenging by correctly identifying easy and difficult questions. Further, calibrated questions could be combined into increasingly large corpora, and these corpora could then be used to provide researchers with immediate performance feedback *in the same metric as the NIST evaluation scale*. The availability of large corpora of equated questions might also provide the basis for the development of methods to predict question difficulty, thus stimulating important theoretical research in QA.

The work presented here only begins to scratch the surface of adopting a probabilistic approach such as the Rasch model for the evaluation of human language technologies. First, as was discussed above, questions displaying unexpectedly large or small *Outfit* values can be identified for further study. The questions themselves can be analyzed in terms of both content and linguistic expression. With the objective of beginning to form a theory of question difficulty, questions can be analyzed in concert with the occurrence of correct answers in the document corpus and the incorrect answers returned by systems. Also, experimentation with more complex scaling models could be conducted to uncover information other than questions' difficulty levels. For example, so-called 2-parameter IRT models (see e.g., Hambleton and Swaminathan, 1985) would allow for the estimation of a *discrimination* parameter together with the *difficulty* parameter for each question. More direct information concerning the diagnosis of systems' skill defects are described in Stout (2002).

It is also possible to incorporate into the model other factors and variables affecting a system's performance. Rasch modeling can be extended to many other HLT evaluation contexts since Rasch measurement procedures exist to deal with multi-level responses, counts, proportions, and rater effects. Of particular interest is application to technology areas that use metrics other than percent of items processed correctly. Measures such as average precision, R-precision and precision at fixed document cutoff, which are used in Information Retrieval (Voorhees and Harman, 1999), metrics such as BiLingual Evaluation Understudy (BLUE) (Papineni et al., 2002) used in Machine Translation, and F-measure (Van Rijsbergen, 1979) commonly used for evaluation of a variety of NLP tasks are just a few of the variety of metrics used for evaluation of language technologies that can benefit from Rasch scaling and related techniques.

References

Bond, T.G. and Fox, C.M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* New Jersey: Lawrence Erlbaum Associates.

Fischer, G.H. (1995). Derivations of the Rasch model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. (pp. 15-38) New York: Springer.

Hambleton, R.K. and Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer – Nijhoff.

Lange, R. (2003). Model Sailplane Competition: From Awarding Points to Measuring Performance Skills. *RC Soaring Digest*, August Issue. (This paper is also available as: http://www.iknowsys.org/Download/Model_Sailplanes.pdf).

Lange, R., Donathan, C.L., and Hughes, L.F. (2002). Assessing olfactory abilities with the University of Pensylvania smell identification test: A Rasch scaling approach. *Journal of Alzheimer's Disease*, *4*, 77-91.

Linacre, J. M. (2003). *WINSTEPS Rasch measurement computer program*. Chicago, IL: Winsteps.com.

Papineni, K., Roukos, S., Ward, T, Henderson, J. and Reeder F. (2002). Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results. *Proceedings of the 2002 Conference on Human Language Technology* (pp. 124-127). San Diego, CA, 2002.

Stout W.F. (2002). Psychometrics from practice to theory and back. *Psychometrika*, *67*, 485-518. http://www.psychometricsociety.org/journal/online/ARTICLEstout2002.pdf

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: MESA Press.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. Dept. of Computer Science, University of Glasgow.

Voorhees, E. M. and Harman, D. K. (eds.). 1999. *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, NIST Special Publication 500-246.

Wright, B.D. and Stone, M.H. (1979). *Best test design*. Chicago, IL: MESA Press.