

A pilot study of job satisfaction in Massachusetts judges

BY DONNA M. NORRIS, M.D., MICHAEL LAMPORT COMMONS, PH.D., PATRICE M. MILLER, ED.D., KATHRYN MARIE ADAMS, B.S., AND THOMAS G. GUTHEIL, M.D.

This pilot study examined the overall job satisfaction of Massachusetts judges, and additionally addressed these judges' views of what might serve to increase their job satisfaction. Results indicated that these judges were highly satisfied with their jobs, and that they viewed increased pay as the most important contributor to increased job satisfaction, followed by improvements in professional support staff. Various other factors, including geographic location and the availability of computer networking, were viewed by these judges as relatively unimportant.

KEY WORDS: *Courts, judges, job satisfaction, Rasch analysis.*

Numerous scientific studies and media reports have focused on occupational stress and job satisfaction for lay and professional workforces (Donkin, 2005; Eisenstat & Felber, 1984; Geller, 2005; Graham, Ramirez, Field, & Richards,

AUTHORS' NOTE: *For additional information about this article, please contact: Donna M. Norris, M.D., 5 Shattuck Street, Boston, MA 02115. E-Mail: dmnorrismd@verizon.net. The authors wish to acknowledge their indebtedness to the Program in Psychiatry and Law, Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School. The views expressed in this article are the sole responsibility of the authors.*

© 2011 by Federal Legal Publications, Inc.

2000). For example, in a survey of psychiatric nurses, the number of stressful events on the job correlated with a negative sense of the work environment, but not with the work experience. When the work was interesting, employees did not experience significant turnover, even in the face of stressful events (Zautra, Eblen, & Reynolds, 1986).

There are relatively few studies that address occupational stress and job satisfaction for the judiciary. In the early 1980s, Walter Menninger surveyed American judges and concluded that judicial stress was not all bad, as “if you learn to deal with stress and master it, stress can make you stronger” (Middleton, 1981, p. 1100). Eells and Showalter (1994) found the most significant stressors for another sample of American judges to be “poorly prepared or disrespectful counsel,” “exercising judicial management and discretion,” and “highly emotional cases under public scrutiny” (p. 71). A study of Canadian judges conducted by Rogers, Freeman, and LeSage (1994) identified child custody cases, jury trials, and sentencing as particularly stressful responsibilities.

The paths to becoming a judge are multidetermined, and often demand personal skills and knowledge unrelated to the practice of law. In some states, candidates for judgeship must stand for election, conduct a successful campaign, receive a majority of the popular vote, and remain subject to public confirmation and recall. Other states employ an appointment system—for example, in Massachusetts, qualified individuals are initially vetted by a nonpartisan Judicial Nominating Commission, after which the selected nominees are appointed by the Governor (Massachusetts Exec. Order No. 420, 2006).

This article addresses the results of a survey of Probate and Superior Court Judges in the Commonwealth of Massachusetts. These courts are of comparable rank. The jurisdiction of the Probate Court includes general equity, estates,

wills, trust, guardianship, conservatorship, competency related to medical treatment decisions, and domestic matters including adoption, paternity, divorce and custody. The jurisdiction of the Superior Court includes felonies and civil matters involving amounts of \$25,000 or more, including equity release.

A questionnaire was designed to assess these judges' perceptions of the following factors: (a) satisfaction or dissatisfaction with the physical environment; (b) financial compensation; (c) presence and adequacy of both professional and nonprofessional staff; and (d) adequacy of other support services provided (e.g., personal computers). Overall, we sought to examine whether or not judges are satisfied with their jobs, despite the variety of personal and professional challenges they face on a daily basis.

Rasch Model

In this pilot study, we employed a Rasch analysis to establish the degree of perceived satisfaction in an objective, empirical manner. The Rasch Model is used in a variety of disciplines and for a wide range of topics, including issues in psychiatry and the law. Instructive examples include studies conducted by Dattilio, Commons, Adams, Gutheil, and Sadoff (2006), and Commons, Miller, and Gutheil (in press).

The Rasch Model produces an objective, additive scale that can be used to analyze a large variety of human sciences data. Through the use of probabilistic equations, raw ratings of items are converted into an equal-interval scale. This linear scale (Wright & Stone, 1979) can be used as a type of objective ruler against which to measure data on the survey items as well as on the respondents (Andrich, 1988). Further details about the Rasch Model are provided in Appendix 1.

Method

Participants In 2004, the first author—an expert witness who often consults in the probate courts of Massachusetts—contacted the Chief Judges of the Probate Courts and the Superior Courts of Massachusetts to discuss the study proposal. Permission was obtained to conduct an anonymous survey of all sitting judges in these Courts regarding their respective court environments.

Of the questionnaires returned, 30 were from Probate Court judges (68.2% of the overall sample, and 60.0% of the 50 judges sitting in these courts), and 14 were from Superior Court judges (31.8% of the overall sample, and 18.6% of the 75 judges from the Superior Courts).

The majority of respondents were male (28, or 63.6%) as opposed to female (16, or 36.4%), and their mean age was 56.4 years ($SD = 6.3$ years).

On the whole, respondents were predictably well experienced in the law, having practiced for an average of 18.6 years ($SD = 8.8$ years), and were also well experienced as judges, having served in that capacity for an average of 9.36 years ($SD = 4.9$ years). Current respondent workloads were variable, but typically quite substantial, with an average assignment of 303.4 cases per month ($SD = 232.0$).

The population served by these courts appeared to represent a broad socioeconomic class range. Respondents estimated that 33.6% ($SD = 20.5$) of litigants were poverty class, that 46.3% ($SD = 19.7$) were working class, and that 26.0% ($SD = 12.4$) were professional class.

Instrument Respondents were asked a variety of questions about their courtroom experiences. Initial questionnaire items were generated on the basis of the collective experience of the forensic experts conducting this study. These items were then reviewed and edited by members of the Program in

Psychiatry and the Law, Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, as well as by the Chief Judges of both courts. The majority of items directed respondents to rate their impressions on a scale of 1 to 6, with 1 as the negative endpoint of the scale (e.g., “never,” “not at all,” or “no”) and 6 as the positive end of the scale (e.g., “always,” “totally,” or “yes”), with a further option to embellish answers in narrative form.

Procedure

The questionnaires, each with a self-addressed, postage prepaid envelope, were sent to the state’s main judicial offices (at the Chief Judges’ suggestion, and to maintain the respondents’ anonymity). From this office they were forwarded to 50 Probate Court judges and to 75 Superior Court judges. The researchers, with their credentials and affiliations, were identified to the Chief Judges but not to the respondents.

To maintain the respondents’ anonymity still further, the completed surveys were returned by mail to the second author—who had not appeared as an expert witness in any of these courts for over 8 years—at a private address not identified with any of the academic settings of the study researchers. To increase the survey return, a second mailing to all of the judges was mailed from the Chief Judges’ offices in exactly the same manner.

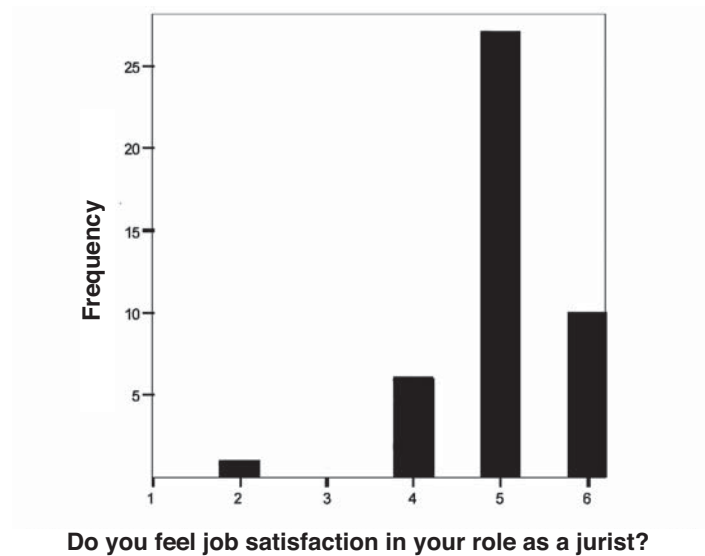
Although such demographic data as age, gender and type of court were collected, there would have been little or no possibility of matching a questionnaire to a particular respondent. Questionnaires were arbitrarily assigned to subject numbers, based upon the order of their receipt.

Results

The primary finding of this study is that judges depicted themselves as highly satisfied. On a scale of 1 (“dissatisfied”) to 6 (“completely satisfied”), the vast majority of the

sample rated their satisfaction as 5 or 6 (see Figure 1). There was only one judge who was quite dissatisfied, giving a rating of just 2.

FIGURE 1 **Frequency of respondent answers to the question, “Do you feel job satisfaction in your role as a jurist?” on a scale of 1 (dissatisfied) to 6 (completely satisfied).**



For the sample as a whole, the mean satisfaction level of 5.02 ($SD = .762$) was significantly higher than the value of 1 or “dissatisfied”: $t(43) = 35.011$, $p = .004$. This item had a large effect size, $d = 5.28$, therefore accounting for a large amount of the item’s variability (Cohen, 1988). Effect size is a measure of how much of the variability is accounted for by the item(s) under study. Variability is made up of two parts. One part is due to the item(s) under study and the second part is due to noise, which is random variation. The effect size tells how large the known source of variability (the item variability) was compared to the random noise. There are several different ways to measure effect size. In this case, effect size was calculated using: $d = (m - c) / \sigma$, where m = the estimated population mean, c = the expected

population mean, and σ = the estimated standard deviation of the population (Wright & Stone, 1979).

There were no differences in overall satisfaction between judges serving in Probate Court ($M = 4.93$; $SD = .583$) and judges serving in Superior Court ($M = 5.21$; $SD = 1.05$; $t(42) = -1.14$, $p = \text{n.s.}$).

While the judges expressed general satisfaction with their work, there could be specific factors that might add to or detract from that satisfaction. The next analysis examined to what extent different aspects of the work situation might make a positive or negative difference in the satisfaction of respondents. The mean ratings for each item and the t test results will be discussed first, and then the Rasch Analysis. The results for both types of analyses are shown in Table 1, and each will be discussed in turn.

For each of the items shown, a one-sample t test was conducted, assessing the mean rating against a fixed value. For the questions that provided a 1–6 scale, where 1 = “never” and 6 = “always,” a fixed value of 1 was used to assess whether or not the item means differed significantly from the lowest value on the scale.

Of eight sample means tested against 1, all differed significantly from that fixed value. The reported p value for each item was adjusted using the Bonferroni correction (Rosenthal & Rosnow, 1985), which is used when multiple statistical tests are calculated at one time. This correction states that the p value for each item should be multiplied by the number of items that were tested against the given t value. Note that the item asking about the overall satisfaction level felt by the judges is also included in this table and was included in the Rasch analysis as well. It serves as a comparison point for the other items.

As seen from the means in Table 1, respondents felt that an improvement in financial compensation would make the

largest positive difference. The next most important improvement was in the area of professional support staff. A more neutral rating was given to improvements in security staff and secretarial support staff. Improvements in networked personal computers were even less important. Lastly, an improvement in geographical area would also reportedly make only a very small difference.

In order to assess overall stress levels felt by participants in the work arena, respondents were also asked “how frequently do you feel that you are under stress?” where 1 = “never” and 6 = “always.” The mean amount of stress was not significantly above the indifference mid-point of 3.5, but was significantly greater than 1 ($M = 3.78$, $SD = 1.338$, $t(31) = 11.762$, $p = .004$, $d' = 2.07$). The amount of stress a respondent reported was not correlated with their job satisfaction in this sample, $r(43) = -0.157$, $p = \text{n.s.}$

Finally, since the questionnaire also asked participants about their caseloads and the typical population with which they worked, this first step in the analysis used correlations (not shown in Table 1) to examine the relationship between job satisfaction and these variables. Caseload, as measured by the number of cases per month that a judge estimated overseeing, was positively correlated with job satisfaction— $r(35) = 0.33$ —but just missed statistical significance ($p = .053$). Caseload was not related to feeling stressed: $r(34) = 0.20$, $p = .26$, n.s. As to the effects of the type of caseloads seen in a judge’s courtroom, in those courts in which judges reported seeing predominantly professional class participants there was a significant positive correlation with job satisfaction, $r(38) = 0.401$, $p < .013$.

To provide more insight into the degree to which a given situation would affect job satisfaction in respondents, a Rasch Analysis using Winsteps software (Linacre, 2004) was carried out on these items. The t test results showed whether the raw ratings of items differed significantly from a fixed

TABLE 1

Mean, standard deviation, T test significance, effect size, Rasch scaled score, and infit MNSQ for each questionnaire item relating to job satisfaction.

<i>Item</i>	<i>Mean</i>	<i>Standard Deviation</i>	<i>p value</i>	<i>Effect Size</i>	<i>Rasch Scaled Score</i>	<i>Infit MNSQ</i>
Do you feel job satisfaction in your role as a jurist?	5.02	0.762	.004	1.99	-0.80	0.78
How much of an improvement in [financial compensation] would make a positive difference in your daily work environment?	4.77	1.309	.004	.970	-0.57	1.39
How much of an improvement in [professional support staff] would make a positive difference in your daily work environment?	4.39	1.528	.004	.582	-0.29	0.86
How frequently do you feel that you are under stress?	3.74	1.329	.004	.181	0.09	0.72
How much of an improvement in [secretarial support staff] would make a positive difference in your daily work environment?	3.74	1.747	.004	.137	0.09	0.90
How much of an improvement in [security staff] would make a positive difference in your daily work environment?	3.64	1.672	.004	.084	0.16	0.71
How much of an improvement in [networked personal computers] would make a positive difference in your daily work environment?	2.89	1.742	.004	-.350	0.57	1.17
How much of an improvement in [geographic location] would make a positive difference in your daily work environment?	2.57	1.771	.004	-.525	0.76	1.38

value, therefore indicating that item's responses had a low probability of occurring by chance. In contrast, the Rasch analysis linearly ordered the items according to degree of satisfaction. This allows an exact determination of the spacing between items, that is, specifically how much more important is one item than another in contributing to satisfaction. For a Rasch analysis, this is a small sample. Nevertheless, a sample size of just 30 assures 95% confidence, so there should not be excessive concern about having "only" 44 respondents.

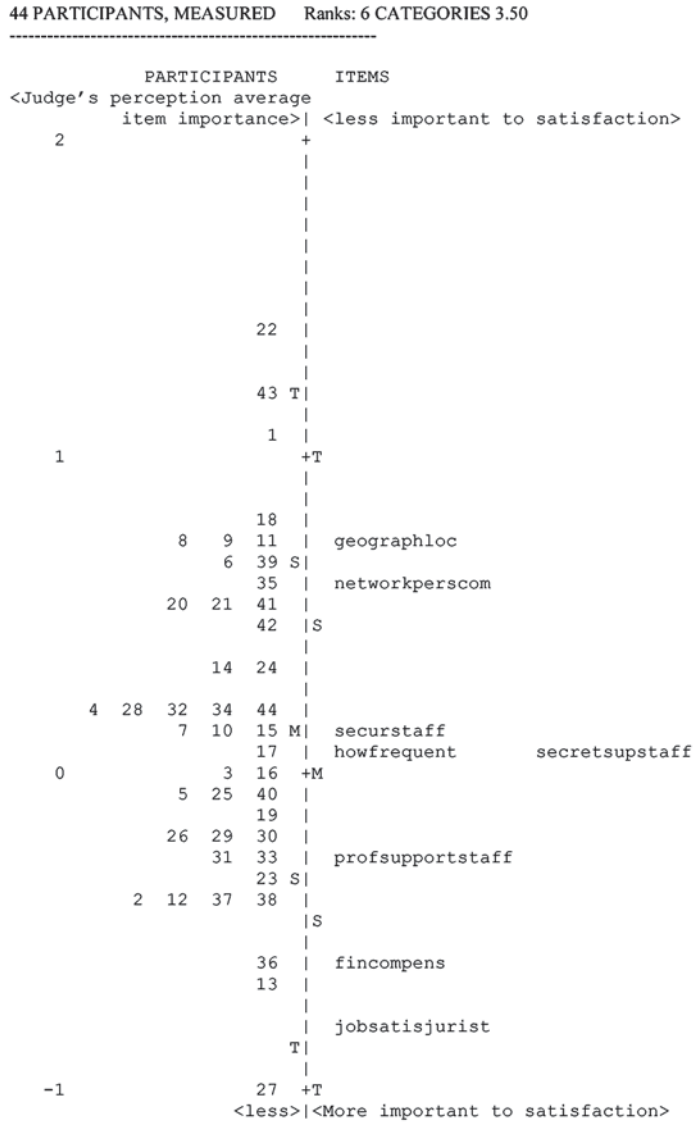
There were eight items entered into the Rasch analysis, as shown in Table 1. These items fit on a single dimension, which was named the degree of importance for satisfaction. This is seen by looking at the Rasch scaled scores and the accompanying Infit scores. To ascertain whether an item fits on the underlying dimension, there is a criterion of rejecting items with infit errors larger than 2.0 (M. Linacre, personal communication, January, 2003). None of the items had an infit error greater than 2.0, indicating that all item-responses fit the scale in a way that allowed for determination of the degree to which they contributed to satisfaction. Furthermore, each of the infit scores was approximately 1.00, indicating that the items fit nearly perfectly. This suggests that participants rated items in a consistent manner.

Whereas Table 1 shows the specific values of the Rasch scaled scores, the nature of the scale itself can be better understood by looking at the positions of the scores, shown in Figure 2. The right side of Figure 2 places each item, according to how positively or negatively—and to what extent—it contributed to job satisfaction. In Figure 2, each item is represented by an abbreviated variable name.

The item that asked participants to rate the following questions on a scale of 1 ("dissatisfied") to 6 ("completely satisfied"), "do you feel job satisfaction in your role as a jurist?" had the lowest scaled score, -0.80, therefore allowing a

FIGURE 2 **Rasch map of job satisfaction in jurists.**

There were 44 participants, 8 items, and 6 categories. On the left of the vertical line are the judges' scaled scores. These indicate the degree to which a judge perceived the average item to make a difference in job satisfaction. The greater the difference, the more negative the score. On the right of the vertical line are the item scaled scores as to perceive determinants of satisfaction of the items. Again, the more negative the score, the more effect upon satisfaction.



determination that this end of the scale (the negative end) represented satisfaction. In other words, the items at the bottom of the scale in Figure 2 (with negative scaled scores) are those that were rated as being the most important in determining satisfaction. The more positive was an item's scaled score, the less this item contributed to job satisfaction of respondents. It was found that the most important item was financial compensation, having a scaled score equal to -0.57. This was followed by improvements in professional support staff (-0.29), secretarial support staff (0.09), and four other items referring to how frequently they felt under stress (0.09), security staff (0.16), networked personal computers (0.57), and geographic location (0.76).

A Rasch analysis provided informative findings regarding the aspects of job satisfaction that this sample of judges valued most. The participant scaled scores were also very informative about the ways in which respondents rated each item. Although the item range was from approximately -1 to +1 (see the right hand side of the scale in Figure 2), the participant range was from approximately -1 to +1.5 (see the left hand side of the scale in Figure 2). This indicates that the majority of respondents rated the items in a similar manner, and their ratings fell within one standard deviation of the mean item rating (indicated by "M," just to the right of the line in the middle of the Figure).

There were a few respondents who felt that none of the items were important to their job satisfaction. In particular, participants 22, 43, and 1 (see Figure 2) had scaled scores that were significantly higher on the scale (more toward the positive end) than the item that was rated as being the least important (geographical location). There were also three participants (36, 13 and 27) who had scaled scores that were much more toward the negative end of the scale, suggesting that they had rated more of the items as needing improvement. Because the distribution of participant scaled scores indicates where respondents fit in regard to item ratings

(participant scaled scores are approximately at the same point of the scale as the item that they rated as being the most important), it is possible to determine that the first three respondents discussed above rated items as being less important than did other respondents, while the second three respondents rated items as being more important than the other respondents.

In order to further illustrate the linearity of the Rasch scale, one can examine the items that are approximately one unit apart on the scale (see Figure 2). The item regarding overall job satisfaction (“jobsatisfjurist”) had a scaled score of -0.80, and was approximately one unit away from the items that asked about the effects of secretarial support staff on job satisfaction (“secretsupstaff”) and the frequency of job-related stress (howfrequent), which both had scaled scores of .09.

Furthermore, the item that asked about the effects of professional support staff on job satisfaction (“profsupportstaff”) had a scaled score of -0.29 and was approximately one unit away from the item that asked about the effects of geographic location on job satisfaction (“geographic”), which had a scaled score of 0.76. This relationship not only infers linearity within the scale, but also gives us a better idea of how the items were rated relative to each other, as well as the magnitude of difference between the ratings.

Discussion

Judges in this study appeared highly satisfied in their work, and conveyed a strong feeling that accelerated pay would improve their satisfaction still further.

Judges possess alternatives. Many could make much more money as lawyers, as arbitrators, or as “private judges.” The last effective salary increase for Massachusetts judges was in 2000. In 2006, Massachusetts judges received their

first salary raise since 2000, from \$117,000 to \$129,000. By way of comparison, the mean salary for comparable positions in other United States jurisdictions at that time was \$119,630, with a range of \$94,093 to \$165,200 (National Center for State Courts, 2006).

Respondents indicated that, in addition to financial compensation, an improvement in professional support staff would increase job satisfaction. Other potential improvements, such as those in secretarial support staff or security staff, were described as having less of a potentially positive effect. Obtaining networked personal computers was also not seen as very important, and geographic location was found to be the least important factor for job satisfaction.

The positive relationship between caseloads and satisfaction, while only significant at the level of a statistical trend, suggested that those with high caseloads were at least as satisfied as those with lower caseloads—an intriguing and as yet unexplained result. One additional aspect of working conditions, the social class background of courtroom participants, was positively correlated with overall satisfaction.

Despite the factors that might have an impact on job satisfaction, the high level of satisfaction reported suggests that most or all of them were not important enough to impact the judges' sense of their job satisfaction to any great negative extent. Perhaps this is because these were individuals who opted to remain working in the judicial system as opposed to seeking opportunities elsewhere.

One might think that working in the intense and adversarial setting of the court, in which judges generally receive little positive feedback or support regarding their decisions, would negatively impact job satisfaction. Within this environment, it may be that judges generally build their own supportive networks among colleagues or friends. It may also be worthy of note that the survey was completed in

2004, prior to a number of high-profile incidents of violence against judges and their families in other jurisdictions (AJC Staff, 2005; Heinzmann & Coen, 2005).

Do these results provide useful information for expert witnesses? It may be that highly satisfied judges are less defensive and more self-assured. Does the degree of a judge's satisfaction alter the fashion in which an expert witness should testify? This is an issue that may merit further investigation. Of course, present results are known to apply only to certain courts within the Massachusetts legal system. Future studies might usefully survey a variety of judges in other jurisdictions, enhancing our understanding of how the way in which courts are designed and staffed may affect judicial satisfaction nationwide.

References

- AJC Staff. (2005, March 12). Courthouse killings: Anatomy of a shooting. *Atlanta Journal-Constitution*, p. 4A.
- Andrich, D. (1988). *Rasch models for measurement*. Thousand Oaks, CA: Sage.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Commons, M. L., Miller, P. M., & Gutheil, T. G. (in press). A pilot Rasch scaling of forensic experts' perceptions of expert bias. *Journal of the American Academy of Psychiatry and the Law*.
- Dattilio, F. M., Commons, M. L., Adams, K. M., Gutheil, T. G., & Sadoff, R. L. (2006). Pilot Rasch scaling of lawyers' perceptions of expert bias. *Journal of the American Academy of Psychiatry and the Law*, 34, 482-491.
- Donkin, R. (2005, April 5). Why competent management means happy workers but poor leadership can destroy the underlying factors that underpin job satisfaction. *The Financial Times*, p.15.
- Eells, T. D., & Showalter, C. R. (1994). Work-related stress in American trial judges. *Bulletin of the American Academy of Psychiatry and the Law*, 22, 71-83.
- Eisenstat, R. A., & Felber, R. D. (1984). Toward a differentiated view of burnout: Personal and organizational mediators of job satisfaction and stress. *American Journal of Community Psychology*, 12, 411-430.

- Geller, A. (2005, February 28). U.S. workers more dissatisfied with jobs. *Associated Press*. Retrieved July 1, 2011, from <http://www.highbeam.com/doc/1P1-105839852.html>
- Graham, J., Ramirez, A. J., Field, S., & Richards, M. A. (2000). Job stress and satisfaction among clinical radiologists. *Clinical Radiology*, 55, 182-185.
- Heinzmann, D., & Coen, J. (2005, February 28). Federal judge's family killed: Husband, mother found slain in basement. *Chicago Tribune*. Retrieved July 1, 2011, from <http://www.com/news/nationworld/chi-0503010123mar01,0,4913954.story>
- Lai, J. S., Cella, D., Chang, C. H., Bode, R., & Heinemann, A. A. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Quality of Life Research*, 12, 485-501.
- Linacre, M. (2004). Winsteps [software]. Chicago, IL: Author.
- Massachusetts Exec. Order No. 420. (2006, February 21). Retrieved July 1, 2011, from <http://www.lawlib.state.ma.us/source/mass/eo/eotext/EO470.pdf>
- Middleton, M. (1981). Judges told how to deal with stress. *ABA Journal*, 67, 1100-1102.
- National Center for State Courts. (2006). *Survey of judicial salaries*. Washington, DC: Author.
- Rogers, J. M., Freeman, S. J., & LeSage, P. (1991). The occupational stress of judges. *Canadian Journal of Psychiatry*, 36, 317-322.
- Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. New York: Cambridge University Press.
- Smith, E. V. (2002). Understanding Rasch measurement: Detecting and evaluating the impact of multidimensional using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3, 205-230.
- Wright, B. D., Linacre, J. M., Gustafson J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: Mesa Press.
- Zautra, A. J., Eblen, C., & Reynolds, K. D. (1986). Job stress and task interest: Two factors in work life quality. *American Journal of Community Psychology*, 14, 377-393.

Appendix 1

The properties of a Rasch scale are important in understanding the relative contribution of various possible factors on Judges' perceived satisfaction. As a result, a change of the degree of contribution to satisfaction of the Rasch scores of 1, is the same going from -2 to -1 as going from 0 to +1. Doubling on the Rasch scale means the same change in degree of contribution to satisfaction anywhere along its linear axis. For example, a perceived satisfaction with a value of 2.3 is half as strong as a perceived satisfaction of 4.6.

After analyzing data with a Rasch model, a number of questions can be answered. First, where on the scale does each independent variable fall (e.g., in this case, how strongly did each item contribute to the satisfaction of the judges surveyed)? Second, what is the range of scaled values between all items for all participants? For example, consider a difference of 1 unit between two scores. For a small range of scaled perceived satisfaction scores, this would be a big difference, whereas for a large range this would be a small difference. Third, for each participant, what is the scaled value of improvement of satisfaction when measured on the same scale? If the participants' scores are very similar, then they would tend to agree about what items are associated with satisfaction.

The main limitation with Rasch Analysis is when items do not fit a unidimensional scale. To test this, item fit is always evaluated using goodness-of-fit statistics, reported as infit mean-squares (MNSQ). MNSQ is the ratio of the observed to the predicted variance for an item, and informs how well each item functions within the scale (Smith, 2002). Items with MNSQ values higher than 2 misfit the scale (Wright, Linacre, Gustafson, & Martin-Lof, 1994), indicates that the item either taps a different dimension or that it differs from other items in its ability to discriminate people (Lai, Cella, Chang, Bode, & Heinemann, 2003).