Training to Improve Judgmental Expertise by Using Decompositions of Judgment Accuracy Measures

Eric R. Stone

Wake Forest University

Overarching Goal

- How can we improve judgment accuracy?

- Two types of judgments:

1. Judgments of discrete events, typically in probabilistic form

   What is the probability the Cardinals will win the World Series?

2. Quantitative judgments of continuous quantities

   How many users of Facebook will there be at the end of 2011?

Overarching Approach

- Rather than develop training techniques designed to increase judgment accuracy generally, our approach targets specific aspects (components) of judgment accuracy.
- Each of these components are related to different skills, which are typically relatively unrelated to each other.
- By focusing intervention efforts on these specific skills, we can train each of the elements underlying overall judgment accuracy, leading to maximal improvement.

Today's Plan

1. **Discrete events**
   - Accuracy measures, including the Brier score ($\overline{PS}$) and its decomposition into the components of calibration and discrimination (Murphy, 1973; Yates, 1982)
   - Training of the component measures (Stone & Opel, 2000)

2. **Continuous events**
   - Accuracy measures, as seen in Extended-MSE analysis (Lee & Yates, 1992)
   - Training of the component measures (Youmans & Stone, 2005)

3. **ACES Project**
   - (Preliminary) instantiation of these ideas in an applied forecasting situation

Judgments of Discrete Events

- Overall accuracy – Brier Score (Mean Probability Score; $(\overline{PS})$)
- $\overline{PS} = \sum (f - d)^2 / n$

  where f = probability judgment

  d = outcome (0 if event does not happen; 1 if it does)

- Example Problem: What is the probability that the home team (e.g., Rangers) will win?
- f = judged probability of Rangers winning (0 to 1)
- d = 1 if Rangers win; 0 if Rangers lose

Judgments of Discrete Events

Mean Probability Score

- $\overline{PS} = \sum (f - d)^2 / n$
- Make judgments of the same type repeatedly

Game 1 -- p (HT wins) = .90; home team does win

Game 2 -- p (HT wins) = .60; home team does <u>not</u> win

Game 3 -- p (HT wins) = .20, home team does <u>not</u> win

$$\overline{PS} = \left[(.9-1)^2 + (.6-0)^2 - (.2-0)^2\right]$$

$$= (.01 + .36 + .04) / 3 = .14$$

Judgments of Discrete Events

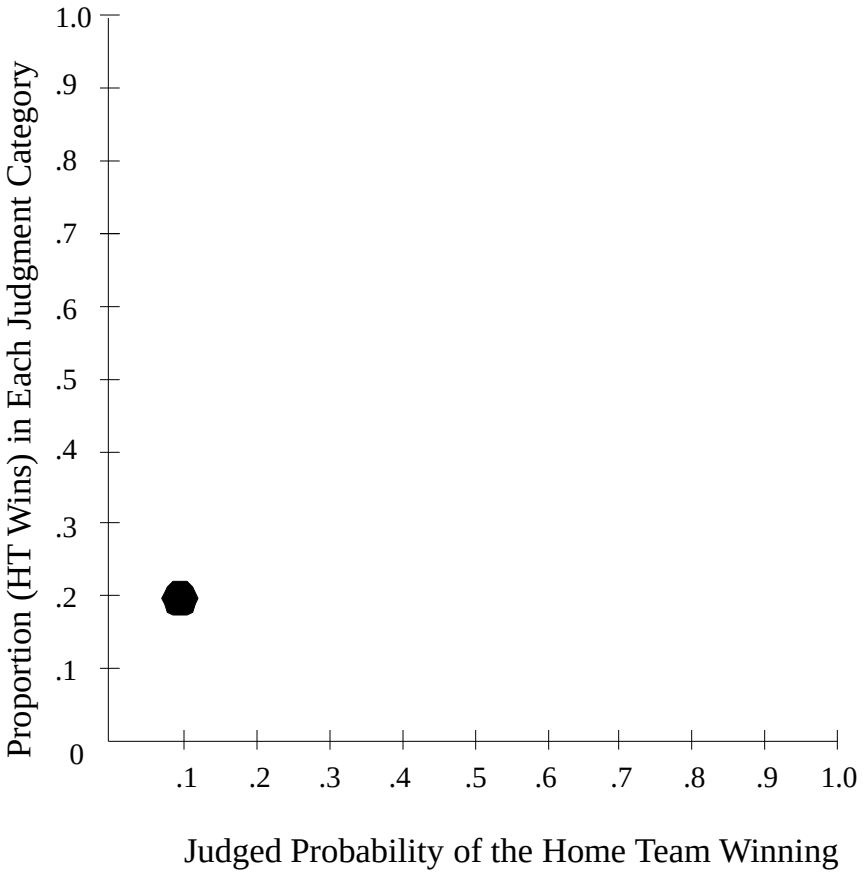
Decomposition of $\overline{PS}$

- $\overline{PS} = \sum (f - d)^2 / n$
- $\overline{PS}$ = fn (calibration, discrimination)

- discrimination (sometimes referred to as resolution) reflects "substantive expertise" – domain-specific knowledge in a specific area
- calibration reflects "calibration expertise" – the ability to assign probabilities that match the percentage of times that the target event actually occurs

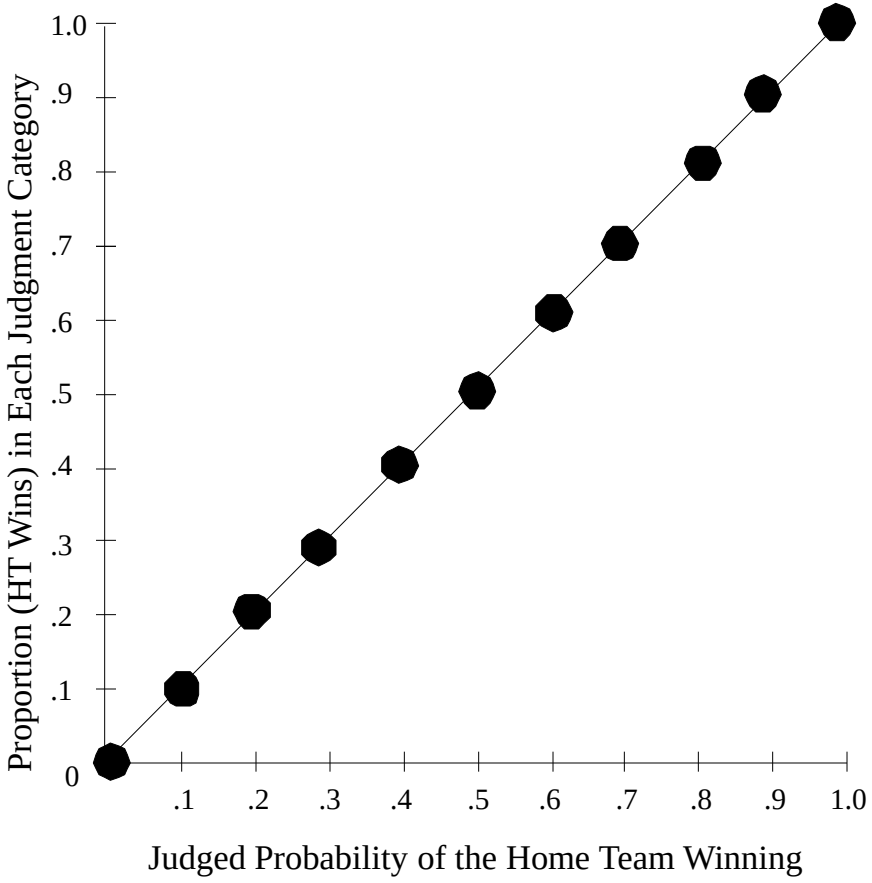
Judgments of Discrete Events

Calibration



Judgments of Discrete Events

Calibration



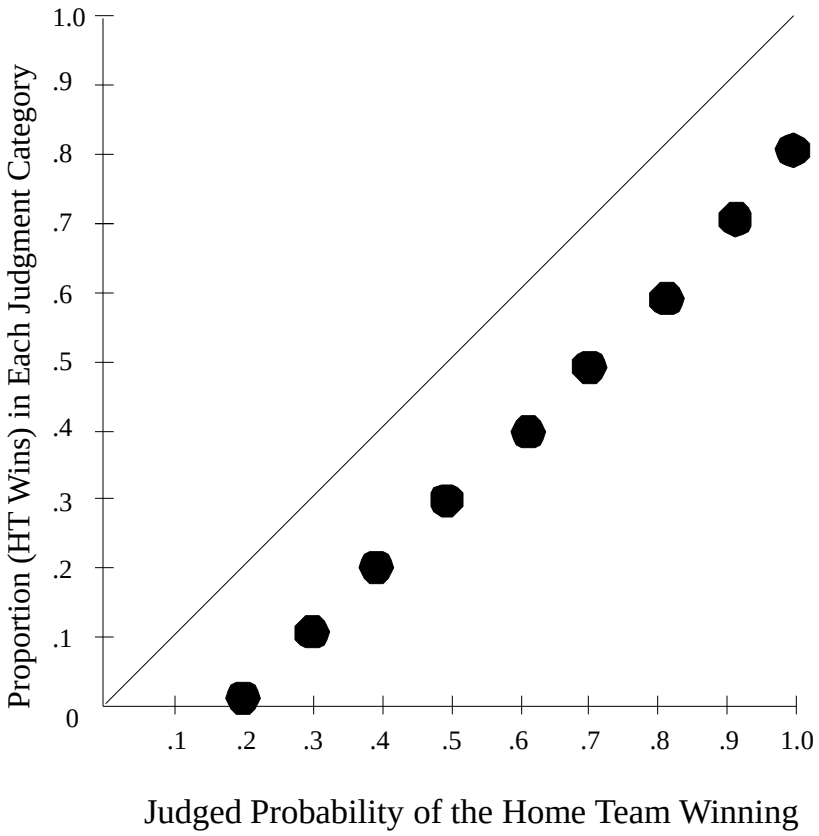
Judgments of Discrete Events

Calibration

- Types of poor calibration
  1) Over (under) estimation
  2) Over (under) confidence

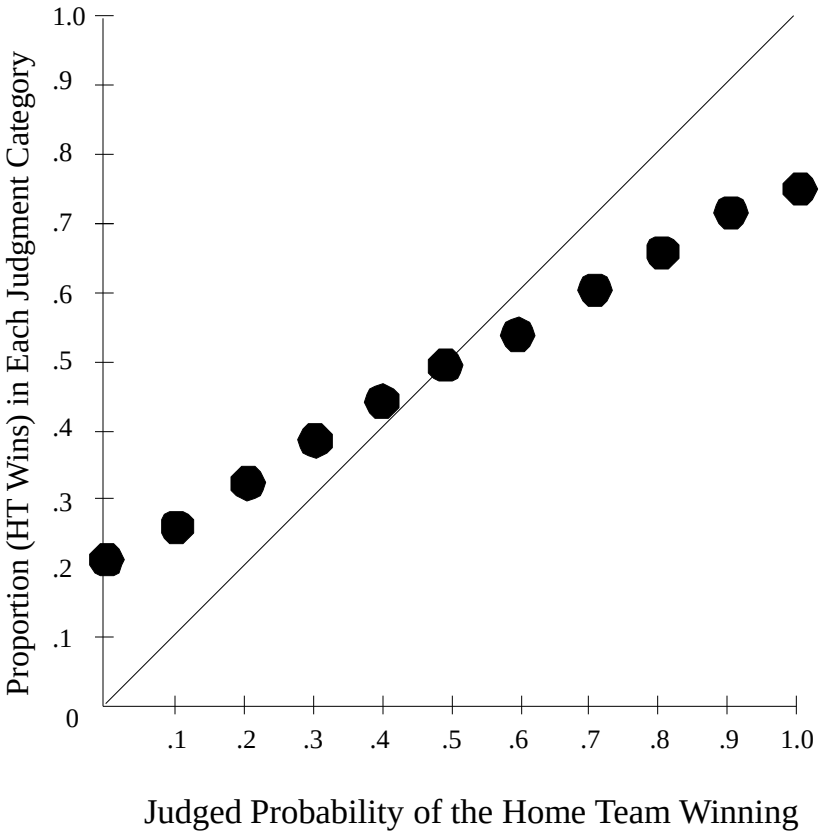
Judgments of Discrete Events

Calibration



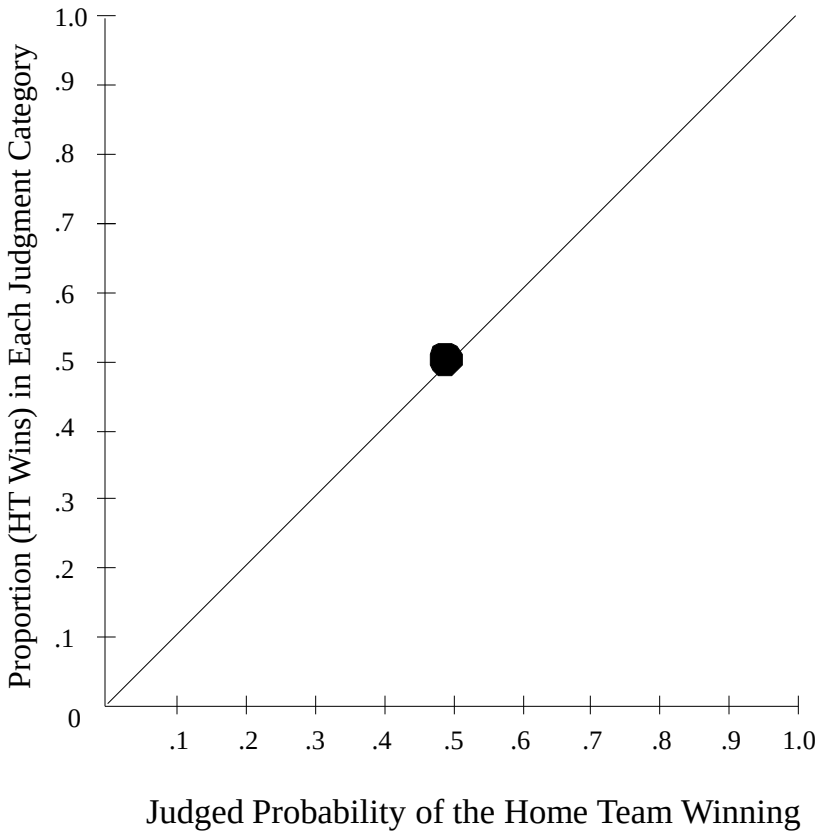
Judgments of Discrete Events

Calibration



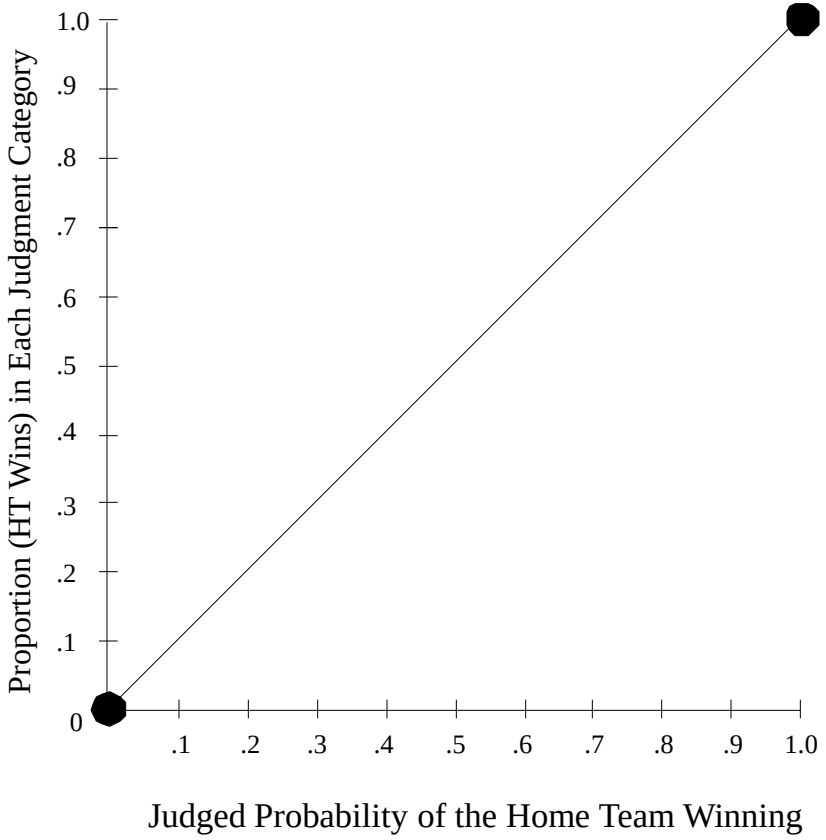
Judgments of Discrete Events

Discrimination



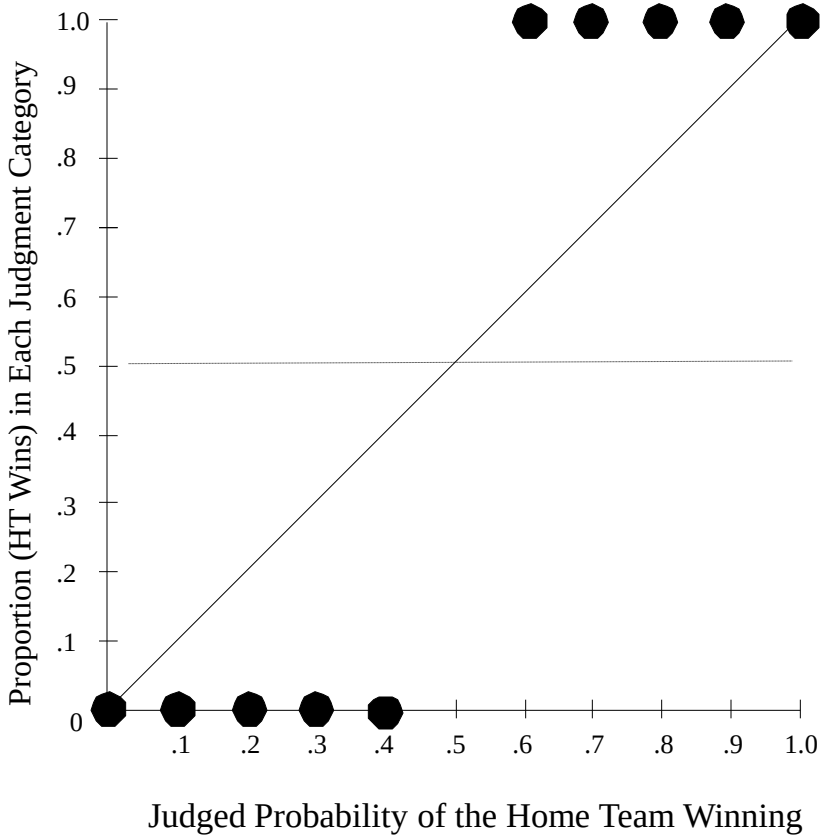
Judgments of Discrete Events

Discrimination



Judgments of Discrete Events

Discrimination



Judgments of Discrete Events

Summary of Decomposition Analysis

- Judgment expertise reflects an ability to make well-calibrated and well-discriminated probability judgments

Next Questions

- What judgment skills underlie good calibration and good discrimination?

- How can we train these skills to lead to an improvement in calibration and discrimination, and ultimately produce the largest possible improvement in $\overline{PS}$? (e.g., Yates, 1982)

Judgments of Discrete Events

Judgment Skills

Underlying good calibration

- translation of a "feeling of confidence" into a probability judgment (Ferrell & McGoey, 1980; Suantak, Bolger, & Ferrell, 1996)
- "the forecaster's ability to assign the 'right' labels to his or her forecasts" (Yates, 1982)
- we refer to this ability as "calibration expertise" (Stone & Hoffman, 1999; Stone & Opel, 2000)

Underlying good discrimination

- "the ability … to discriminate individual occasions on which the event of interest will and will not take place" (Yates, 1982)
- requires substantive knowledge about the events of interest
- we refer to this ability as "substantive expertise" (Stone & Hoffman, 1999; Stone & Opel, 2000)

Judgments of Discrete Events

Judgment Training: Calibration

- Many types of poor calibration (e.g., overconfidence) are resistant to training techniques (e.g., Sieck & Arkes, 2005).
- In particular, providing general advice seems to have little effect, presumably because people dismiss this advice as not relevant to them.
- The approach that seems to be most fruitful is to provide *performance feedback* about past judgment sessions (e.g., Lichtenstein & Fischhoff, 1980). This performance feedback entails providing more than outcome feedback; typically it entails providing calibration graphs of one's performance.
- The approach is particularly useful in reducing judgments that are overly extreme (Lichtenstein, Fischhoff, & Phillips, 1982).
- To maximize the effectiveness of this approach, we both present people with their calibration graphs and provide assistance in the interpretation of it (Stone & Opel, 2000; Stone, Rittmayer, & Parker, 2004).

Judgments of Discrete Events

Judgment Training: Discrimination

- To improve discrimination, one needs to provide substantive information related to the task at hand, or to train people to better use the information they have.
- Because it requires actual substantive information, discrimination is sometimes referred to as a more fundamental skill (e.g., Yates, 1982).
- Thus, discrimination training entails providing *environmental feedback*, i.e., information about the environment one is making predictions in / about.

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

- Goal: Do calibration expertise (e.g., calibration) and substantive expertise (e.g., discrimination) reflect two conceptually distinct skills that need to be trained separately?
- Basic Approach: Provide performance feedback to train calibration and environmental feedback to train discrimination, and examine the effect of each on the other measure.
- Performance feedback – Present participants with information related to their performance, in terms of a calibration diagram and accompanying individual feedback (e.g., you were overconfident…).
- Environmental feedback – Present participants with substantive information regarding the task at hand

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Design

- All participants responded twice, once during pretraining (baseline) and once during posttraining.
- Between pretraining and posttraining participants received either:

    1) No feedback

    2) Performance feedback

    3) Environmental feedback

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Materials

- Example question:

What period was this slide from?
a) Medieval (earlier period)
b) Renaissance (later period)

Probability from the later time period:
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Materials

- 100 hard slides (mean = 60% from the pretest).
- 100 easy slides (mean = 80% from the pretest).
- Participants saw 50 easy and 50 hard slides in both the pretraining and posttraining test phases.

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Procedure

• Participants arrived in groups of 15, and received a brief (10-15 minute) lecture on calibration and discrimination.

• All participants went through the pretraining phase, responding to 50 hard and 50 easy slides.

• Participants were split into groups of 5, and underwent the appropriate training technique.

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Procedure

• Performance feedback group -- Provided calibration diagrams, and individualized feedback.

• Environmental feedback group -- Given lecture on art history.

• No Feedback -- No intervention.

• Participants reconvened in the main room, and responded to another 50 hard and 50 easy slides.

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Results – Hard Slides: *Mean Probability Score*

Performance Feedback

- Scores decreased from .293 to .259 **

Environmental Feedback

- Scores decreased from .287 to .233 **

No Feedback

- Scores stayed the same, going from .286 to .273

***** indicates p < .01***

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Results – Hard Slides: *Calibration*

Performance Feedback

- Scores decreased from .095 to .067 **

Environmental Feedback

- Scores stayed the same, going from .094 to .090

No Feedback

- Scores stayed the same, going from .094 to .093

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Results – Hard Slides: *Overconfidence*

Performance Feedback

- Scores decreased from .188 to .106 **

Environmental Feedback

- Scores stayed the same, going from .176 to .184

No Feedback

- Scores stayed the same, going from .188 to .168

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Results – Hard Slides: *Discrimination*

Performance Feedback

- Scores stayed the same, going from .048 to .039

Environmental Feedback

- Scores increased from .054 to .092 **

No Feedback

- Scores stayed the same, going from .055 to .056

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Results – Hard Slides: *Percent Correct*

Performance Feedback

- Scores stayed the same, going from 56.2% to 58.3% correct

Environmental Feedback

- Scores increased from 57.0% to 71.4% correct **

No Feedback

- Scores stayed the same, going from 58.2% to 60.4% correct

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Results – Easy Slides

The results with the easy slides were similar to those with the hard slides.

In particular:

- Performance feedback reduced overconfidence (and improved calibration generally, although not significantly), but did not influence discrimination or percent correct.
- Environmental feedback improved discrimination and percent correct, but did not improve calibration or overconfidence.

Additionally, the provision of environmental feedback actually led to an increase in overconfidence. This result is in keeping with many studies (e.g., Oskamp, 1965) that show that providing information can increase *perception* of knowledge more than actual knowledge.

Judgments of Discrete Events

Judgment Training: Stone & Opel (2000)

Conclusions

- Calibration and discrimination appear to reflect two separate skills (which we call calibration and substantive expertise), which require separate training techniques.
- Further, there is some risk that training of one skill can actually cause decrements in the other skill.

Judgments of Continuous Events

- Overall accuracy – Mean Squared Error (MSE)
- $MSE = \sum (f - d)^2 / n$

  where f = quantity judgment

  d = quantitative outcome

- Example Problem: What will be the high temperature on Tuesday October 23rd?
- f = judged temperature
- d = actual temperature

Judgments of Continuous Events

Mean Squared Error

- $MSE = \sum (f - d)^2 / n$
- Make judgments of the same type repeatedly

Day 1 – predicted high = 71 degrees; actual high =75

Day 2 – predicted high = 68 degrees; actual high =74

Day 3 – predicted high = 67 degrees; actual high =66

$$MSE = \left[(71-75)^2 + (68-74)^2 - (67-66)^2 \right]$$

$$= (16 + 36 + 1) / 3 = 17.67$$

Judgments of Continuous Events

Decomposition of MSE

- $MSE = \sum (f - d)^2 / n$
- Extended-MSE analysis (Lee & Yates, 1992)

$$MSE = (\bar{Y}_s - \bar{Y}_e)^2 + (S_{Y_s} - S_{Y_e})^2 + 2(1 - r_a) S_{Y_s} S_{Y_e}$$

where $\bar{Y}_s$ = the average judged value

$\bar{Y}_e$ = the average criterion value

S_{Y_s} = the standard deviation of the judge's values

S_{Y_e} = the standard deviation of the criterion values

r_a = achievement

Judgments of Continuous Events

$E-MSE$ Analysis

- Combines the decomposition below with a lens model approach to decompose achievement.

$$MSE = (\bar{Y}_s - \bar{Y}_e)^2 + (S_{Y_s} - S_{Y_e})^2 + 2(1 - r_a)S_{Y_s}S_{Y_e}$$

where $\bar{Y}_s$ = the average judged value

$\bar{Y}_e$ = the average criterion value

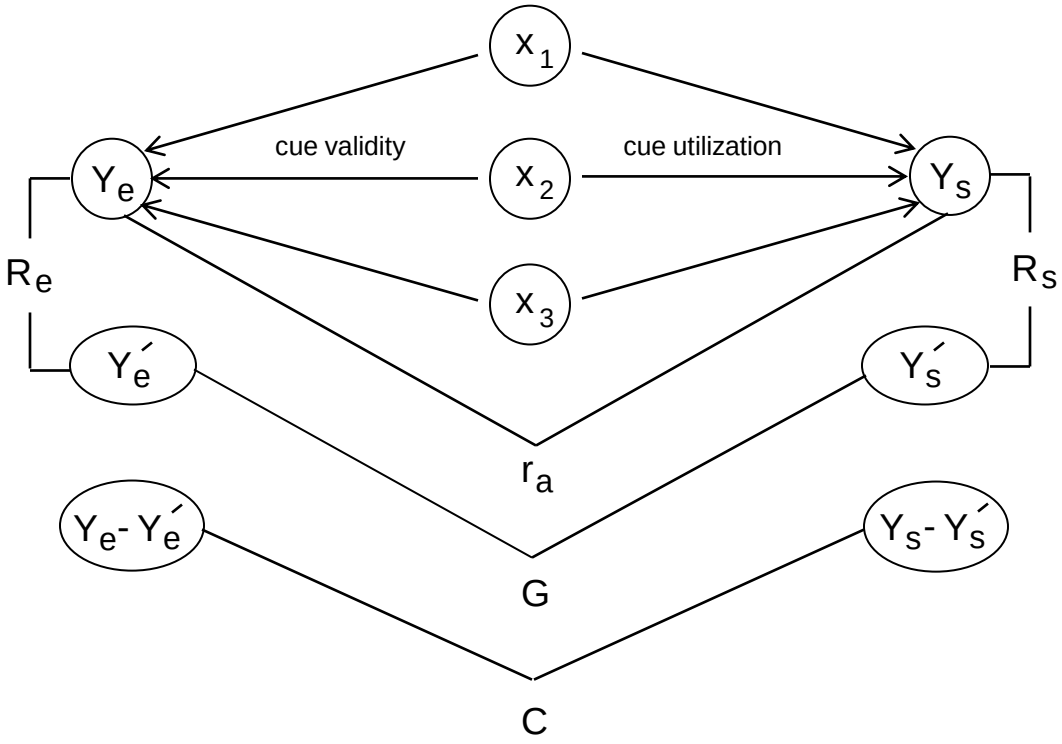
S_{Y_s} = the standard deviation of the judge's values

S_{Y_e} = the standard deviation of the criterion values

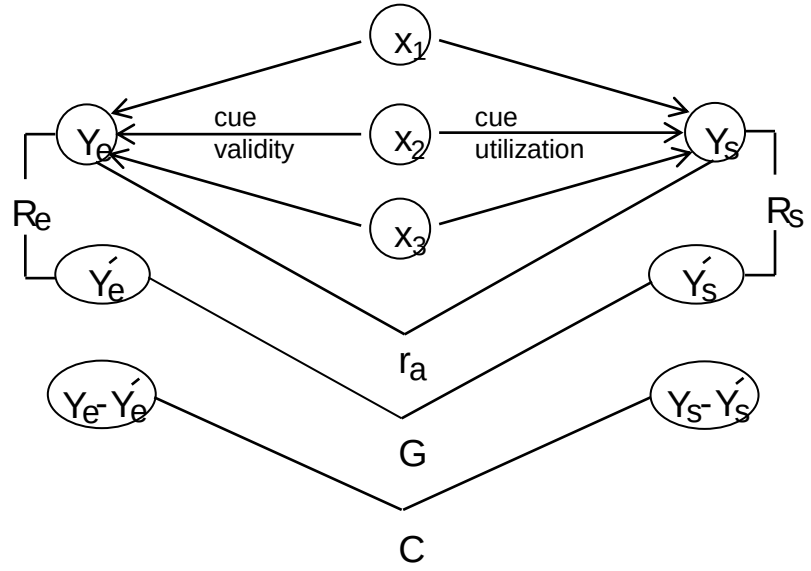
r_a = achievement

Judgments of Continuous Events

Lens Model Analysis



Judgments of Continuous Events



r_a = achievement, the correlation between the judgments and the criterion values

G = linear knowledge, the correlation between the best linear prediction of the judge and the best linear prediction of the criterion

C = non-linear knowledge, the correlation between the residual from the best linear prediction of the judge and the best linear prediction of the criterion

R_s = linear consistency, the correlation between the judgments and best linear prediction of the judgments

R_e = environmental predictability, the correlation between the criterion and best linear prediction of the criterion

Judgments of Continuous Events

$E-MSE$ Analysis

- From a lens model analysis:

$$r_a = GR_eR_s + C\left(1-R_e^2\right)^{1/2}\left(1-R_s^2\right)^{1/2}$$

- Recall: $MSE = \left(\bar{Y}_s - \bar{Y}_e\right)^2 + \left(S_{Y_s} - S_{Y_e}\right)^2 + 2\left(1-r_a\right)S_{Y_s}S_{Y_e}$

- Thus:

$$MSE = \left(\bar{Y}_s - \bar{Y}_e\right)^2 + \left(S_{Y_s} - S_{Y_e}\right)^2 + 2\left(1-[GR_eR_s + C\left(1-R_e^2\right)^{1/2}\left(1-R_s^2\right)^{1/2}]\right)S_{Y_s}S_{Y_e}$$

Judgments of Continuous Events

$E-MSE$ Analysis

$$MSE = \bar{Y}_s - \bar{Y}_e^{\,2} + S_{Y_s} - S_{Y_e}^{\,2} + 2\left(1 - [CR_eR_s + C\left(1 - R_e^2\right)^{1/2}\left(1 - R_s^2\right)^{1/2}]\right) S_{Y_s} S_{Y_e}$$

- Controllable elements:

1) $\bar{Y}_s$ = average judged value

2) S_{Y_s} = the standard deviation of the judge's values

3) G = linear knowledge

4)  = linear consistency

C?

Judgments of Continuous Events

$E-MSE$ Analysis

$$MSE = \left(\bar{Y}_s - \bar{Y}_e\right)^2 + \left(S_{Y_s} - S_{Y_e}\right)^2 + 2\left(1 - \left[C R_e R_s + C\left(1 - R_e^2\right)^{1/2}\left(1 - R_s^2\right)^{1/2}\right]\right) S_{Y_s} S_{Y_e}$$

1) $\bar{Y}_s$ = average judged value

- The goal here is to match the average judged value to the average criterion value to the extent possible
- Thus, helping the judge to be aware of the average criterion value should reduce bias

Judgments of Continuous Events

$E-MSE$ Analysis

$$MSE = \bar{Y}_s - \bar{Y}_e^{\,2} + S_{Y_s} - S_{Y_e}^{\,2} + 2\left(1 - \left[C R_e R_s + C\left(1 - R_e^2\right)^{1/2}\left(1 - R_s^2\right)^{1/2}\right]\right) S_{Y_s} S_{Y_e}$$

2) S_{Y_s} = the standard deviation of the judge's values

- The standard deviation influences accuracy in two ways:

  -- in comparison to the criterion standard deviation

  -- in an absolute sense

- In combination, the standard deviation of the judge's values should generally be smaller than the standard deviation of the criterion values, but how much smaller depends on achievement. Specifically, MSE is maximized when:

  $S_{Y_s} = r_a \times S_{Y_e}$ (Gigone & Hastie, 1997)

Judgments of Continuous Events

$E-MSE$ Analysis

$$MSE = \left(\bar{Y}_s - \bar{Y}_e\right)^2 + \left(S_{Y_s} - S_{Y_e}\right)^2 + 2\left(1 - \left[GR_eR_s + G\left(1 - R_e^2\right)^{1/2}\left(1 - R_s^2\right)^{1/2}\right]\right)S_{Y_s}S_{Y_e}$$

3) G = linear knowledge

- To have good knowledge, one needs to have diagnostic information in the relevant domain.
- Thus, the same factors that influence discrimination (environmental feedback) should influence knowledge.

Judgments of Continuous Events

$E-MSE$ Analysis

$$MSE = (\bar{Y}_s - \bar{Y}_e)^2 + (S_{Y_s} - S_{Y_e})^2 + 2(1-[(C(R_e R_s) + C(1-R_e^2)^{1/2}(1-R_s^2)^{1/2}])S_{Y_s}S_{Y_e}$$

4)  = linear consistency

- Linear consistency can be low for at least two reasons:
  -- applying a judgment policy inconsistently (i.e., having random error in one's judgments)
  -- including irrelevant cues in one's judgments
- Thus, any interventions that target either of the above should improve consistency.

Judgments of Continuous Events

Judgment Training: Youmans & Stone (2005)

- Goal: Investigate the effects of task information and cognitive information feedback on each of the controllable measures in E-MSE analysis

- Basic Approach: We provided participants with a prediction task, and then gave them feedback in terms of either task information (TI), cognitive information (CI), or both. Participants then made another set of judgments.

- Task information (TI) – Information about the policies that should be followed to make good judgments (a type of environmental feedback).

- Cognitive information (CI) – Information about an individual's judgment policy.

Judgments of Continuous Events

Judgment Training: Youmans & Stone (2005)

Design

- All participants responded twice, once during pretraining (baseline) and once during posttraining.
- Between pretraining and posttraining participants received either:

    1) No feedback

    2) TI feedback

    3) CI feedback

    4) TI + CI feedback

Judgments of Continuous Events

Judgment Training: Youmans & Stone (2005)

The Task

- Participants judged the income level of respondents to the General Social Survey on a scale from 1 (under $5,000) to 12 ($75,000 or over).

- Participants were told the average income level of all participants.

- To make their predictions, participants were given information about three cues:

    1) education level (diagnostic cue)

    2) time spent socializing with relatives (non-diagnostic cue)

    3) attitude toward easy listening music (non-diagnostic cue)

Judgments of Continuous Events

Judgment Training: Youmans & Stone (2005)

Feedback

- TI Group: Participants were given the non-standardized regression weights between the cues and the criterion. Specifically, these were .52 for education, .007 for time with relatives, and -.08 for easy music listening.

- CI Group: Participants were given the non-standardized regression weights corresponding to their predictions, i.e., their linear judgment policy

- TI + CI Group: Participants were given both pieces of information above (i.e., the actual regression weights and their judgment policy)

Judgments of Continuous Events

Judgment Training: Youmans & Stone (2005)

Results: *Standard deviation of participants' judgments*

- At pretest, the average standard deviation was 2.42 (optimal level = 1.09)

Change in S_{Y_s}

	No TI	TI
No CI	.058	-.438 **
CI	.014	-.775 **

Judgments of Continuous Events

Judgment Training: Youmans & Stone (2005)

Results: *Linear knowledge*

- At pretest, linear knowledge was .98.

Change in G

	No TI	TI
No CI	.004	.006
CI	.000	.019 *

Judgments of Continuous Events

Judgment Training: Youmans & Stone (2005)

Results: *Linear consistency*

- At pretest, linear consistency was .87.

Change in R_s

	No TI	TI
No CI	.031**	.100**
CI	.032**	.097**

Judgments of Continuous Events

Judgment Training: Youmans & Stone (2005)

Conclusions

- CI alone did not produce improvements on any of the measures.
- TI alone produced improvements in the standard deviation of the judgments and in linear consistency.
- CI when added to TI led to improvement in the standard deviation of the judgments and in knowledge vs. TI alone.
- These results thus provide a starting point for learning how to train the various components in E-MSE analysis.

Judgments of Continuous Events

Judgment Training: Youmans & Stone (2005)

Limitations

- Although TI (and environmental feedback more generally) are helpful for training various components, constructing this feedback is not straightforward in many applied situations.

- Thus, the development of other techniques for training specific components would be very beneficial.

ACES

(PI: Dirk Warnaar, Applied Research Associates)

- Response to an IARPA announcement designed to improve intelligence forecasting.
- Provides an important applied situation for testing many of the ideas described previously.
- In particular, our approach is to provide training related to the various components. Our main focus so far has been on discrete events but we will extend training to continuous events in the future.
- All training techniques have to be part of the ACES architecture and thus be automated. This provides considerable challenges in adapting what we have done previously.

Automated Calibration Training

Eric Stone (WFU), Jason Luu (WFU), & Ben Simpkins (ARA)

Purpose

- Develop an automated procedure to provide the feedback given in Stone and Opel (2000)
- Doing so would allow us to utilize this feedback in the ACES project

Additional Goals

- Determine if this feedback is successful in a forecasting task
- Examine the impact on aggregated forecasts as well as on individual forecasts

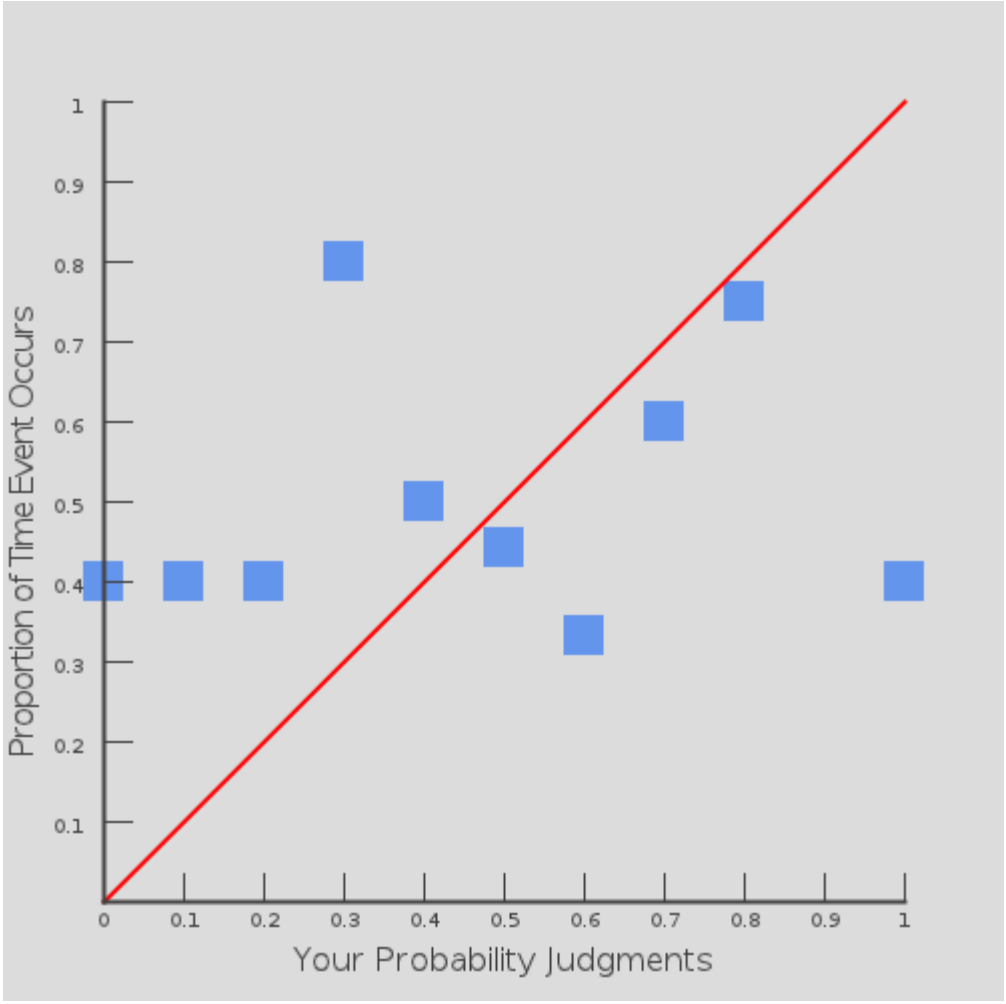
Task

- Predict the winner of baseball games
- What is the probability of the home team winning?

Experimental Design

- Independent variable: Provision of calibration training (yes vs. no)
- Dependent variables: calibration, overconfidence (both individual and aggregated)

Automated Calibration Training



Automated Calibration Training

Example of individualized feedback – accompanying text

[Note: This text was provided below the calibration diagram, so both were simultaneously visible]

Your responses indicate substantial overconfidence: you were much too certain about which team would win. This overconfidence can be seen in results that are typically below the line for judgments greater than 50%, and/or above the line for judgements less than 50%. To increase the accuracy of your predictions, we recommend that you make less confident predictions.

In particular, your responses were too extreme, especially when you stated you were certain the home team would win or lose. Specifically, as you can see on your calibration diagram, when you said you were 100% sure the home team would win, the home team won only 40% of the time. The same observation holds when you said you were very sure the home team would *not* win. When you said there was no chance the home team would win, the home team actually won 40% of the time. This was also true when you said there was a 90% or 10% chance of the home team winning, but the overconfidence was most serious when you stated you were certain the home team would win or lose. To improve your predictions, we recommend that you be more reluctant to make these types of highly confident predictions.

Future Goals

Training Discrimination

- We are presently testing information sharing procedures, which should increase substantive knowledge and hence discrimination, at least at an individual level.
- We have discussed additional procedures for improving discrimination, including 1) a database of relevant questions and answers for an item, and 2) providing contributors the option of asking questions about an item that can be answered by other contributors.

Improving Judgment of Continuous Quantities

- Many of the procedures designed to improve discrimination should also increase linear knowledge.
- At the same time that we are trying to increase access to relevant information, we also are trying to implement procedures for decreasing use of irrelevant information, which should improve linear consistency in particular.
- Due to the use of aggregation procedures in this project, however, linear consistency may be of less relevance than other components.

Thanks!