Reviving Lord-McNemar True Gain Score in the Modern World

John Denbleyker
Houghton Mifflin Harcourt

Ye Ma
University of Iowa

Ideas in Testing Seminar

Nov 2, 2018

Outline of Presentation

- Why? Lord (1962)
- CTT
- Estimate of True Score Gain
- Kelley formula
- Example calculation
- Results/Comparisons
- Bayesian formulation

ELEMENTARY MODELS FOR MEASURING CHANGE

Frederic M. Lord

Paper presented at the Invitational Conference on Problems in Measuring Change, Madison, Wisconsin, April 30, 1962.

Educational Testing Service

Princeton, New Jersey

May 1962

Lord (1962, p.10)

Thoughts of observed gain score…

"It is a widespread fault in speech and in thought to substitute the observe value for the true value…"

"This sort of thinking can frequently be used without serious results because of the fact that in a group the rank order of the observed measurements often provide a reasonably good approximation to the rank order of observed scores. This approximation usually falls down, however, when we dealing with measurement of change. **It is for this reason that a consideration of errors of measurement is specifically important here**.

"For those who like a common-sense, operational approach to problems, for those who dislike the use of hypothetical constructs, problems in the measurement of change should provide a special challenge, since the usual common-sense notions can be shown to be inadequate here and **models involving unobservable variables seem to be of great practical use**."

Classical Test Theory (CTT) and Estimated True Score Gain

Data

- A longitudinal set of within-year scores (Fall to Spring) from grade 5 examinees taking the HMH Math Inventory 2.7.
- Student's scores from two testing occasions, denoted as opportunity X (Fall) and Y (Spring) were gathered and merged using a unique identifier variable.
- There were 18,484 matched scores from students testing at least twice within the school year
- These examinees were merged with the Winsteps calibration p-file to extract their empirical grade-level probit ability and transformed to the HMH Math Inventory 3.0 reporting scale via the constructed scoring tables to allow the reporting of Quantiles.
- Summary statistics were calculated on these data such as reliability, score averages, standard deviations, and correlation.

Classical Test Theory

$X_i = T_i + E_i$ *CTT model specifies X_i as an additive combination of two components that may vary over examinees, where T_i is a true score with mean μ_T and variance σ_T^2 and there is an error for an examinee with mean 0 and variance σ_E^2*

$\mu_x = \mu_T$ *the mean of the observed scores is equal to the mean of the true scores*

$\rho_{TE} = 0$ *errors are uncorrelated with true scores in the population*

$\rho_{(E_1,E_2)} = 0$ *correlation between the errors across people are independent*

$\rho_{(E_1,T_2)} = 0$ *error for a person is uncorrelated with the true score of any other person*

$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 - 2\sigma_{TE} = \sigma_T^2 + \sigma_E^2$ *observed score variance σ_X^2 is the summation of true score variance and error variance where the covariance between true score and errors is zero*

$\rho = \sigma_T^2 / \sigma_X^2 = \sigma_T^2 / \sigma_T^2 + \sigma_E^2 = \rho_{xT}^2$ *canonical definition of reliability is the squared correlation between observed and true scores*

Kelley's Regressed Estimate True Score

$$\hat{T}_i = \rho x_i + (1- \rho)\mu_x$$
$$= \mu_x + \rho(x_i - \mu_x)$$

True scores for an examinee may be estimated via the Kelley formula that dates back to at least 1923. The Kelley estimate can be framed as (1) the regression of the true scores on observed scores, and/or (2) an estimate for an examinee's true score by starting with the mean, and then moving away from the mean in the direction of their observed score in proportion to the score reliability.

Kelley's formula contradicts CTT in that we should use the examinee's observed score as the (unbiased) estimate of the true score ($\hat{T}_i = x_i$). See Brennan (2012) for further discussion of some fundamental inconsistencies with certain traditional assumptions and results in classical test theory vis-à-vis Kelley's formula.

Note that the Kelley formula will yield the observed score as the estimate of the true score when $\rho = 1$.

Kelley's Regressed Estimate True Score (1947, p.409)

"This is an interesting equation that it expresses the estimate of the true ability as the weighted sum of two separate estimates,-one based upon the individual's observed score, X_1 *, and the other based upon the mean of the group to which he belongs,* M_1. *If the test is highly reliable, much weight is given to the test score and little to the group mean, and vice versa."*

Lord/McNemar Estimated True Gain (CTT)

Lord (1962, p12) noted the estimated value of the true change $\hat{G}$ as the following:

$$\hat{G} = \bar{G} + b_{Gx.y}(x - \overline{x}) + b_{Gy.x}(y - \overline{y}) \quad \rightarrow \quad \widehat{T2 - T1}$$

Where $\bar{G} = \bar{y} - \bar{x}$

$$b_{Gx.y} = (1 - r_{yy'})\, r_{xy} s_y / s_x - r_{xx'} + r_{xy}^2 \,/\, 1 - r_{xy}^2$$

$$b_{Gy.x} = r_{yy'} - r_{xy}^2 - (1 - r_{xx'}) r_{xy} s_x / s_y - r_{xx'} \,/\, 1 - r_{xy}^2$$

As noted by O'Conner (1971), the Lord–McNemar approach can be expressed in more general terms. To estimate the true gain $G = Y - X$, with some estimator, $\hat{G} = ky + mx$, where k and m are weights, t is defined as the error of estimate, i.e., difference between the true value G, and our estimate, G:

$$t = G - \hat{G}$$

Lord/McNemar Estimated True Gain (CTT)

Components

Reliability estimate of score X $r_{xx'}$

Reliability estimate of score Y $r_{yy'}$

Standard deviation of scores for X s_x

Standard deviation of scores for Y s_y

Correlation between scores X and Y r_{xy}

Mean of scores X $\overline{x}$

Mean of scores Y $\overline{y}$

From these set of parameters partial regression coefficients $b_{Gx.y}, b_{Gy.x}$ are estimated and used as weights in a multiple regression equation to predict an examinee's true score gain

Brennan (2006) noted this is likely the best estimate of a gain score available.

Lord/McNemar Estimated True Gain (CTT)

The reliability of the estimated true gain score is

$$\rho_{\widehat{G}\widehat{G}'} = 1 - (b_{Gx.y}{}^2\ \sigma_x^2(1 - r_{xx'}) + b_{Gy.x}{}^2\ \sigma_y^2\ (1 - r_{yy'}))/\sigma_{\widehat{G}}^2.$$

where $\sigma_{\widehat{G}}^2 = (b_{Gx.y}{}^2\sigma_x^2 + b_{Gy.x}{}^2\sigma_y^2) + 2(b_{Gx.y}{}^2 b_{Gy.x}{}^2 r_{xy}\sigma_x^2\sigma_y^2)$

Lord/McNemar Estimated True Gain

McNemar (1962, 1969) stated an observed difference score is considered dependable if:

$$|D_i| = |Y_i| - |X_i| > 1.96\ \sigma_{E_D}$$

σ_{E_D} = *the standard error of the difference score*

Standard Error of Estimate

> Variance of true scores equals the variance of obtained scores minus the error of measurement variance, $\sigma^2_{G_t} = \sigma^2_G - \sigma^2_{e_g}$

$$\sigma^2_{G_t} = (\sigma^2_X + \sigma^2_Y - 2\,\rho_{xy}\sigma_x\sigma_y) - (\sigma^2_{e_x} + \sigma^2_{e_y})$$

$$\sigma_{G_{t.xy}} = \sigma_{G_t}\sqrt{1 - \rho^2_{G_{t.xy}}}$$

Example: Grade 5 MI Fall / Spring Scores

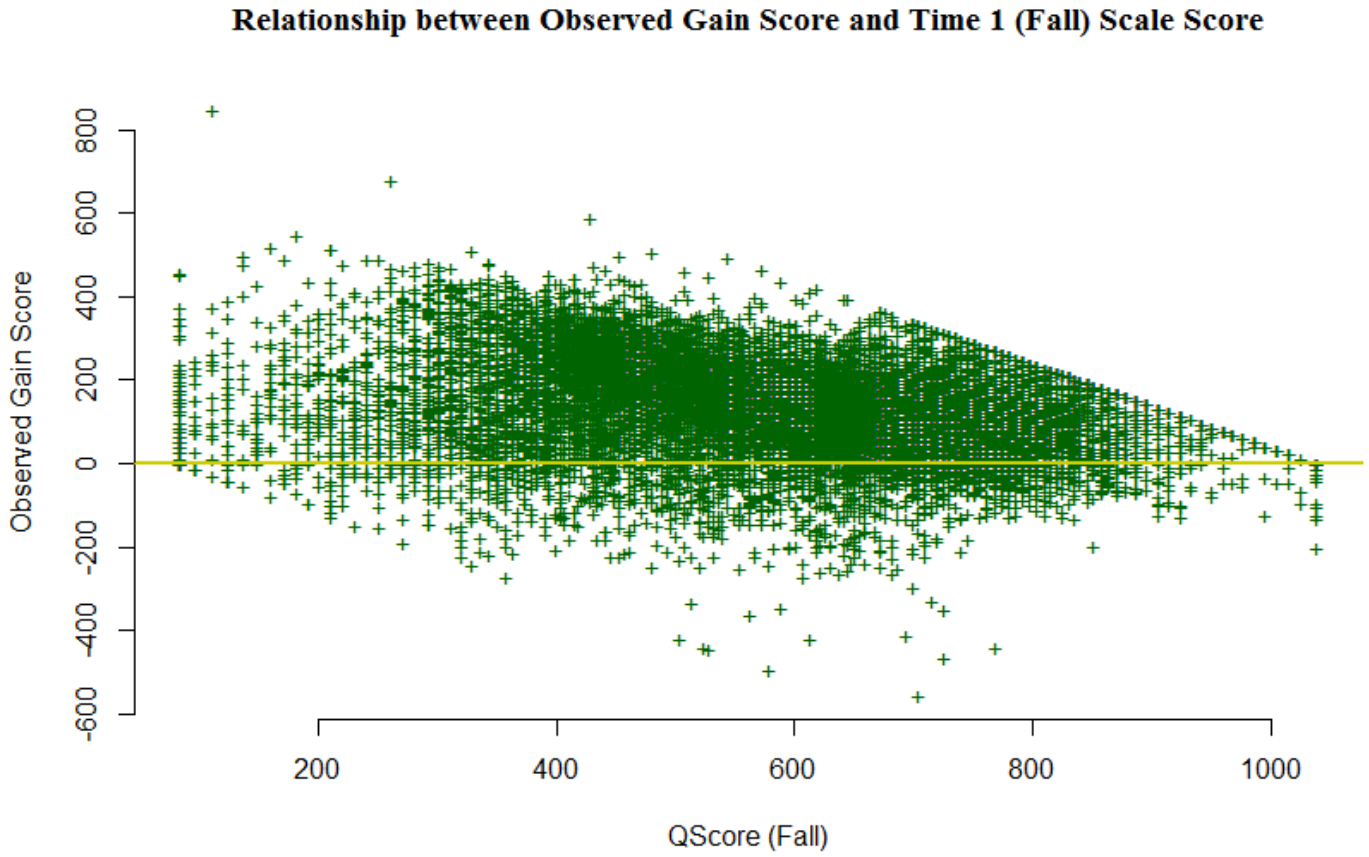
$r_{xx'} = 0.8567$ $r_{yy'} = 0.8567$ $s_x = 157.23$ $s_y = 157.176$

$r_{xy} = 0.7755$ $\bar{x} = 598.15$ $\bar{y} = 725.796$

```
LN <- function(x){
 W1 <- 1/(1-cor^2)*((((cor * sd2)/sd1)*(1-rel.2))-rel.1+(cor^2))
 W2 <- 1/(1-cor^2)*((((cor * sd1)/sd2)*(rel.1-1))+rel.2-(cor^2))
 Z <- (m2-m1) + W1 * (x$QSCORE-m1) + W2 * (x$QSCORE2-m2)
out <- data.frame(w1=W1, w2=W2, Z=Z)
 return(out)
}
```

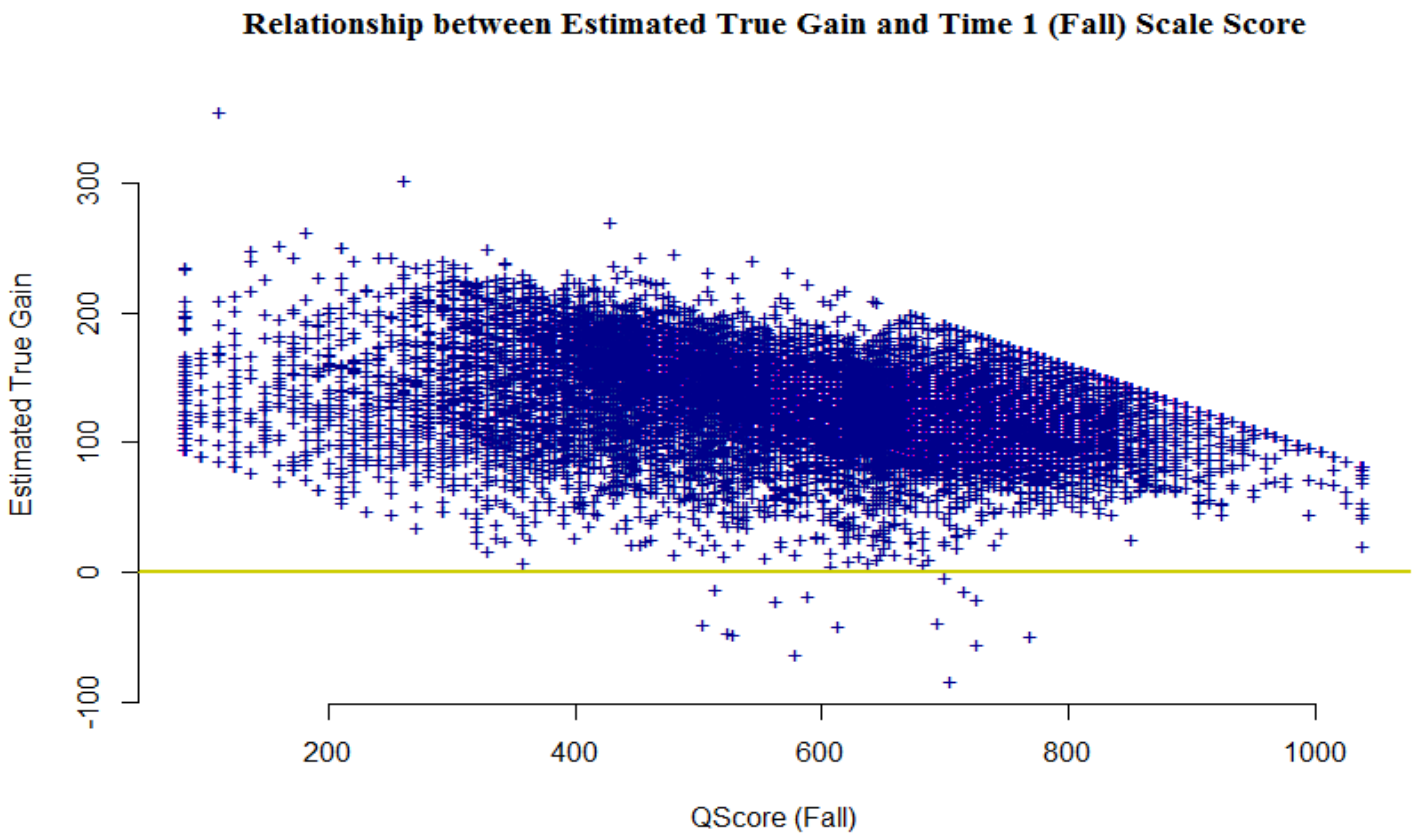
```
LN(jj3[1,])
        w1        w2       Z
1 -0.4391351 0.4322391 176.6696
```

Results



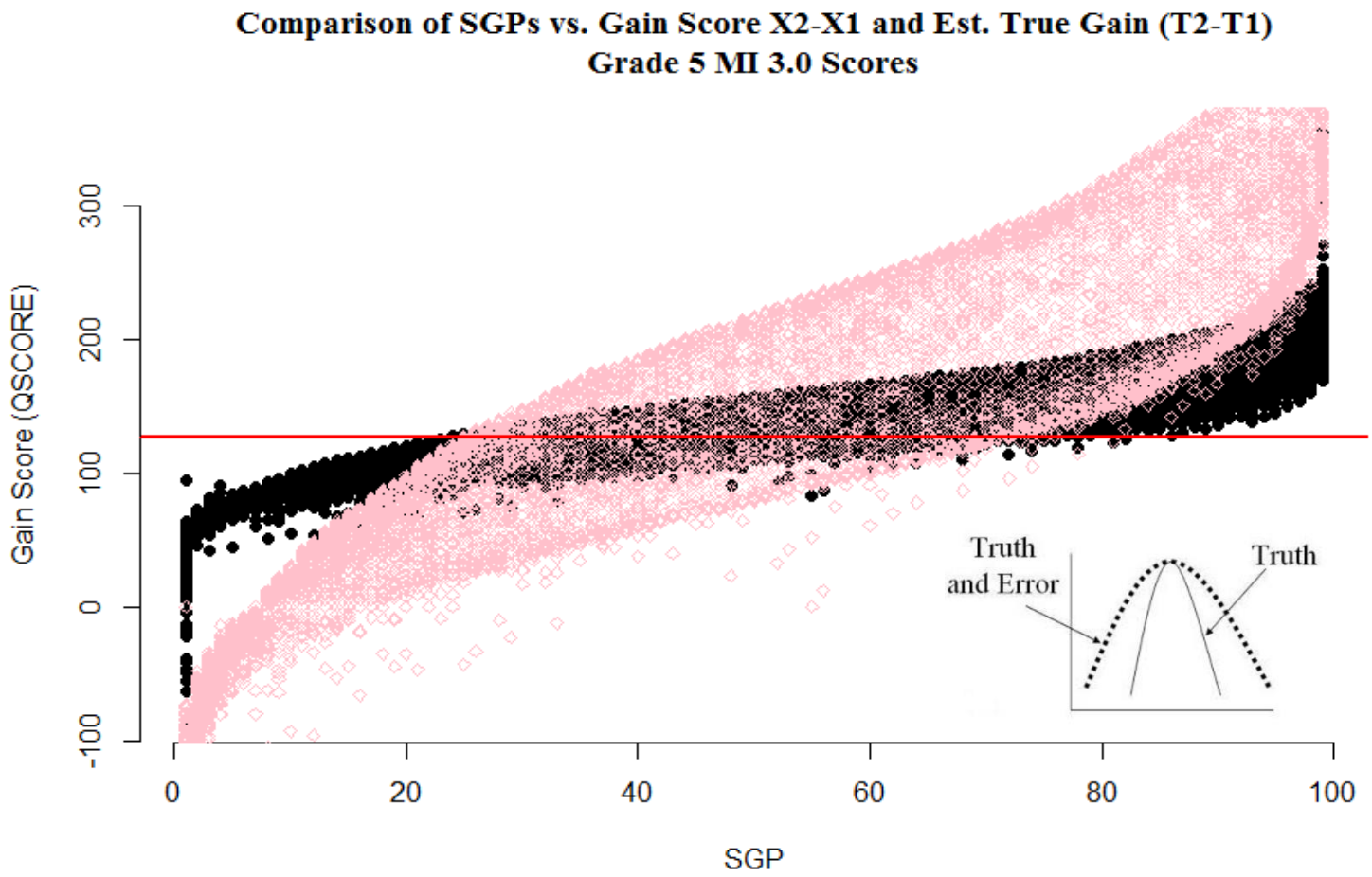
Houghton Mifflin Harcourt

Results



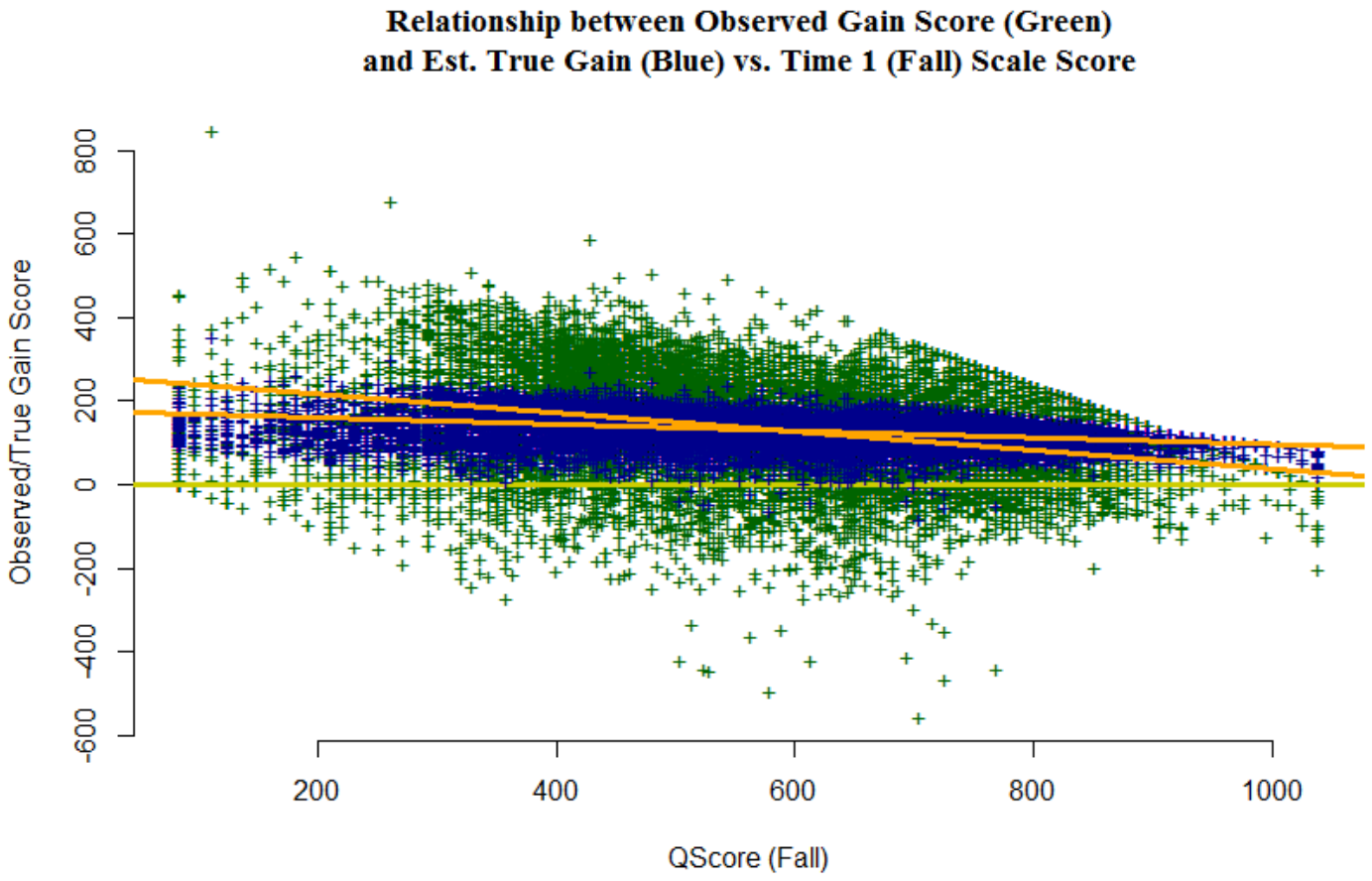
Houghton Mifflin Harcourt

Results



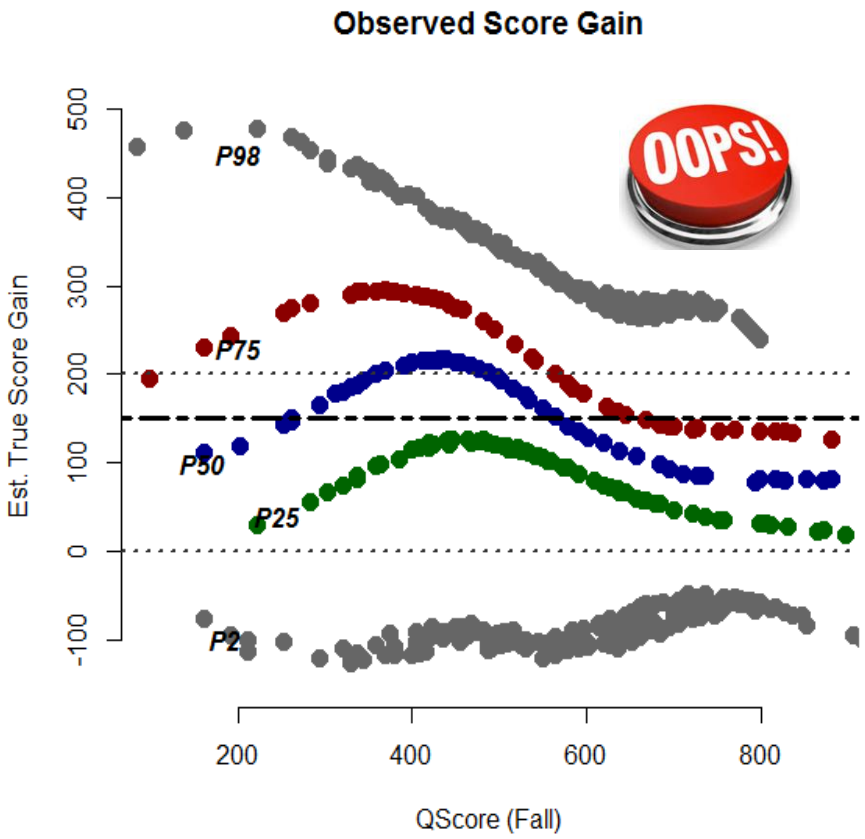
Houghton Mifflin Harcourt

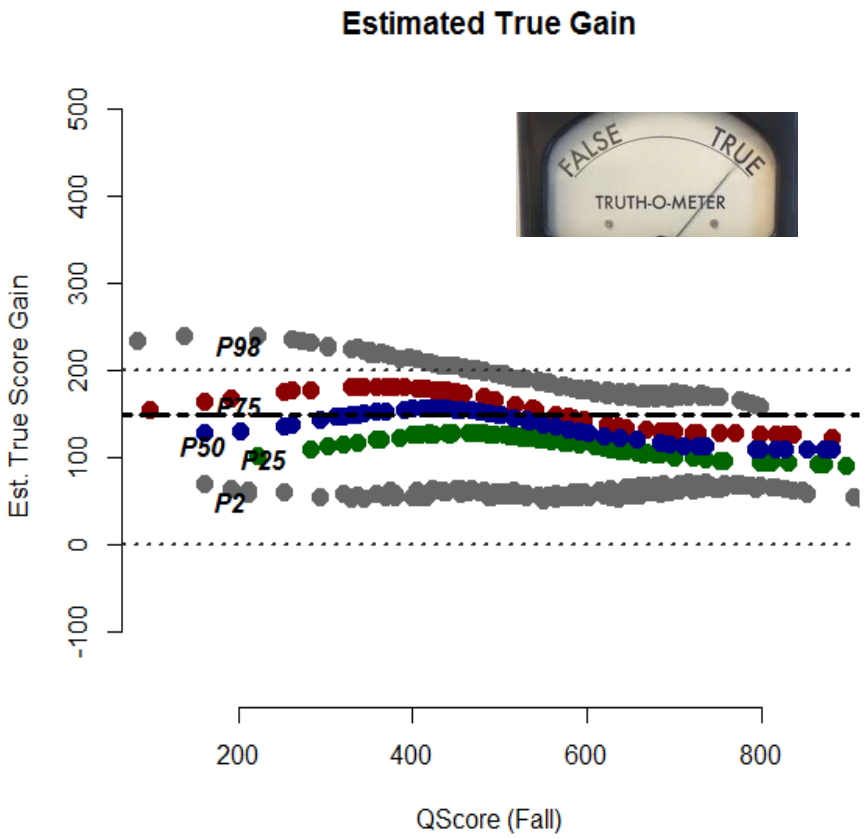
Results



Houghton Mifflin Harcourt

Comparison of Selected SGPs and Gains





Houghton Mifflin Harcourt

Extensions

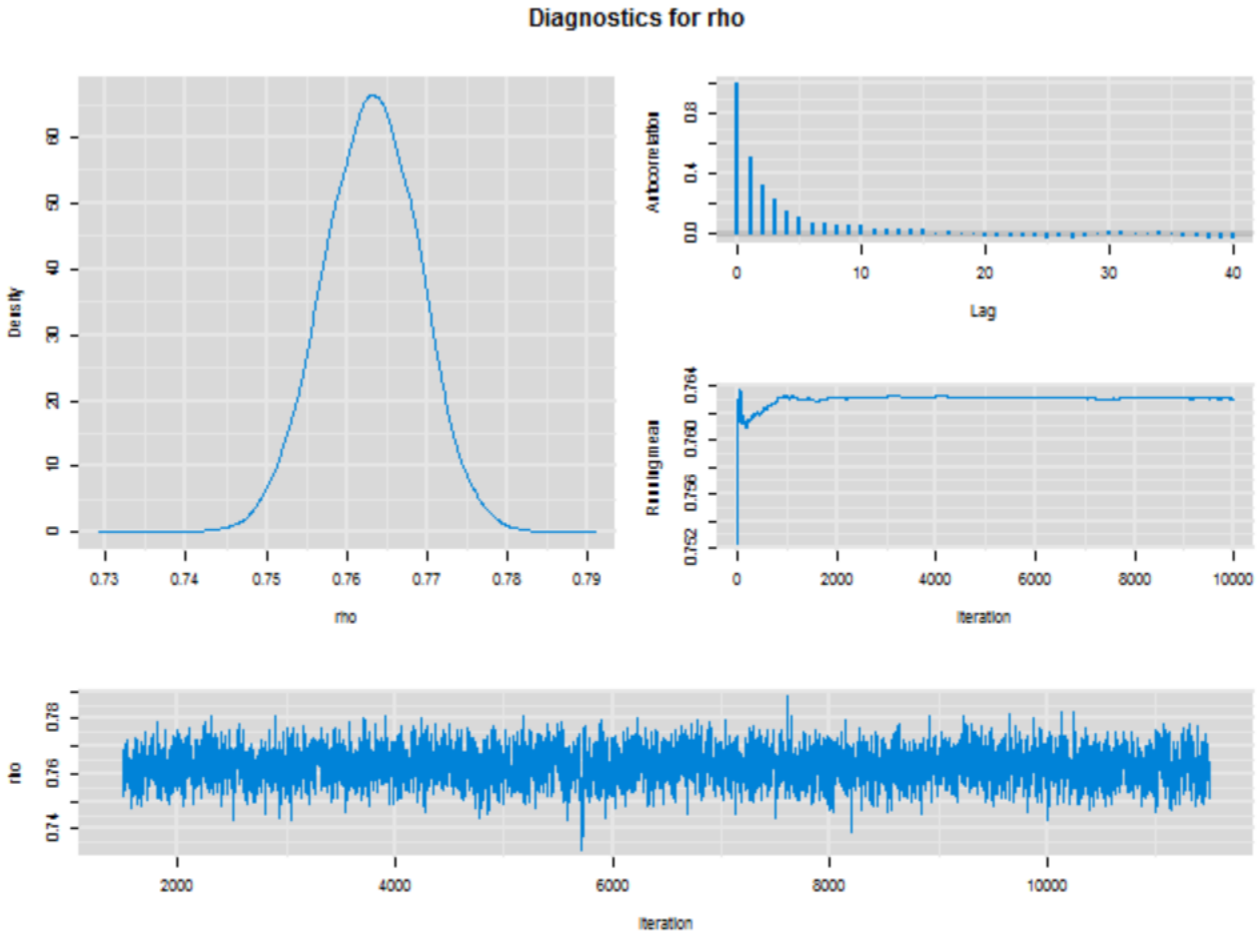
Lord-McNemar in Bayesian framework

- Instead of a point estimate of true score we can derive a posterior distribution of credible values for an estimated gain score from which to make inferences.

Lord-McNemar Bayesian Formulation in JAGS

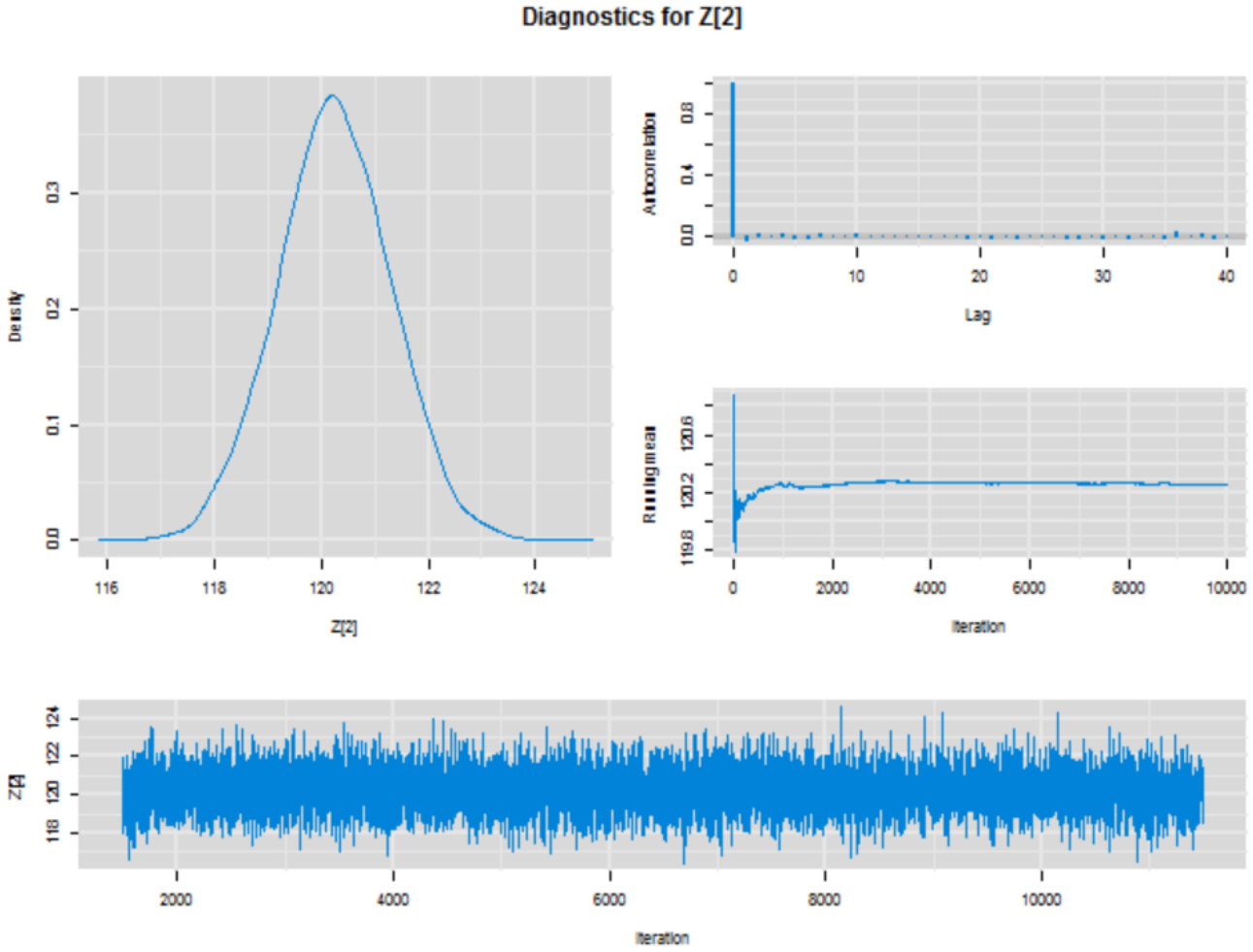
```
model_string <- "
 model {
  for(i in 1:n) {
   x[i,1:2] ~ dmnorm(mu[], prec[ , ])

  Z[i] <- (mu[2]-mu[1]) + W1 * (x[i,1]-mu[1]) + W2 * (x[i,2]-mu[2]) ## Lord-McNemar estimated gain

  }

rel.1 <- 0.857  ## score reliability for test 1 and test 2 are assumed known
rel.2 <- 0.835
W1 <- 1/(1-rho^2)*((((rho * sqrt(sigma[2]))/ sqrt(sigma[1]))*(1-rel.2))-rel.1+(rho^2)) ## weight 1
W2 <- 1/(1-rho^2)*((((rho * sqrt(sigma[1]))/ sqrt(sigma[2]))*(rel.1-1))+rel.2-(rho^2)) ## weight 2

  # Constructing the covariance matrix and the corresponding precision matrix.
  prec[1:2,1:2] <- inverse(cov[,])
  cov[1,1] <- sigma[1] * sigma[1]
  cov[1,2] <- sigma[1] * sigma[2] * rho
  cov[2,1] <- sigma[1] * sigma[2] * rho
  cov[2,2] <- sigma[2] * sigma[2]

  # Diffuse priors on all parameters which could, of course, be made more informative.
  sigma[1] ~ dunif(0, 10000)
  sigma[2] ~ dunif(0, 10000)
  rho ~ dunif(0, 1)
  mu[1] ~ dnorm(600, 0.00001)
  mu[2] ~ dnorm(730, 0.00001)

  # Generate random draws from the estimated bivariate normal distribution
  x_rand ~ dmnorm(mu[], prec[ , ])

 }
```

MCMC Diagnostics for the Correlation Parameter

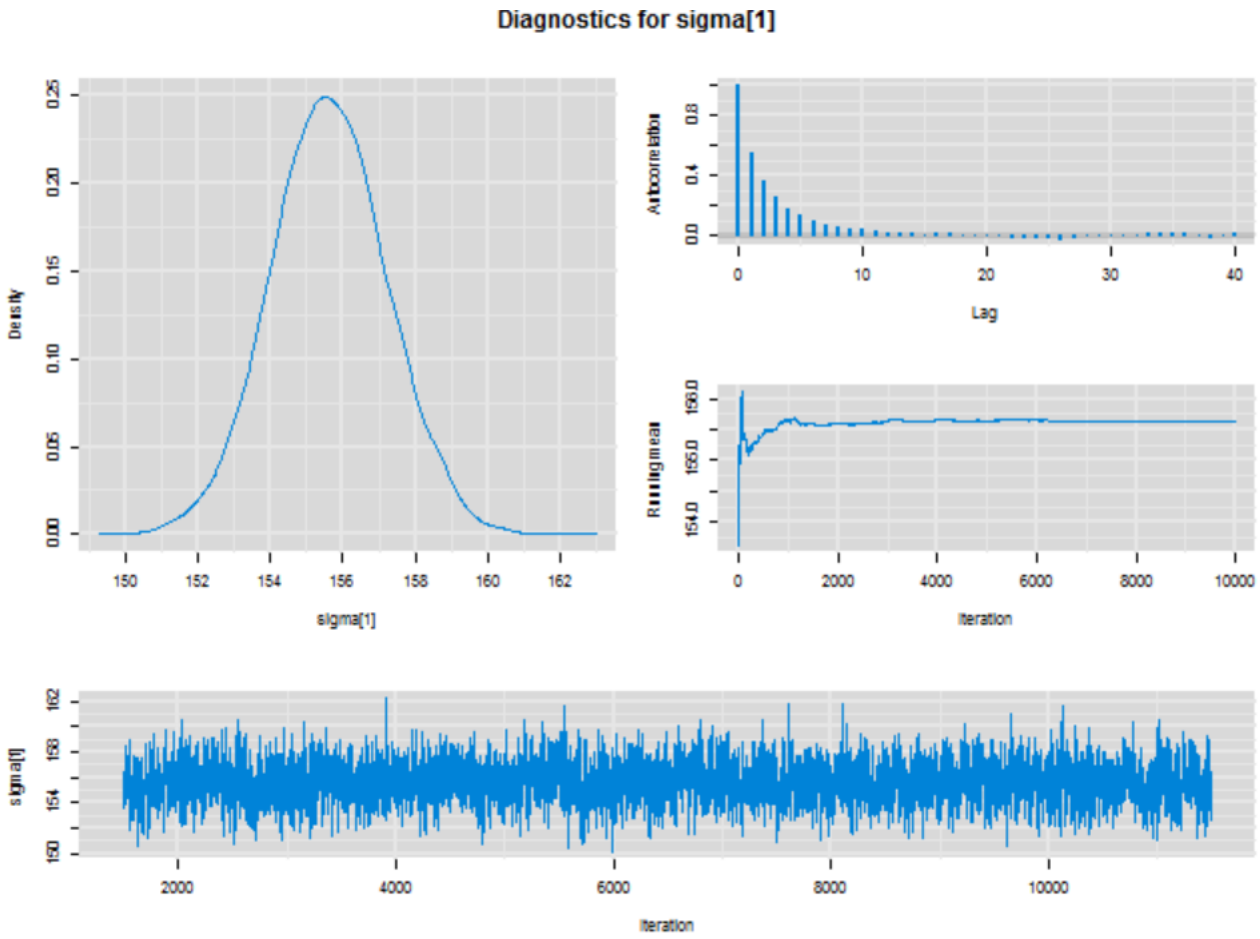


Houghton Mifflin Harcourt

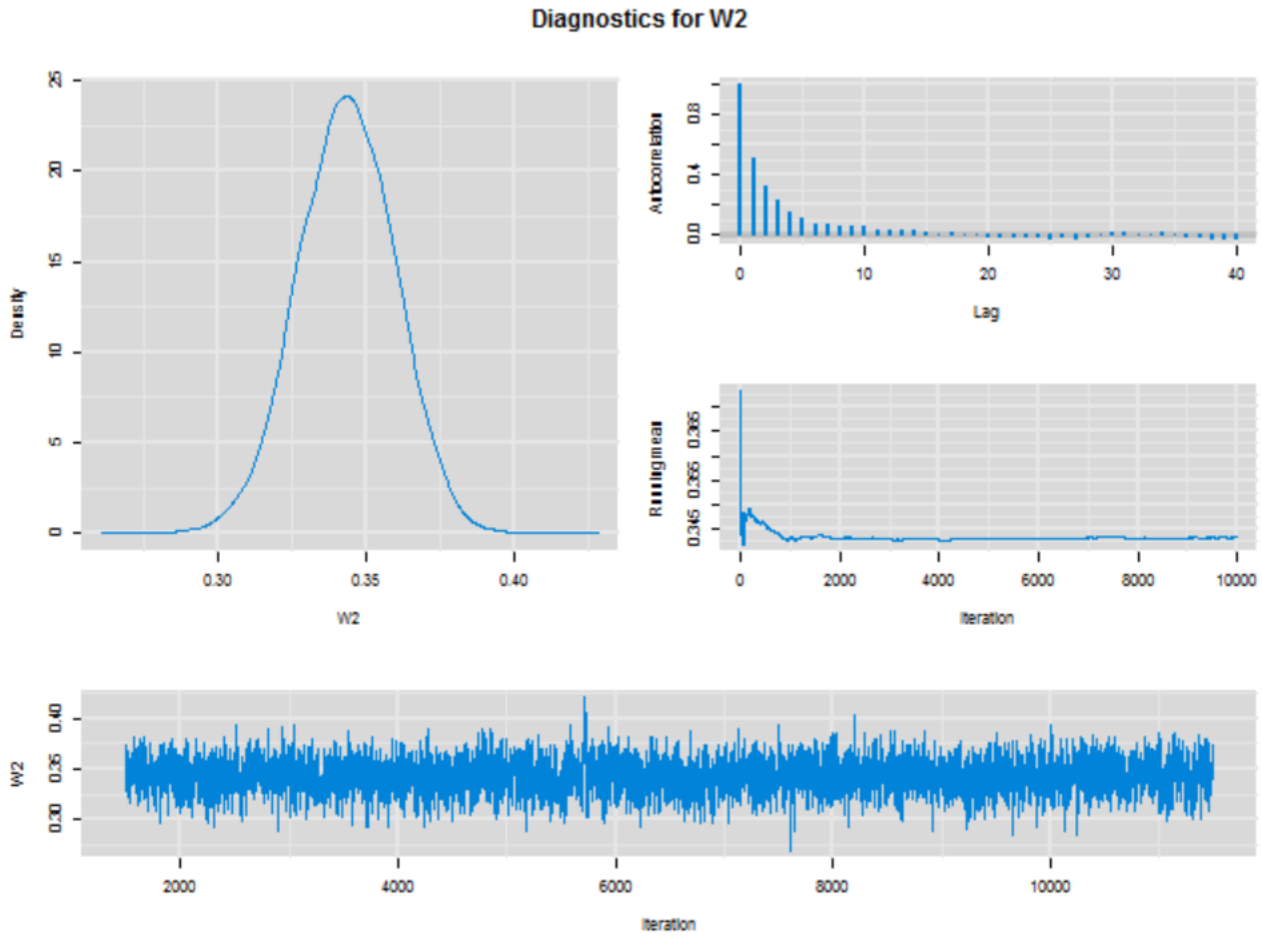
MCMC Diagnostics for a single estimated gain score



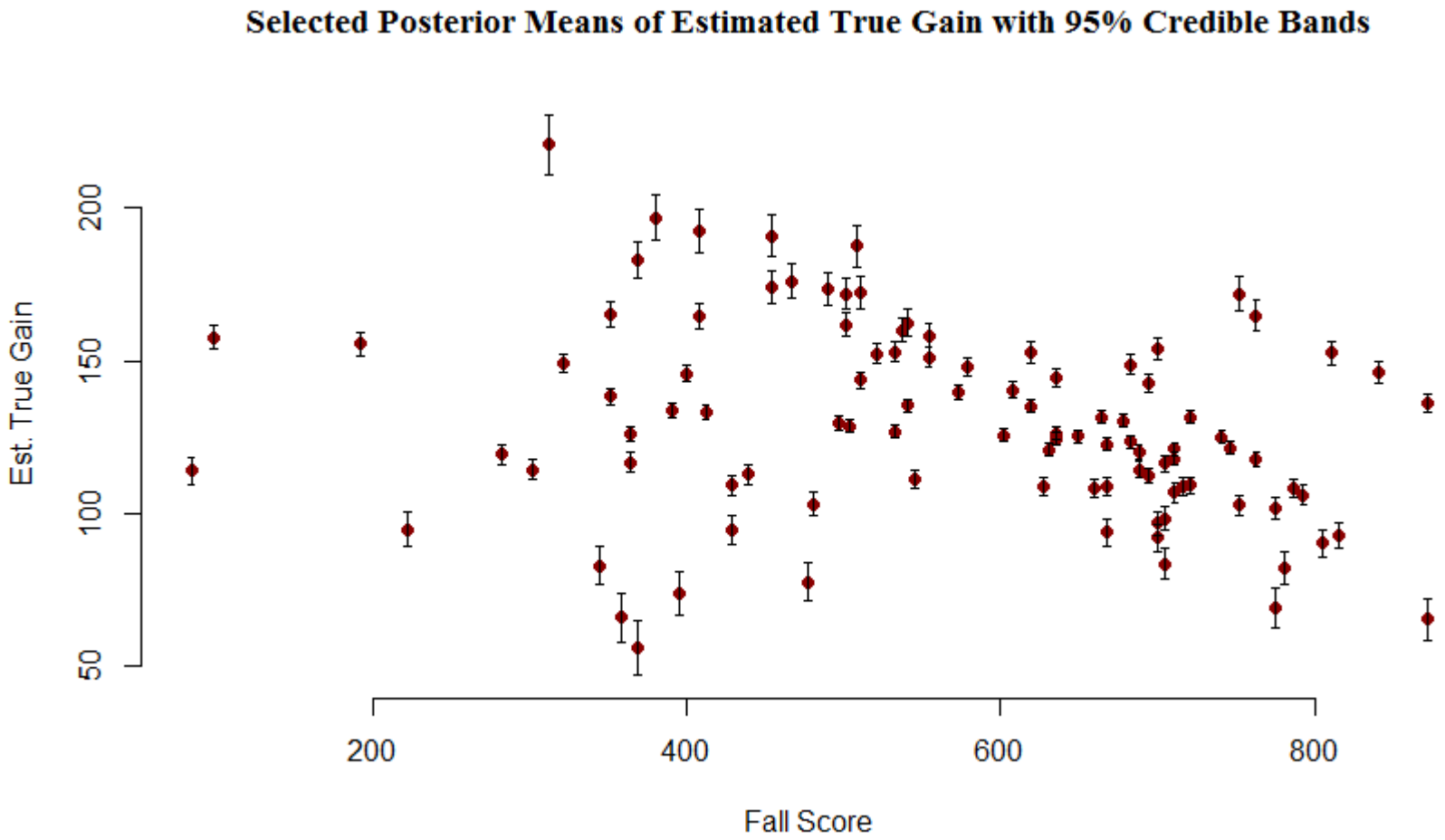
MCMC Diagnostics for the variance

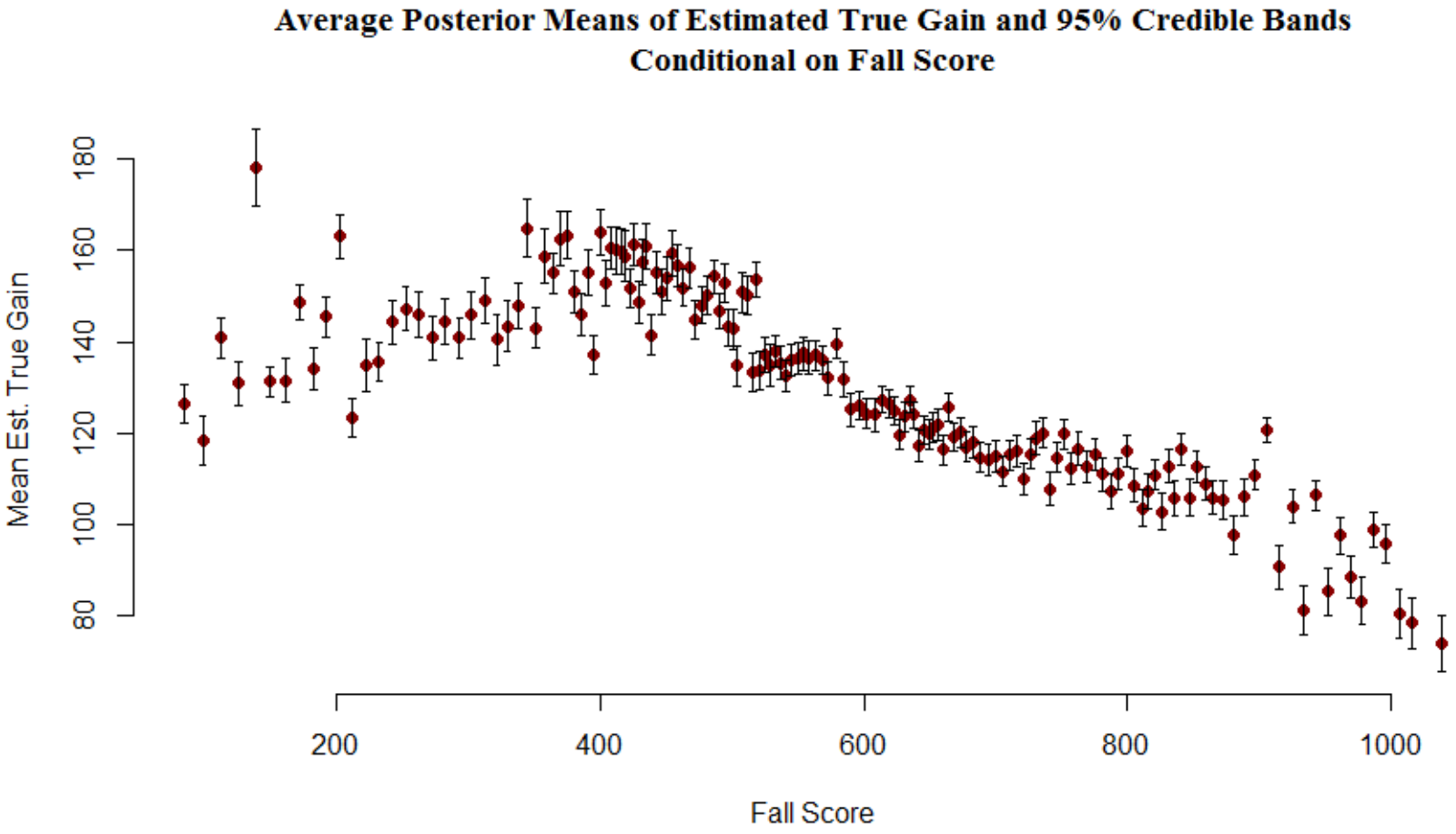


MCMC Diagnostics for the 2nd partial regression coefficient



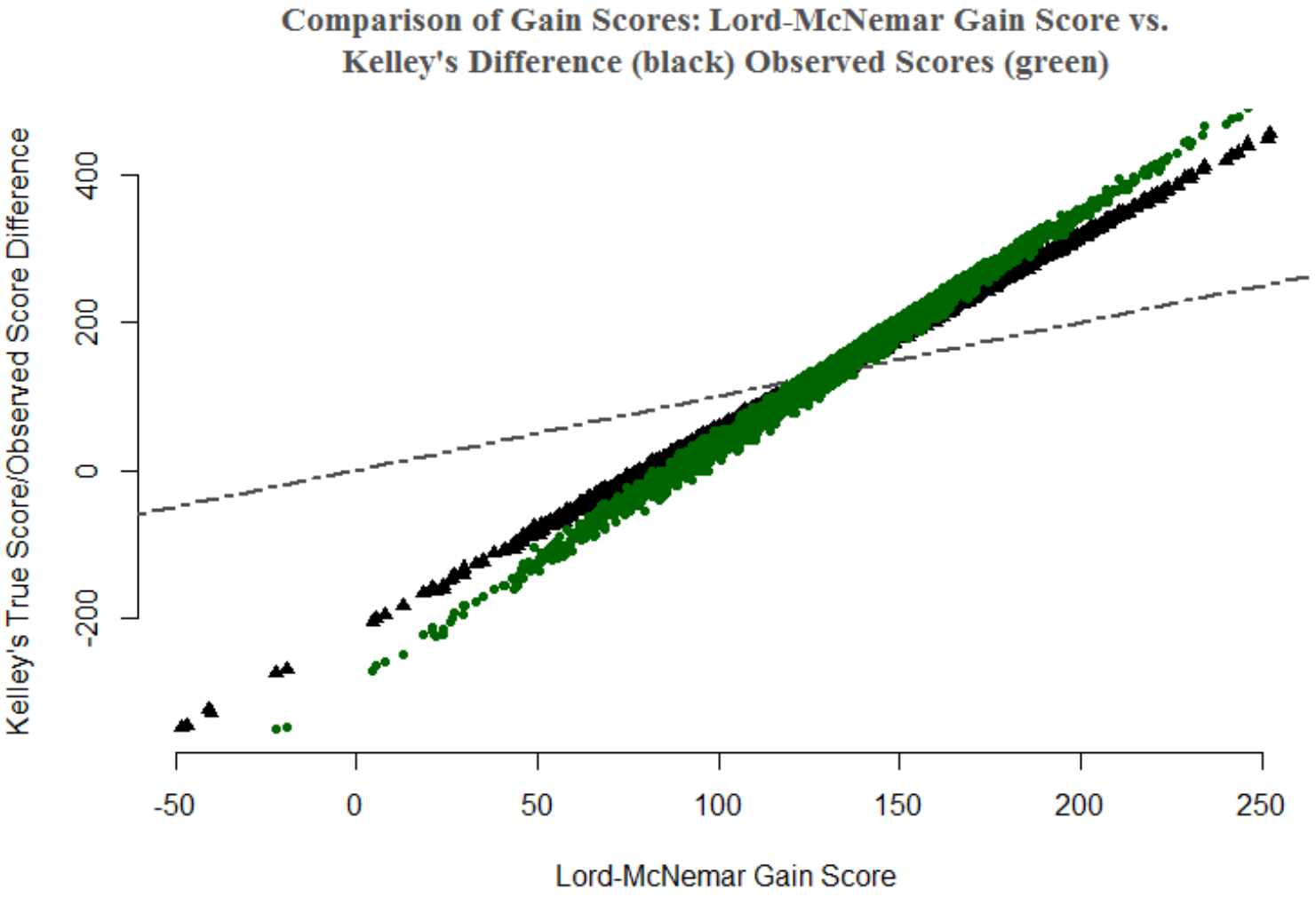
Houghton Mifflin Harcourt





Kelley's Bayesian Model in JAGS
Unknown Mean, Known Variances

```
model_string.K1 <- "model {
mu.T ~ dnorm(600, 0.0001)

tau.t <- 0.0000474;
tau.e <- 0.000284

sigma2.t <- 1/tau.t
sigma2.e <- 1/tau.e

   for (i in 1:n){
T[i] ~ dnorm(mu.T, tau.t)
x[i] ~ dnorm(T[i], tau.e);

}

reliability <- sigma2.t/(sigma2.t + sigma2.e); # reliability
}
"
```



谢谢

감사합니다