Measuring the Power of Learning.®

Research Memorandum
ETS RM–19-05

Documentation for the ScoreDiff R Function

Sandip Sinharay

June 2019

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Documentation for the ScoreDiff R Function

Sandip Sinharay
Educational Testing Service, Princeton, New Jersey

June 2019

Corresponding author: S. Sinharay, E-mail: ssinharay@ets.org

Suggested citation: Sinharay, S. (2019). *Documentation for the ScoreDiff R function* (Research Memorandum No. RM-19-05). Princeton, NJ: Educational Testing Service.

Action Editor: Gautam Puhan

Reviewers: Carol Eckerly and Jonathan Weeks

Abstract

This research memorandum introduces documentation for ScoreDiff, which is a function written in R. The function can be used to compute three statistics that can be used to implement score differencing for mixed-format tests, that is, tests that include both dichotomous and polytomous items. The three statistics include the Wald or Z statistic, the signed likelihood ratio (SLR) statistic, and the modified SLR statistic.

Key words: Modified signed likelihood ratio test; signed likelihood ratio test; score differencing; test fraud; Wald test

Acknowledgments

The work was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305D170026. The author would like to express his gratitude to Professor Cees Glas, who provided one of the real data sets discussed in this manual. The author is grateful to Carol Eckerly, Jonathan Weeks, and Gautam Puhan for their helpful comments and to Kim Fryer for editorial help. The author makes no warranty whatsoever with respect to this R function, whether express or implied by law, course of dealing, course of performance, usage of trade, or otherwise.

Wollack and Schoenig (2018) categorized the statistical methods to detect cheating (on tests) into six categories. One of these categories is *score differencing*—this category of methods focuses on a test of the hypothesis of equal performance or equal ability of an examinee over two sets of items. It is possible to apply score differencing to detect several types of test fraud, including fraudulent erasures, fraudulent and large gain scores, and item preknowledge, as demonstrated by Sinharay and Jensen (2019). Although test security has lately been an active area of research, there are currently no publicly available software packages to perform score differencing.

ScoreDiff is a function written in R (R Core Team, 2019) to compute three statistics that can be used to implement score differencing for mixed-format tests, that is, tests that include both dichotomous and polytomous items. The three statistics include the Wald or Z statistic (Fischer, 2003), the signed likelihood ratio (SLR) statistic (Sinharay, 2017), and the modified SLR (MSLR) statistic (Sinharay & Jensen, 2019).

Background: The Wald, Signed Likelihood Ratio, and Modified Signed Likelihood Ratio Statistics for Mixed-Format Tests

Consider a test comprising I items, each of which can be a dichotomously or polytomously scored item. Let the possible scores on item i be $0, 1, \ldots, m_i$. Let us assume that one is interested in applying score differencing to test the hypothesis of the equality of performance on item sets S_1 and S_2 for an examinee whose true overall ability is θ. The item sets S_1 and S_2 are nonoverlapping and together constitute all items on the test. Typically, the alternative hypothesis in score differencing is that the performance on one item set is better than that on the other due to reasons such as test fraud. Let us assume, without loss of generality, that score differencing has to be applied to test against the alternative hypothesis that the performance on S_2 is better than that on S_1 for the examinee. This alternative hypothesis is equivalent to the alternative hypothesis that the ability based on S_2 is larger than that based on S_1 for the examinee. Let $\boldsymbol{X} = (X_1, X_2, ...X_I)$ denote the scores for the examinee on the I items of the test. Let $\boldsymbol{X}_j$ denote the collection of the scores of the examinee on the items in Set $j, j = 1, 2$. Thus $\boldsymbol{X}_1 = \{X_i, i \in S_1\}$ and

$\boldsymbol{X}_2 = \{X_i, i \in S_2\}$. Let $P_{ij}(\theta) = P(X_i = j)$. For example, for the generalized partial credit model (GPCM; Muraki, 1992),

$$P_{ij}(\theta) = \frac{\exp[\sum_{h=0}^{j} a_i(\theta - b_{ih})]}{\sum_{c=0}^{m_i} \exp[\sum_{h=0}^{c} a_i(\theta - b_{ih})]},$$

where a_i is the slope parameter and b_{ih}s are the location or threshold parameters of item i (b_{i0} is assumed to be 0 for the GPCM, so there are effectively m_i location parameters for item i).

The log-likelihood of the item scores of the examinee, denoted as $l(\theta; \boldsymbol{X})$, can be computed as

$$l(\theta; \boldsymbol{X}) = \sum_{i=1}^{I} \sum_{j=0}^{m_i} d_j(X_i) \log P_{ij}(\theta), \tag{1}$$

where

$$d_j(X_i) = \begin{cases} 1 \text{ if } X_i = j \\ 0 \text{ otherwise.} \end{cases}$$

Note that if the ith item is dichotomous, then $m_i = 1$, and

$$d_0(X_i) = 1 - X_i, d_1(X_i) = X_i, P_{i0}(\theta) = P(X_i = 0)$$
$$P_{i1}(\theta) = P(X_i = 1).$$

For example, if the two-parameter logistic (2PL) model is used for item i, then

$$P_{i1}(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]} \text{ and } P_{i0}(\theta) = \frac{1}{1 + \exp[a_i(\theta - b_i)]},$$

where b_i is the difficulty parameter of item i.

For an examinee, let us define the maximum likelihood estimate (MLE) or the weighted maximum likelihood estimate (WLE; Warm, 1989) of the examinee ability

computed from the scores on item set S_1 as $\hat{\theta}_1$, that from the scores on S_2 as $\hat{\theta}_2$, and that from the scores on all the items as $\hat{\theta}$.

The Wald or Z statistic (e.g., Fischer, 2003) for testing the null hypothesis of equality of the examinee ability over S_1 and S_2 versus the alternative hypothesis that the ability based on S_2 is larger than that based on S_1 is given by

$$Z = \frac{\hat{\theta}_2 - \hat{\theta}_1}{\sqrt{\hat{V}(\hat{\theta}_2) + \hat{V}(\hat{\theta}_1)}}, \tag{2}$$

where, for example, $\hat{V}(\hat{\theta}_1)$ is the estimated variance of $\hat{\theta}_1$ and can be computed as the reciprocal of the estimated information under the item response theory (IRT) model. The Z statistic is assumed to follow the standard normal distribution under the null hypothesis (e.g., Fischer, 2003). A large value of the statistic leads to the rejection of the null hypothesis of no difference in performance over S_1 and S_2 and indicates a significantly better performance on item set S_2 compared to S_1.

The likelihood ratio test (LRT) statistic (e.g., Finkelman, Weiss, & Kim-Kang, 2010; Guo & Drasgow, 2010) for testing the null hypothesis of equality of the examinee ability over S_1 and S_2 is given by

$$\begin{aligned}
\Lambda &= 2\left[l(\hat{\theta}_1; \boldsymbol{X}_1) + l(\hat{\theta}_2; \boldsymbol{X}_2) - l(\hat{\theta}; \boldsymbol{X})\right] \\
&= 2\left[\sum_{i \in S_1}\sum_{j=0}^{m_i} d_j(X_i) \log P_{ij}(\hat{\theta}_1) + \sum_{i \in S_2}\sum_{j=0}^{m_i} d_j(X_i) \log P_{ij}(\hat{\theta}_2) - \sum_{i=1}^{I}\sum_{j=0}^{m_i} d_j(X_i) \log P_{ij}(\hat{\theta})\right] \\
&= 2\left[\sum_{i \in S_1}\sum_{j=0}^{m_i} d_j(X_i) \log \frac{P_{ij}(\hat{\theta}_1)}{P_{ij}(\hat{\theta})} + \sum_{i \in S_2}\sum_{j=0}^{m_i} d_j(X_i) \log \frac{P_{ij}(\hat{\theta}_2)}{P_{ij}(\hat{\theta})}\right].
\end{aligned}$$

To test the null hypothesis of equality of the examinee ability over S_1 and S_2 versus the alternative hypothesis that the ability based on S_2 is larger than that based on S_1,

Sinharay (2017) suggested the SLR statistic, given by

$$L_s = \begin{cases} \sqrt{\Lambda} \text{ if } \hat{\theta}_2 \geq \hat{\theta}_1 \\ -\sqrt{\Lambda} \text{ if } \hat{\theta}_2 < \hat{\theta}_1. \end{cases}$$

The statistic L_s has an asymptotic standard normal distribution (e.g., Cox, 2006; Sinharay, 2017, p. 104) under the null hypothesis. A large value of L_s leads to the rejection of the null hypothesis of no difference in performance over S_1 and S_2 and indicates a significantly better performance on item set S_2 compared to S_1.

Define the statistic Z' as

$$Z' = (\hat{\theta}_2 - \hat{\theta}_1) \sqrt{\frac{\sum_{i \in S_1} a_i^2 \operatorname{Var}(X_i|\hat{\theta}_1) \sum_{i \in S_2} a_i^2 \operatorname{Var}(X_i|\hat{\theta}_2)}{\sum_{i \in S_1} a_i^2 \operatorname{Var}(X_i|\hat{\theta}) + \sum_{i \in S_2} a_i^2 \operatorname{Var}(X_i|\hat{\theta})}}, \tag{3}$$

where, for example, $\operatorname{Var}(X_i|\hat{\theta}_1)$ is the conditional variance of X_i computed at $\theta = \hat{\theta}_1$.[1]

Under the GPCM, the statistic Z' can be shown to be very similar to the Z statistic shown in Equation 2. Sinharay and Jensen (2019) defined the MSLR statistic as

$$r^* = L_s + \frac{1}{L_s} \log \frac{Z'}{L_s},$$

and proved that if the GPCM fits the data, then the statistic has an asymptotic standard normal distribution under the null hypothesis of no difference in performance over S_1 and S_2. A large value of r^* leads to the rejection of the null hypothesis in favor of the alternative hypothesis that the performance on S_2 is better than that on S_1 for the examinee.

As several researchers have demonstrated (e.g., Fischer, 2003; Guo & Drasgow, 2010; Sinharay, 2017; Sinharay & Jensen, 2019), the Wald/Z, SLR/L_S, and MSLR/r^* statistics can be used to detect several types of test fraud, including fraudulent erasures, fraudulent and large gain scores, and item preknowledge.

The ScoreDiff Function, the Input, and the Output

The appendix shows the ScoreDiff function in its entirety.

The inputs to the function ScoreDiff are as follows:

1. *mirtoutput*, the output from fitting an IRT model to a data set using the mirt package (Chalmers, 2012)—note that mirtoutput would include information such as the data set and the item parameter estimates, and it is possible to extract additional information, such as the estimated examinee abilities, from mirtoutput

2. *id*, the row number (in the data set) of the examinee for whom score differencing has to be performed

3. *S2*, the second item set S_2, which would be used for the score differencing, in the form of a numerical vector; the complement of S_2 would be used as S_1

4. *est*, equal to ML or WLE, depending on whether the user wants to use the MLE or the WLE (Warm, 1989) as the ability estimate[2]

None of the above inputs has any default values.

The output produced by the function ScoreDiff is a vector consisting of the following six quantities for the examinee, corresponding to the row number that was provided as the input to the function: (a) $\hat{\theta}_2$; (b) $\hat{\theta}_1$; (c) $\hat{\theta}$; (d) the Wald statistic (e.g., Fischer, 2003); (e) the SLR statistic (Sinharay, 2017); and (f) the MSLR statistic (Sinharay & Jensen, 2019).

The MSLR statistic should be used (maybe in combination with the Wald and SLR statistics) only when a combination of the 2PL model and the GPCM (or their special cases, such as the Rasch model) is used to calibrate the data, because the asymptotic standard normal null distribution of the MSLR statistic holds only for a combination of the 2PL model and the GPCM (e.g., Sinharay & Jensen, 2019). Only the Wald and SLR statistics should be used when other IRT models (such as the 3PL model or the probit models) are used either for the dichotomous items or the polytomous items, or both. Therefore, for models other than a combination of the 2PL model and the GPCM, the ScoreDiff function prints "NA" for the value of MSLR. The SLR and MSLR statistics are typically very close to (say, within 0.05 of) each other, and the Wald statistic is typically close to (say, within 0.2 of) the other two statistics for almost all examinees.

When the absolute value of the SLR is small, the MSLR is unstable (e.g., Sinharay & Jensen, 2019); in the ScoreDiff function, MSLR is set equal to SLR whenever SLR is smaller than 0.05 in absolute value. However, this instability should not affect the conclusions much, because an examinee for whom SLR is small in absolute value has performed very similarly on S_1 and S_2, and the null hypothesis is not rejected for such an examinee.

Real Data Examples

Example 1: Detecting Item Preknowledge

The included data file ScoresDich includes the scores of 1,644 examinees on 170 dichotomous items. The data set is simulated and has the same number of items and examinees as a data set described in Cizek and Wollack (2017, p. 14). The latter data set was analyzed by Sinharay (2017) and Sinharay and Jensen (2019). A set of 61 items on the test is known to have been compromised for the original data set.

An annotated R script that can be used to apply the ScoreDiff function to examine the performance difference on the compromised and noncompromised items for the first examinee in the data set is provided in Figure 1. First, the ScoreDiff function is loaded into R from the file ScoreDiff.R.[3] Then, the ability estimate is set to be the MLE, and the number of items is set to 170. The item scores for all the examinees on the items are read from a file SimScoresDich. The set S_2 is defined as the set of the first 61 items. Then, the examineeNo variable is set to 1, indicating that score differencing would be performed for the first examinee in the data file. The columns of the file with item scores are then named I1, I2, . . . , I170.

In the next step, the MIRT software (Chalmers, 2012) is used to fit the 2PL model to the data set—the output is the object *mod*. In the next step, the command "out=ScoreDiff(mod,examineeNo,S2,est)" uses the function ScoreDiff to compute several quantities, including the Wald, SLR, and MSLR statistics for the first examinee in the data file, and saves them in "out." Finally, the contents of "out" are printed on the screen. The output for the first three examinees in the data set should look like that in Table 1.

```
source("ScoreDiff.R")
est="ML"
#Example 1: Detecting Item Preknowledge (dichotomous items)
nitem=170
scores=matrix(scan("SimScoresDich"),,nitem,byrow=T)#Read item scores
S2=1:61
examineeNo=1
colnames(scores)=paste("I",1:nitem,sep="")
mod=mirt(scores,1,itemtype=rep('2PL',nitem))
out=ScoreDiff(mod,examineeNo,S2,est)
cat(out,"\n")#Write the output for the examinee
```

Figure 1. **R codes for Example 1.**

Table 1
Output for Three Examinees for the First Real Data Example

$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}$	Z	SLR	MSLR
−0.46	−0.79	−0.64	1.03	1.05	1.00
2.57	1.46	1.55	1.21	1.47	1.32
2.51	1.90	1.95	0.68	0.76	0.61

Note. MSLR = modified signed likelihood ratio; SLR = signed likelihood ratio.

For the first three examinees, all three statistics are smaller than the 95th percentile of the standard normal distribution, so that Table 1 does not provide any evidence of item preknowledge for the first three examinees.

Example 2: Evaluating Performance Difference Over Two Subtests

The included real data file ScoresPoly includes the scores for 1,168 examinees on 16 five-category polytomous items on the first subtest of the neuroticism domain of the NEO Personality Inventory. These data were considered in Glas and Dagohoy (2007) and Sinharay and Jensen (2019). An annotated R script that can be used to apply the ScoreDiff function to examine the performance difference on the first half and second half of the test[4] for the first examinee in the data set is provided in Figure 2.

The GPCM is used for the analysis. The steps in the analysis are very similar to those in Figure 1—the item type is set to gpcm while using the mirt package to fit the

```
source("ScoreDiff.R")
est="ML"
nitem=16
scores=matrix(scan("ScoresPoly"),,nitem,byrow=T)
colnames(scores)=paste("I",1:nitem,sep="")
mod=mirt(scores,1,itemtype=rep('gpcm',nitem))
S2=1:8#The performance on the first and second halves will be compared
examineeNo=1
out=ScoreDiff(mod,examineeNo,S2,est)
cat(out,"\n")#Write the output for the examinee
```

Figure 2. **R codes for Example 2.**

GPCM to the data set. The output for the first three examinees in the file should look like that in Table 2.

Table 2
Output for Three Examinees for the Second Real Data Example

$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}$	Z	SLR	MSLR
2.47	0.38	0.88	2.24	2.39	2.37
2.26	0.47	0.92	1.97	2.08	2.06
−0.42	−0.15	−0.24	−0.41	−0.40	−0.44

Note. MSLR = modified signed likelihood ratio; SLR = signed likelihood ratio.

All three statistics are larger than the 95th percentile of the standard normal distribution for the first two examinees, so Table 2 provides some evidence of better performance on the first eight items compared to the last eight items for the first two examinees.

Program Availability and Condition of Use

The R function ScoreDiff (shown in the appendix), the two data sets discussed in the two data examples, and the R scripts to apply ScoreDiff to the two data sets (shown in Figures 1 and 2) are available at no cost for academic and other noncommercial use and can be downloaded from https://github.com/EducationalTestingService/ScoreDiff

References

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chang, H. H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52. https://doi.org/10.1007/BF02294469

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of detecting cheating on tests*. Washington, DC: Routledge.

Cox, D. R. (2006). *Principles of statistical inference*. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511813559

Finkelman, M., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement*, *34*, 238–254. https://doi.org/10.1177/0146621609344844

Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement*, *27*, 3–26. https://doi.org/10.1177/0146621602239474

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*, 159–180. https://doi.org/10.1007/s11336-003-1081-5

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored Internet tests: The Z-test and the likelihood ratio test. *International Journal of Selection and Assessment*, *18*, 351–364. https://doi.org/10.1111/j.1468-2389.2010.00518.x

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. https://doi.org/10.1177/014662169201600206

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: Author.

Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score

test. *Journal of Educational and Behavioral Statistics*, *42*, 46–68. https://doi.org/10.3102/1076998616673872

Sinharay, S., & Jensen, J. L. (2019). Higher-order asymptotics and its application to testing the equality of the examinee ability over two sets of items. *Psychometrika*, *84*, 484–510. https://doi.org/10.1007/s11336-018-9627-8

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. https://doi.org/10.1007/BF02294627

Wollack, J. A., & Schoenig, R. W. (2018). Cheating. In B. B. Frey (Ed.), *The Sage encyclopedia of educational research, measurement, and evaluation* (pp. 260–265). Thousand Oaks, CA: Sage.

Appendix: The ScoreDiff Function

```
library(mirt)
ScoreDiff=function(mod,id,S2,est)
{   nitem=mod@Data$nitems
    n2=length(S2)#Number of items in Item Set 2
    n1=nitem-n2
    S1=setdiff(1:nitem,S2)
    x=mod@Data$data[id,]
    x1=x
    x1[S2]="NA"
    scr=fscores(mod,method=est,response.pattern=x1)
    theta1=as.numeric(scr[(nitem+1)]) # Ability estimate based on  Item Set 1
    SE1=as.numeric(scr[(nitem+2)])#Standard error of Ability estimate for Item Set 1
    x2=x
    x2[S1]="NA"
    scr=fscores(mod,method=est,response.pattern=x2)
    theta2=as.numeric(scr[(nitem+1)]) # Ability estimate based on Item Set 2
    SE2=as.numeric(scr[(nitem+2)])#Standard error of Ability estimate for Item Set 2
    Z = (theta2-theta1)/sqrt(SE2^2+SE1^2)#Z or Wald Statistic
    scr=fscores(mod,method=est,response.pattern=x)
    theta=as.numeric(scr[(nitem+1)]) # Ability estimate based on all items
    L1=Loglike(mod,id,S1,theta1)# Log-likelihood based on Item Set 1
    L2=Loglike(mod,id,S2,theta2) # Log-likelihood based on Item Set 2
    Lall=Loglike(mod,id,1:nitem,theta) # Log-likelihood based on all items
    LRT=2*(L2+L1-Lall) #The likelihood ratio statistic for 2-sided test
    LRT[LRT<0]=0 # Reset the LRT's with negative values to 0
    Ls=ifelse(theta2>=theta1,sqrt(LRT),-sqrt(LRT)) # Calculate SLR Statistic
 # Calculate the MSLR for GPCMs
   MSLR="NA"
    if (sum(mod@Model$itemtype
    {itparms=coef(mod,IRTpars=TRUE)
    V=rep(0,3)
    for (i in 1:n1)
     {oneitem=extract.item(mod,S1[i])
      a=eval(parse(text=paste("itparms$I",S1[i],sep="")))[1]
      V[1]=V[1]+(a**2)*var.item(oneitem,theta1)
      V[3]=V[3]+(a**2)*var.item(oneitem,theta)}
    for (i in 1:n2)
     {oneitem=extract.item(mod,S2[i])
      a=eval(parse(text=paste("itparms$I",S2[i],sep="")))[1]
      V[2]=V[2]+(a**2)*var.item(oneitem,theta2)
      V[3]=V[3]+(a**2)*var.item(oneitem,theta)}
    q = (theta2-theta1)*sqrt((V[1]*V[2])/V[3])#Equation 23, Sinharay & Jensen, 2018
```

```
  # Calculate the MSLR (Equation 12, Sinharay & Jensen, 2018)
    MSLR = round(ifelse(abs(Ls)<0.05,Ls,Ls + log(q/Ls)/Ls),2)}
    return(c(round(c(theta2,theta1,theta,Z,Ls),2),MSLR))}
#
Probtrace = function(items, Theta){
 traces = lapply(items, probtrace, Theta=Theta)
   ret = do.call(cbind, traces)
   ret}
#
Loglike=function(mod,id,s,theta)
{
 items = vector('list', length(s))
 for(i in 1:length(s)) items[[i]] = extract.item(mod, s[i])
 traces = Probtrace(items, theta)
 fd <- mod@Data$fulldata[[1L]]
 f=NULL
 for (j in s)
  {f=cbind(f,fd[,grepl(paste("Item.",j,"_",sep=""),colnames(fd))])}
 f2=f[id,]
 LL <- rowSums(f2 * log(traces))
 return(LL) }
 #
 var.item <- function(x, Theta){
    if(missing(x)) missingMsg('x')
    if(missing(Theta)) missingMsg('Theta')
    if(is(Theta, 'vector')) Theta <- as.matrix(Theta)
    if(!is.matrix(Theta)) stop('Theta input must be a matrix', call.=FALSE)
    tmp <- try(x@nfact, TRUE)
    if(!is(tmp, 'try-error'))
        if(ncol(Theta) != x@nfact)
            stop('Theta does not have the correct number of dimensions', call.=FALSE)
    P <- probtrace(x=x, Theta=Theta)
    Emat <- matrix(0:(x@ncat-1), nrow(P), ncol(P), byrow = TRUE)
    Var <- rowSums(P * Emat * Emat) - (expected.item(x,Theta))^2
    return(Var)}
```

Notes

[1] For example, if the 2PL model is used for item i, then
$\text{Var}(X_i|\hat{\theta}_1) = P_{i1}(\hat{\theta})[1 - P_{i1}(\hat{\theta})] = \frac{\exp[a_i(\hat{\theta}_1 - b_i)]}{(1+\exp[a_i(\hat{\theta}_1 - b_i)])^2}$.

[2] Given the asymptotic results of Chang and Stout (1993), the ScoreDiff function may provide reasonable results if *est* is set equal to "EAP" or "MAP," but the asymptotic results of Sinharay (2017) and Sinharay and Jensen (2019) do not hold, and the ScoreDiff function was not tested for "EAP" or "MAP" estimates.

[3] This line of the program should be changed appropriately to provide the correct file path. For example, on a Windows machine, the correct path could be
C:/Users/mmcfly/Documents/Test Security/ScoreDiff.R

[4] Glas and Dagohoy (2007) and Sinharay and Jensen (2019) considered this partitioning in applying their tests of performance difference over two subtests for these data.