International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Use of Full Hierarchy Consistency Index to Assess Response Consistency

Lokman Akbay, Mustafa Kılınç

To cite this article: Akbay, L., & Kılınç, M. (2018). Use of Full Hierarchy Consistency Index to Assess Response Consistency. *International Journal of Assessment Tools in Education*, 5(1), 105-118. DOI: 10.21449/ijate.350499

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

This article may be used for research, teaching, and private study purposes.

Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

Authors alone are responsible for the contents of their articles. The journal owns the copyright of the articles.

The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of the research material.

Full Terms & Conditions of access and use can be found at
http://ijate.net/index.php/ijate/about

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 105-118

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Use of Full Hierarchy Consistency Index to Assess Response Consistency

Lokman Akbay*, **Mustafa Kılınç**

Educational Measurement and Evaluation, Mehmet Akif Ersoy University, Burdur, Turkey

Abstract: Measurement models need to properly delineate the real aspect of examinees' response processes for measurement accuracy purposes. To avoid invalid inferences, fit of examinees' response data to the model is studied through *person-fit* statistics. Misfit between the examinee response data and measurement model may be due to invalid models and/or examinee's aberrant response behavior such as cheating, creative responding, and random responding. Hierarchy consistency index (HCI) was introduced as a person-fit statistics to assess classification reliability of particular cognitive diagnosis models. This study examines the HCI in terms of its usefulness under nonhierarchical attribute conditions and under different item types. Moreover, current form of HCI formulation only considers the information based on correct answers only. We argue and demonstrate that more information could be obtained by incorporating the information that may be obtained from incorrect responses. Therefore, this study considers the full-version of the HCI (i.e., FHCI). Results indicate that current form of HCI is sensitive to misfitting item types (i.e., basic or more complex) and examinee attribute patterns. In other words, HCI is affected by the attribute pattern an examinee has as well as by the item s/he aberrantly responded. Yet, FHCI is not severely affected by item types under any examinee attribute pattern.

ARTICLE HISTORY

Received: 01 October 2017
Revised: 07 November 2017
Accepted: 20 November 2017

KEYWORDS

Person-fit,
Attribute Hierarchy Index,
Cognitive Diagnosis,

1. INTRODUCTION

Measurement models must play an important role in test construction and result interpretation processes of educational assessments. As a recent measurement model, cognitive diagnosis modeling has drawn great attention on the grounds of incorporating cognitive psychology in testing practices. Cognitive diagnosis models (CDMs) are the statistical models used to identify the knowledge and skills students mastered or failed to master in a particular domain. To accomplish this, associations between the test items and the measured knowledge or skills must be predefined. These measured knowledge, skills, cognitive processes, and problem solving steps are referred to as *attributes* (de la Torre, 2009; de la Torre & Lee, 2010) and the matrix reflecting items-by-attributes association is called *Q-matrix* (Tatsuoka, 1983). For example, if an item requires the first two attributes out of three attributes measured by a test, q-vector of this item is specified as [110] in the Q-matrix. Here 1 stands for required

*Corresponding Author E-mail: lokmanakbay@gmail.com

ISSN: 2148-7456 online /© 2018 DOI: 10.21449/ijate.350499

attribute and 0 indicates not required attribute. This vector signifies the fact that examinees are expected to be mastered the first two attributes to reach correct answer.

Starting with the pioneering work of Tatsuoka (1983), various approaches integrating cognitive theory into psychometric practices have been proposed. The rule space methodology (RSM: Tatsuoka, 1983), attribute hierarchy method (AHM: Leighton, Gierl, & Hunka, 2004), deterministic input, noisy "and" gate (DINA: Junker & Sijtsma, 2001), and generalized-DINA (GDINA: de la Torre, 2011) are among the examples of CDMs. In general, based on the presence of absence of K measured attributes, at most 2^K latent classes can be formed by a CDM where K indicates the number of attributes to be measured. For instance, when a test developed for cognitively diagnosis assessment measures three attributes, CDM analysis classifies examinees into, at most, eight possible latent classes (i.e., {000}, {100}, {010}, {001}, {110}, {101}, {011}, {000}). When an examinee is classified in {100} latent group, his/her estimated attribute pattern becomes [100], which indicates that the examinee has mastered the first attribute and has not mastered the second and third. The ultimate purpose of CDMs is to provide feedback on students' strengths and weaknesses based on the attribute pattern, which could be helpful to modify teaching and learning activities.

To evaluate examinees' performance, CDMs establish the relations between examinees' response data and their mastery status of attributes within measured domain. Probability of an examinee's correct response to a test item is modeled as a function of item parameters and examinee's mastery of the attributes (Cui & Leighton, 2009). For example, the DINA model assumes that an examinee correctly responds to an item as long as the examinee has mastered all the required attributes required for that item. Thus, for one item, examinees are spread into two distinct groups (i.e., examinees who have mastered all required attributes for the item and examinees lacking at least one required attribute). This group-specific deterministic response can be defined by

$$\eta_{lj} = \sum_{k=1}^{K} \alpha_{lk}^{q_{jk}}$$

where, η_{lj} is deterministic response of group l by item j (i.e., 1 or 0); K indicates total number of attributes measured by the test; α_{lk} is the group l's mastery status of attribute k; and q_{jk} is the kth element in the q-vector of item j, which indicates whether or not attribute k is required for correct response of item j.

Item response function (IRF) of the DINA model has a probabilistic component, which allows possibility of *guessing* (i.e., responding correctly when not all attributes are mastered) and *slip* (i.e., giving an incorrect response when all required attributes are mastered). Given examinee i's observed response to item j (i.e., X_{ij}), these two item parameters are denoted as $g_j = P(X_{ij} = 1|\eta_{ij} = 0)$ and $s_j = P(X_{ij} = 0|\eta_{ij} = 1)$ for guessing and slip parameters, respectively. Given the item parameters, the IRF of the DINA model is written as

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i) = P(X_{ij} = 1|\eta_{ij}) = g_j^{(1-\eta_{ij})}(1 - s_j)^{\eta_{ij}}$$

where $\boldsymbol{\alpha}_i$ is the attribute pattern of examinee i; η_{ij} is the expected response of examinee i to item j; X_{ij} is examinee i's observed response to item j; and g_j and s_j are the guessing and slip parameters of item j (de la Torre, 2009). For further information on the estimation and classification of the DINA model, readers may refer to de la Torre (2009).

Measurement accuracy of examinees is directly related to appropriateness of measurement model, which need to properly delineate the real aspect of examinees' response

processes (Cui & Leighton, 2009). For instance, when attributes hold a hierarchical structure (i.e., some of the attributes are prerequisite to master others), not all 2^K latent classes are permissible. Therefore, examinees' response data should be analyzed accordingly. Thus, identification of the attributes, attribute structure, and attribute specifications in the Q-matrix must be precise. Otherwise, invalid inferences about examinees' knowledge states could be made. Furthermore, to avoid invalid inferences, fit of examinees' response data to the model is studied through 'person-fit' statistics. By means of person-fit statistics, examinees who are not being measured well by the test are identified (Cui & Leighton, 2009). Misfit between the examinee response data and measurement model may be due to invalid models and/or examinee's aberrant response behavior (e.g., cheating, creative responding, and random responding).

Cui and Leighton (2009) have introduced a person-fit index to assess classification reliability of specific cognitive diagnosis models (e.g., attribute hierarchy model [AHM: Leighton, Gierl, & Hunka, 2004]). This person-fit index is referred to as hierarchy consistency index (HCI) as it was also used by Cui (2007) to measure the accuracy of specified hierarchical structure of attributes in AHM. More information on the index is provided below.

1.1. Hierarchy consistency index (HCI)

Cui and Leighton (2009) introduced a person-fit statistic to detect misfit between item responses and the cognitive model. This fit statistic is called hierarchy consistency index (HCI) and ranges from -1.0 to 1.0. Statistics close to 1.0 indicate good fit between examinee responses and the model whereas statistics close to -1.0 indicate misfit. Definition of HCI is given in equation 1, which is borrowed from Cui and Leighton (2009), p 436. As it would be seen from the formula on Figure 1, HCI operates based on the match between an examinee's observed item responses and expected item responses based on a hierarchical relationships among measured attributes.

$$HCI_i = 1 - \frac{2\sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{ij}(1 - X_{ig})}{N_{C_i}}$$

where X_{ij} is examinee i's binary response to item j where 0 indicates incorrect response and 1 stands for a correct response; $S_{correct_i}$ is an index set that includes items requiring the subset of attributes required by item j when examinee's response to item j is correct; X_{ig} is examinee i's response to item g where item g belongs to $S_{correct_i}$; and N_{C_i} is the total number of comparisons for all the items correctly responded by examinee i.

2. ARGUMENT

When index is computed solely for the correct responses, some correct responses require less comparison than others. For example, imagine a test measuring three hierarchically structured attributes, in which attribute-1 (A1) is the most basic and attribute-3 (A3) is the most complex attribute. Here, when an item requiring A3 is correctly answered by an examinee, all other responses of the examinee are also expected to be correct. Thus, all other item responses are considered in index computation. Yet, when an examinee correctly responses an item requiring only A1 (i.e., the most basic attribute) only, only the items requiring sole A1 are considered for HCI computation. The potential problems in this regard are depicted below in a scenario where three hierarchical attributes are measured by a 10-items test, for which the Q-matrix is given in Table 1 and hierarchical structure of attribute is given in Figure 1.

Table 1. Q-matrix for 10-items test

Items	A1	A2	A3
1	1	0	0
2	1	1	0
3	1	1	1
4	1	0	0
5	1	1	0
6	1	1	1
7	1	0	0
8	1	1	0
9	1	1	1
10	1	0	0

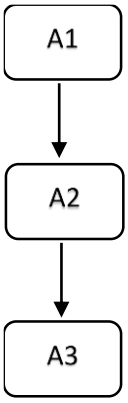

Figure 1. Linear hierarchy of three attributes

When an examinee's true attribute pattern is [000], expected responses of the examinees to all items becomes incorrect (i.e., 0). However, because of probabilistic component of the models, this examinee may correctly respond to one item. When we consider this *guessed* item only in HCI computation, all the comparisons we do will yield a misfit. Thus the computed HCI will be -1, which will, in turn, indicate that this examinee's responses do not fit to model. In fact, there is only one response that contradicts with the model expectancy. Imagine another examinee whose true attribute pattern is [111]. In this case expected responses of this examinee will be all correct. When the examinee misses one item, then only the comparisons due to that item will be left. Moreover, among the all comparisons conducted for the correct responses, only this incorrect response will yield misfit. There will be some reduction in the HCI due to this one misfit, yet the impact of this *slipped* item will not be as large as it is in previous case. Furthermost, because it will change the comparisons counted toward HCI, items missed by the examinee also matter.

Table 2. Two examinees and their HCI indices based on hypothetical response patterns

Examinees	Attribute profile	Response data	HCI
E1	000	**1**000000000	-1.000
E1	000	00**1**0000000	-1.000
E2	111	**0**111111111	0.667
E2	111	11**0**1111111	0.917

This scenario and resulting HCIs are summarized in Table 2. When E1 (i.e., an examinee with an attribute pattern [000]) guesses only one item, than HCI becomes -1. When E2 (i.e., an examinee with an attribute pattern [111]) slips one item, than HCI becomes smaller than 1.0, yet impact of slipped item is determined by the q-vector of the item. In other words, whether slipped item requires basic attribute or complex attribute matters. In above case, when an item requiring the most basic attribute is missed, HCI becomes .667. Impact of missed item when it requires the most complex attribute is relatively smaller (i.e., computed HCI is .917). As can be seen, although there is only one misfitted item in all cases, their impact on examinees' response consistency is different under different conditions.

2.1. Full Hierarchy Consistency Index

It should be noted here that *guessing* does not necessarily mean random guessing in cognitive diagnosis modelling framework, rather it means completing a task employing any other strategy that is not specified by the model. Therefore, guessing and slip behaviour of

examinees may be different for items requiring basic ore more complex attributes. From this point of view, consistency index should not be dramatically affected by the attribute-and-item specification of misfitted item. One possible way to control this is to consider all items for examinee response fit, which can be implemented by adding a second component to HCI that includes comparisons for item sets consists of items that are expected to be incorrectly responded by the examinee. Then the *full* version of the index may be represented as

$$FHCI_i = 1 - \frac{2[\sum_j \sum_{j' \in S_{j-correct}} X_{ij}(1 - X_{ij'}) + \sum_j \sum_{j'' \in S_{j-incorrect}} (1 - X_{ij}) X_{ij''}]}{N_{C_i}}$$

where X_{ij} is examinee i's binary response to item j where 0 indicates incorrect response and 1 stands for a correct response; $S_{j-correct}$ is an index set that includes items requiring the subset of attributes required by item j when examinee's response to item j is correct; $S_{j-incorrect}$ is an index set that includes items requiring all the attributes required by item j when the item incorrectly answered by the examinee; $X_{ij'}$ is examinee i's response to item j' where item j' belongs to $S_{j-correct}$; $X_{ij''}$ is examinee i's response to item j'' where item j'' belongs to $S_{j-incorrect}$; and N_{C_i} is the total number of comparisons for all the items responded by examinee i. This full version of the index will be referred to as *full hierarchy consistency index (FHCI)* throughout this paper. Computed FHCI indices for two previous examinees with certain response patterns are given in Table 3. Results based on FHCI are quite acceptable under all conditions.

Table 3. Two examinees and their FHCI indices based on the response patterns

Examinees	Attribute profile	Response data	HCI	FHCI
E1	000	**1**000000000	-1.000	0.765
E1	000	00**1**0000000	-1.000	0.438
E2	111	**0**111111111	0.667	0.429
E2	111	11**0**1111111	0.917	0.840

This study aims to focus on the following question:

- How successfully HCI is used under nonhierarchical attribute conditions (i.e., unstructured attribute cases) to identify aberrantly responded examinees,
- What is the impact of q-vector of a misfitting item on the HCI. More specifically, this study aims to unveil the impact of a misfitting item on HCI when it measures basic or more complex attributes,
- What is the distribution of misfitting examinees when number of misfits is equal across all permissible latent classes,
- Current form of HCI formulation only considers the information based on correctly answered items. Thus, more information could be obtained by incorporating the information that may be obtained from incorrect responses. Therefore, this study considers the Full-version of the HCI such that examinees' all responses rather than only correct responses are taken into account for consistency index computation.

3. METHOD

A simulation study and a real data analysis were conducted. In the simulation study, number of examinees, number of items and number of attributes were fixed to 2000, 20, and 6; respectively. Corresponding Q-matrix (i.e., item-by-attribute matrix) is given in Table 4. Corresponding Q-matrices for linear and divergent cases are given in Appendices. In the item response data generation, uniform examinee distribution was assumed. Two types of

hierarchical structures (i.e., linear and divergent) and an unstructured attribute case were considered. These hierarchical attribute structures can be seen in Figure 2. Four types of item misfits were considered:

Table 4. Generating Q-matrix

	Attributes							Attributes					
Items	A1	A2	A3	A4	A5	A6	Items	A1	A2	A3	A4	A5	A6
1	1	0	0	0	0	0	11	0	0	0	0	1	1
2	0	1	0	0	0	0	12	1	0	0	0	0	1
3	0	0	1	0	0	0	13	1	1	1	0	0	0
4	0	0	0	1	0	0	14	0	1	1	1	0	0
5	0	0	0	0	1	0	15	0	0	1	1	1	0
6	0	0	0	0	0	1	16	0	0	0	1	1	1
7	1	1	0	0	0	0	17	1	0	0	0	1	1
8	0	1	1	0	0	0	18	1	1	0	0	0	1
9	0	0	1	1	0	0	19	1	0	0	0	0	0
10	0	0	0	1	1	0	20	0	0	0	0	0	1

1. Creative responding (high guessing and slip in items requiring basic attributes)
2. Difficult (high slip in the complex items only)
3. Logical (high guessing in the items requiring basic attributes and high slip in the items requiring complex attributes)
4. Uniform (distribution of guessing and slip is uniform across all items)

For the *creative response* items, the lowest and highest success probabilities (i.e., P(0) and P(1)) were generated from U(0.20, 0.30) and U(0.70, 0.80), respectively, for items requiring basic attributes. These probabilities drawn from U(0.10, 0.20) and U(0.80, 0.90), respectively, for items requiring complex attributes. Lowest success probability of both basic and complex items in the *difficult* item case were generated from U(0.10, 0.20). In contrast, the highest success probabilities were generated from U(0.80, 0.90) and U(0.70, 0.80), respectively, for the basic and complex items. In the *logical* item case, the lowest and highest success probabilities were generated from U(0.20, 0.30) and U(0.80, 0.90), respectively, for items requiring basic attributes. Corresponding distributions for the complex item case were U(0.10, 0.20) and U(0.70, 0.80), respectively. Lastly, the lowest and highest success probabilities of examinees for both basic and complex items were generated from U(0.10, 0.20) and U(0.80, 0.90), respectively. These conditions are summarized in Table 5.

HCI and FHCI were employed to demonstrate extra information that can be obtained from incorrect responses. The data generation was based on the DINA model (de la Torre, 2009; Junker and Sijtsma, 2001). Throughout the study data generation performed using the OxMetrics programming language (Doornik, 2011) and index computation was performed in R-version 3.3.3. Simulation study is followed by a real data analysis. Data consist of 2922 examinees' binary responses to the 28 items in the grammar section of the ECPE examination. The test was developed and administered in University of Michigan English Language Institute in 2003. The dataset and the Q-matrix are available in and obtained from the `CDM' package (Robitzsch, Kiefer, George, & Uenlue, 2014) in R software environment.

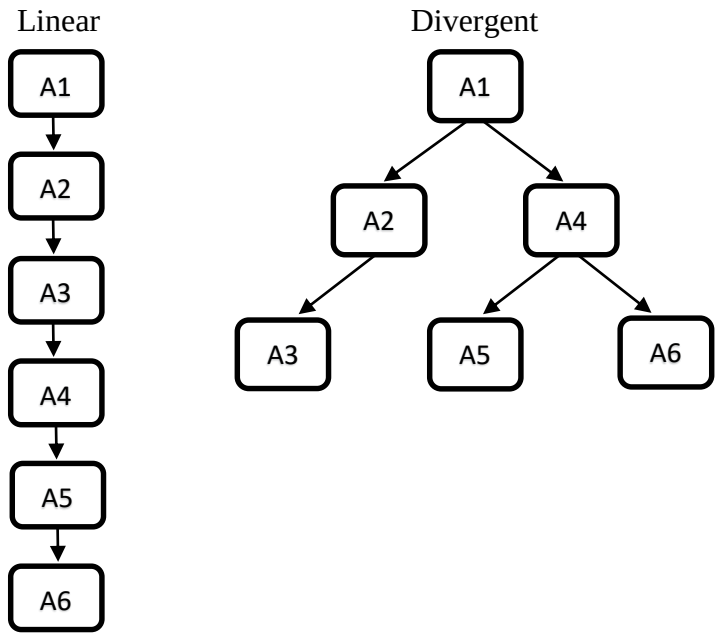

Figure 2. *Linear and divergent hierarchical structures.*

Table 5. Success probability distributions of item types

Item Types	Items with basic attributes		Items with complex attributes	
	P(0)	*P(1)*	*P(0)*	*P(1)*
Creative response	U(0.20, 0.30)	U(0.70, 0.80)	U(0.10, 0.20)	U(0.80, 0.90)
Difficult	U(0.10, 0.20)	U(0.80, 0.90)	U(0.10, 0.20)	U(0.70, 0.80)
Logical	U(0.20, 0.30)	U(0.80, 0.90)	U(0.10, 0.20)	U(0.70, 0.80)
Uniform	U(0.10, 0.20)	U(0.80, 0.90)	U(0.10, 0.20)	U(0.80, 0.90)

4. RESULTS

4.1. Simulation Results

Simulation results based on the HCI are given in Figure 3 as a matrix of scatterplots depicting HCI distribution of 2000 examinees where examinees are ordered based on the number of attributes they mastered. For instance, first a few hundreds of examinees in the linear case have the generating attribute pattern of [000000]; while very last a few hundreds have the generating attribute pattern of [111111]. Considering this order and the fact that all examinees' fit levels are approximately equal, it's very clear from the figure that HCI tends to be negative when an examinee has mastered smaller number of measured attributes. This reality emerges from the fact that when examinee guesses an item all other items requiring the subset of attributes specified in the guessed item are counted toward comparisons employed in index computation. HCI may be a good indicator of person fit when examinee has mastered most of the attributes, however, it may not be a good indicator for examinees who have lack of many attributes.

It can also be observed from Figure 3 that when number of latent classes decreases (i.e., hierarchy becomes more stringent) variance of HCI distribution shrinks. For example, in all types of item cases, HCI variance across attribute patterns is smaller when attributes are linearly structured. When attributes have no hierarchical structure (i.e., unstructured attribute case), HCI for examinees in any latent class are more disperse. Although item types do not make substantial differences, slight changes in the scatter plots by item types are observed. For

instance, in the difficult item case (i.e., high slips in the complex items only), HCI distribution of examinees who mastered more than half of the attributes are more disperse than the distribution of examinees who mastered a few attributes. Similarly, when creative item types are administered, variance of HCI of examinees lacking complex attributes elevates. These results are not surprising because when probabilistic component of item responses increases, examinees' observed responses deviate from the expected responses such that person-fit reduces.

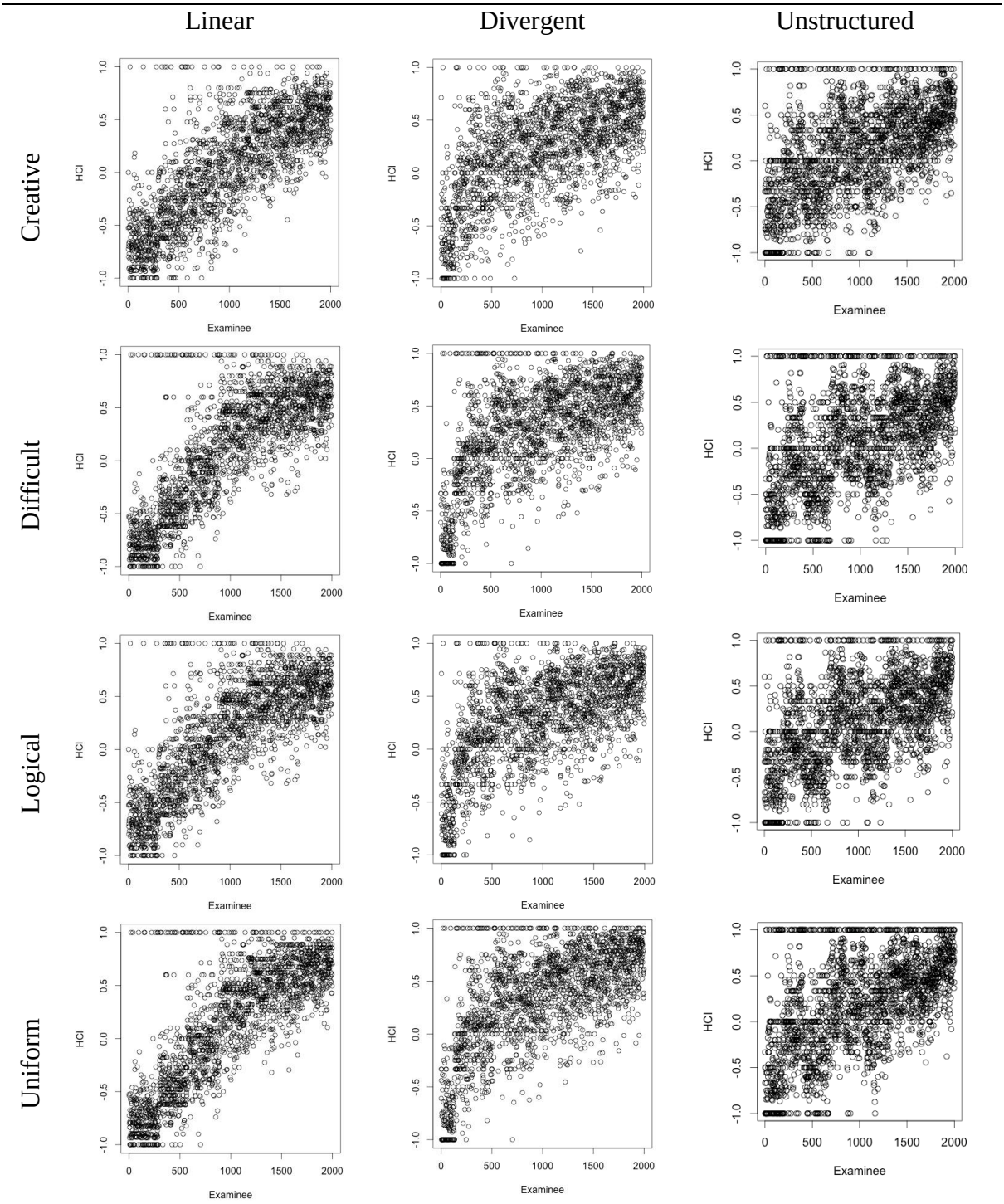

Figure 3. Matrix of scatterplots of HCI under various item types and attribute hierarchies

One major purpose of this study was to unveil the general improvement in identifying person-fit when not only correct responses but also incorrect responses are considered in person fit index computation. Results based on the FHCI are given in Figure 4. It can easily be seen at

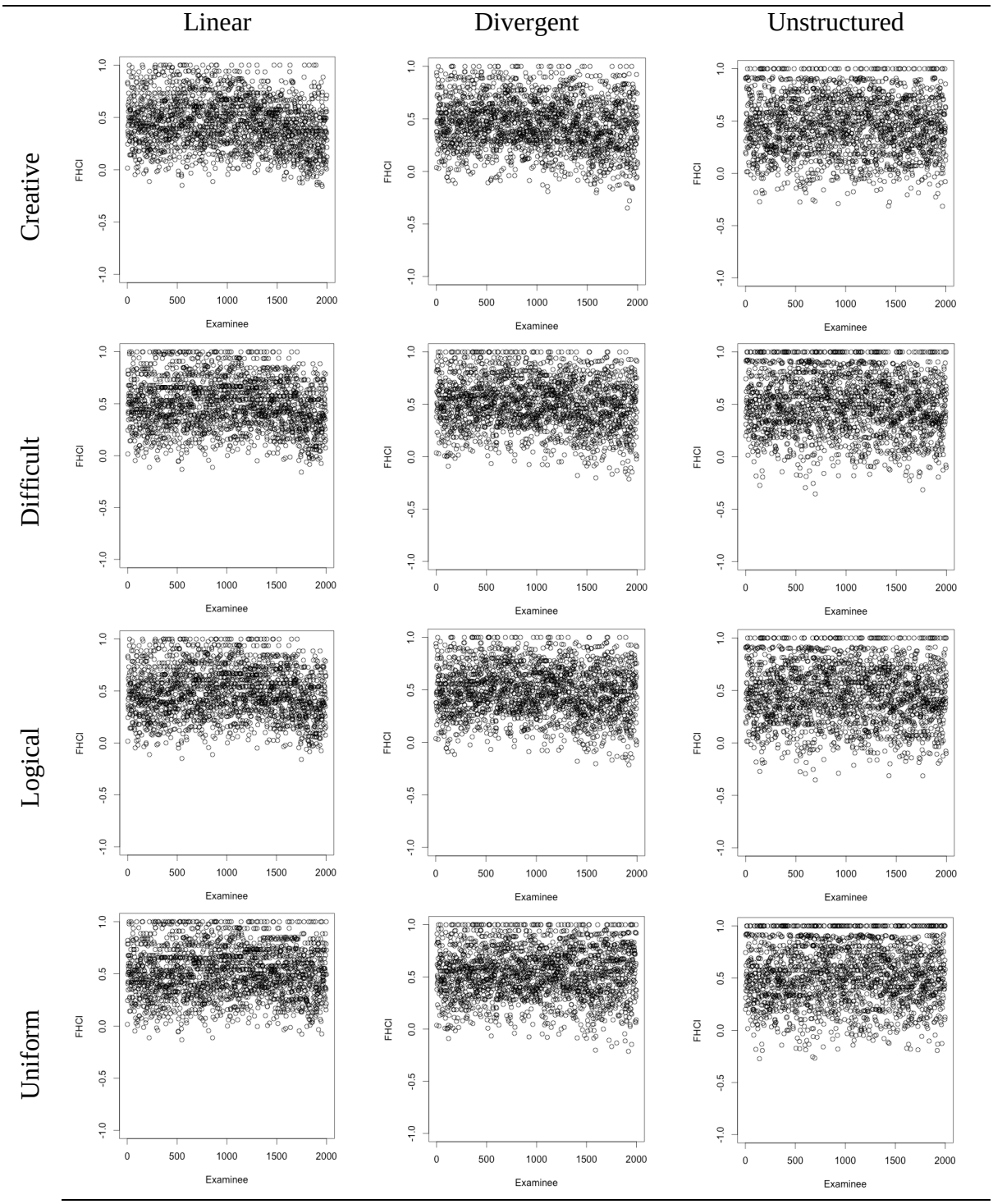

Figure 4. Matrix of scatterplots of FHCI under various item types and attribute hierarchies

first glance that, regardless of item type, attribute structure, and latent class an examinee is in, person-fit approximately falls between 0.00 and 1.00. This result suggests that FHCI may be considered as a more accurate person-fit index as it is not affected by examinees' attribute

pattern distribution (i.e., it measures fit in the same level of accuracy when examinee has mastered all or none of the measured attributes). Moreover, attribute structure does not significantly affect the results (i.e., variance of fit indices in the scatterplots are almost equal across linear, divergent, and unstructured attribute cases). Lastly, when FHCI is employed, small differences arising out of item types (i.e., creative, difficult, logical, and uniform) also diminished or even disappeared.

4.2. Real Data Analysis

Binary responses of 1922 examinees to 28 grammar items in the examination for the certificate of proficiency in English (ECPE) examination were analyzed in terms of examinees' person-fits. Q-matrix of the test and the data were obtained from 'CDM' package in R software environment. The data were analyzed previously by Templin and Bradshaw (2014) and specified a linear hierarchy among the three attributes (i.e., lexical rules, cohesive rules, and morphosyntactic rules) test is measuring. Scatter plots of examinees' person-fit results obtained by employment of HCI and FHCI are given Figure 5. When we look at the figure, FHCI result consistent with the simulation results, while HCI shows relatively better person-fit than what was observed in the simulation results.

However, remember that HCI fails to detect true person-fit when examinees did not master measured attributes. Assuming that the test truly measured aforementioned attributes and Q-matrix is correctly specified, correct answer proportions (proportion-corrects) of items may reflect attribute-pattern distribution of examinees. Proportion-correct of items are given in Table 6. Minimum and maximum proportion-corrects are .45 and .90, respectively. Moreover, 19 out of 28 items have been correctly answered by and over 70% of examinees, while only three items have been correctly answered by less than 50% of examinees. These results imply that many examinees in the sample have mastered two to three attributes. In the light of above information, person-fit result based on HCI could be more reflective of simulation results if there were more examinees lacking more than half of the attributes in the sample.

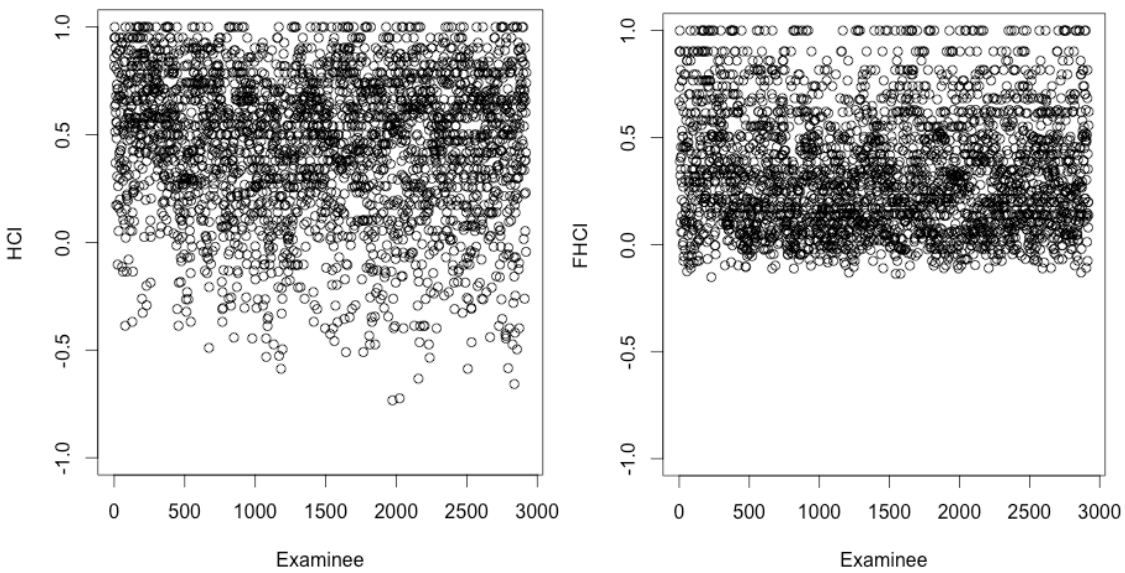

Figure 5. Scatter plots obtained by HCI and FHCI for ECPE data

Table 6. Proportion correct

Items	Proportion correct	Items	Proportion correct	Items	Proportion correct	Items	Proportion correct
1	.80	8	.90	15	.88	22	.63
2	.83	9	.70	16	.70	23	.81
3	.58	10	.66	17	.89	24	.53
4	.71	11	.72	18	.85	25	.62
5	.89	12	.43	19	.71	26	.70
6	.85	13	.75	20	.46	27	.45
7	.72	14	.65	21	.76	28	.82

min.=.43; mean=.71; max.=.90

5. CONCLUSION

HCI and FHCI have been employed under various conditions in this research. In data generation procedure guessing and slip for any item types did not exceed .30 (i.e., maximum P(0) = U(.20, .30) and minimum P(1) = U(.70, .80)). Thus, all examinees with different attribute patterns fit to the model equally well. Results suggested that HCI is a good indicator of person-fit as long as examinee has mastered most of the attributes. However, it fails to capture fitting examinees when examinees lack of many attributes. Conversely, FHCI may be considered as a more accurate person-fit index as it is not affected by examinees' attribute pattern distribution (i.e., it measures fit in the same level of accuracy when examinee has mastered all or none of the measured attributes).

Furthermore, FHCI is robust to different types of items such that impacts of misfit on basic and complex items are comparable. Therefore, more correct results yielding accurate inferences may be obtained by employment of FHCI. Study results demonstrated that regardless of item type, attribute structure, and latent class an examinee is in, FHCI approximately falls between 0.00 and 1.00. These results may be considered to form a cut-off to make a decision when FHCI is used to determine whether an examinee's responses fit to model. So, as long as an examinee's FHCI is positive (i.e., larger than .00), we may postulate this person's fit to model as acceptable. Lastly, in cases where we use FHCI as a measure of hierarchy consistency (i.e., whether assumed hierarchy for the model is acceptable), we should look for the distribution of examinees' FHCI, which need to be ranging from .00 to 1.00.

Disclosure Statement

No potential conflict of interest was reported by the authors.

6. REFERENCES

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46,* 429-449.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34,* 115-130.

Doornik, J. A. (2011). *Object-oriented matrix programming using Ox* (Version 6.20). London: Timberlake Consultants Press.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258-272.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuokas rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187-212.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2014). *CDM: Cognitive diagnosis modeling*. R package version 5.8-9. https://CRAN.R-project.org/package=CDM

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79,* 317-339.

7. APPENDICES

Appendix A. Q_matrix by the linear attribute structure

	Attributes							*Attributes*					
Items	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*	*Items*	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*
1	1	0	0	0	0	0	*11*	1	1	1	1	1	1
2	1	1	0	0	0	0	*12*	1	1	1	1	1	1
3	1	1	1	0	0	0	*13*	1	1	1	0	0	0
4	1	1	1	1	0	0	*14*	1	1	1	1	0	0
5	1	1	1	1	1	0	*15*	1	1	1	1	1	0
6	1	1	1	1	1	1	*16*	1	1	1	1	1	1
7	1	1	0	0	0	0	*17*	1	1	1	1	1	1
8	1	1	1	0	0	0	*18*	1	1	1	1	1	1
9	1	1	1	1	0	0	*19*	1	0	0	0	0	0
10	1	1	1	1	1	0	*20*	1	1	1	1	1	1

Appendix B. Q_matrix by the divergent attribute structure

	Attributes							*Attributes*					
Items	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*	*Items*	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*
1	1	0	0	0	0	0	*11*	1	0	0	1	1	1
2	1	1	0	0	0	0	*12*	1	0	0	1	0	1
3	1	1	1	0	0	0	*13*	1	1	1	0	0	0
4	1	0	0	1	0	0	*14*	1	1	1	1	0	0
5	1	0	0	1	1	0	*15*	1	1	1	1	1	0
6	1	0	0	1	0	1	*16*	1	0	0	1	1	1
7	1	1	0	0	0	0	*17*	1	0	0	1	1	1
8	1	1	1	0	0	0	*18*	1	1	0	1	0	1
9	1	1	1	1	0	0	*19*	1	0	0	0	0	0
10	1	0	0	1	1	0	*20*	1	0	0	1	0	1

Appendix C. R Scripts written to compute HCI and FHCI

```
########### HCI #########
setwd("~/Desktop/FHCI/data")
data<-read.table("ResponseData.txt", header=F, sep="")
q<-read.table("Q_matrix.txt", header=F, sep="")
p<-matrix(NA,1,nrow(data)) # person sayisisi kadar (samplesize)
for(i in 1:nrow(data)){
J=nrow(q)
m<-matrix(NA,1,nrow(q)) # misfit: madde sayisisi kadar
nci<-matrix(NA,1,nrow(q)) # total number of comparison: madde sayisisi kadar
for(j in 1:J){
   c<-matrix(NA,1,nrow(q)) # comparison for item j
   for(l in 1:J){
   c[,l]<-ifelse(data[i,j]==1,(ifelse(sum(ifelse(q[j,]>=q[l,],1,0))==ncol(q),1,0)),0)}
   cj<- (sum(c)-(sum(data[i,]*c))) # number of misfit by item j
   m[,j]<-ifelse(data[i,j]==1,cj,0)
   nci[,j]<-sum(c)  # item j is compared with itselft, which should not be counted}
   HCIi<-1-(2*(sum(m)/(sum(nci)-sum(data[i,])+.000001)))  # .0001 is to avaid NaN result for 0
response vectors
   p[,i]<-HCIi}
plot(p[1,], xlab="Examinee", ylab="HCI")
```

```
########### FHCI #########
setwd("~/Desktop/FHCI/data")
data<-read.table("ResponseData.txt", header=F, sep="")
q<-read.table("Q_matrix.txt", header=F, sep="")
data1<-matrix(NA,nrow(data),nrow(q))
for(i in 1:nrow(data)) {
  for(j in 1:nrow(q)){
  data1[i,j]<-ifelse(data[i,j]==0,1,0)}}
 p<-matrix(NA,1,nrow(data)) # person sayisisi kadar (samplesize)
 for(i in 1:nrow(data)){
 J=nrow(q)
 m<-matrix(NA,1,nrow(q)) # misfit: madde sayisisi kadar
 nci<-matrix(NA,1,nrow(q)) # total number of comparison: madde sayisisi kadar
 m1<-matrix(NA,1,nrow(q)) # misfit: madde sayisisi kadar
 nci1<-matrix(NA,1,nrow(q)) # total number of comparison: madde sayisisi kadar
 for(j in 1:J){
   c<-matrix(NA,1,nrow(q)) # comparison for item j
   c1<-matrix(NA,1,nrow(q)) # comparison for item j
   for(l in 1:J){
   c[,l]<-ifelse(data[i,j]==1,(ifelse(sum(ifelse(q[j,]>=q[l,],1,0))==ncol(q),1,0)),0)
   c1[,l]<-ifelse(data1[i,j]==1,(ifelse(sum(ifelse(q[j,]<=q[l,],1,0))==ncol(q),1,0)),0)}
   cj<- (sum(c)-(   sum(data[i,]*c))) # number of misfit by item j
   m[,j]<-ifelse(data[i,j]==1,cj,0)
   nci[,j]<-sum(c)  # item j is compared with itselft, which should not be counted
   cj1<- (sum(c1)-(sum(data1[i,]*c1))) # number of misfit by item j
   m1[,j]<-ifelse(data1[i,j]==1,cj1,0)
   nci1[,j]<-sum(c1)  # item j is compared with itselft, which should not be counted}
HCIi<-1-(2*((sum(m)+sum(m1))/(sum(nci)-sum(data[i,])+sum(nci1)-sum(data1[i,])+.000001)))  #
.0001 is to avaid NaN result for 0 response vectors
p[,i]<-HCIi}
plot(p[1,], xlab="Examinee", ylab="FHCI", ylim=c(-1,1))
```