Quality & Quantity
https://doi.org/10.1007/s11135-019-00957-8

Statistical measurement of trees' similarity

Sahar Sabbaghan[1] · Cecil Eng Huang Chua[2] · Lesley A. Gardner[3]

Abstract
Diagnostic theories are fundamental to Information Systems practice and are represented in trees. One way of creating diagnostic trees is by employing independent experts to construct such trees and compare them. However, good measures of similarity to compare diagnostic trees have not been identified. This paper presents an analysis of the suitability of various measures of association to determine the similarity of two diagnostic trees using bootstrap simulations. We find that three measures of association, Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma (J Am Stat Assoc 49(268):732–764, 1954) each behave differently depending on what is inconsistent between the two trees thus providing both measures for assessing alignment between two trees developed by independent experts as well as identifying the causes of the differences.

Keywords Diagnostic theory · Tree · Threshold building

1 Introduction

Diagnostic theories are theories about the appropriate corrective action to take when given a set of observable conditions (Reiter 1987). They are fundamental to Information Systems (IS) practise (Webster and Watson 2002; Rooney and Van den Heuvel 2004; Clauset et al. 2008). For example, when an Information Technology (IT) system fails, it can fail for a myriad of reasons. The tree of symptoms and diagnoses associated with an IT system failure is a diagnostic theory. Similarly, when a new IT product is launched, there can be many reasons why users do not adopt it. Again, that tree of possible causes is a diagnostic theory. Indeed, many expert systems operate based

✉ Sahar Sabbaghan
sabbaghs@lsbu.ac.uk

Cecil Eng Huang Chua
cchua@mst.edu

Lesley A. Gardner
l.gardner@auckland.ac.nz

[1] Management, Marketing and People, London South Bank University, London, UK

[2] Business & Information Technology Department, College of Arts, Social Science and Business, Missouri University of Science & Technology, Rolla, MO, USA

[3] ISOM, University of Auckland, Auckland, New Zealand

Published online: 03 January 2020

on diagnostic theories. For instance, Mycin (Shortliffe 2012) and other expert systems navigate a decision tree to identify the root cause of a problem. Despite the fact that diagnostic theories are core to IS practise, little attention has been paid in IS research to diagnostic theories. Perhaps this is because diagnostic theories are not explanatory theories like variance or process theories (Webster and Watson 2002), but instead are prescriptive theories- they directly inform decision making. Nevertheless, like all theory, diagnostic theories need to be validated.

One way of creating diagnostic theories is to have an expert creating the theory. To validate the diagnostic theory, a second expert in the same area creates another diagnostic theory, and the two are compared. However, good measures for the correspondence of two diagnostic theories are essentially non-existent. This study aims to develop measures useful for comparing two diagnostic theories. Existing measures for trees such as edit-distance are not suitable for diagnostic theories because they are sample size dependent. An edit-distance of 20 is very bad when comparing two trees with 40 nodes each but is not so bad if the two trees have over 1000 nodes. Ratios of edit-distance (e.g., 10% of the tree are different) are also not suitable, because a lack of correspondence near the root of a diagnostic tree is a more severe issue than a lack of correspondence near the base of the tree-an idea a ratio does not capture.

To address our problem, we performed a set of bootstrap simulations to measure how various statistics change as a hypothetical diagnostic tree deviates from a “true” version. We apply traditional statistical measures in a new way to measure tree similarity. In particular, we transform the tree into a contingency table and employ traditional contingency table statistics to evaluate similarity. Our contribution is the discovery that three measures of association, Goodman and Kruskal’s Lambda (λ), Cohen’s Kappa (ƙ), and Goodman and Kruskal’s Gamma (γ) (Goodman and Kruskal 1954) together provide information useful for assessing the similarity of two diagnostic theories. Each of these three statistics behaves differently depending on what is inconsistent between the two trees thus providing both metrics for assessing alignment between two diagnostic theories developed by different experts as well as identifying the causes of the differences.

The paper is constructed in the following manner. First, we present the limitations of previous work. Then, we attempt to address those problems by providing a process for developing good thresholds for the construct validity of diagnostic tree and diagnosing their differences. We conclude with a discussion of diagnosing inter-rater reliability.

2 Diagnostic theory

A diagnostic theory is represented by a tree. For instance, Hopp et al. (2007) used a diagnostic tree for evaluating and improving production line performance. A diagnostic tree consists of a root, which corresponds to the problem domain (Geoffrion 1989). The root of the tree is unpacked to represent broad classes of diagnoses. As one traverses down the tree, the classes become narrower until we reach the tree’s base, where specific potential solutions are identified. For example, consider Fig. 1 which presents a diagnostic tree to identify why users have low Instagram self-efficacy—i.e., what is it about Instagram they find most hard to use? In this example, the top-level nodes

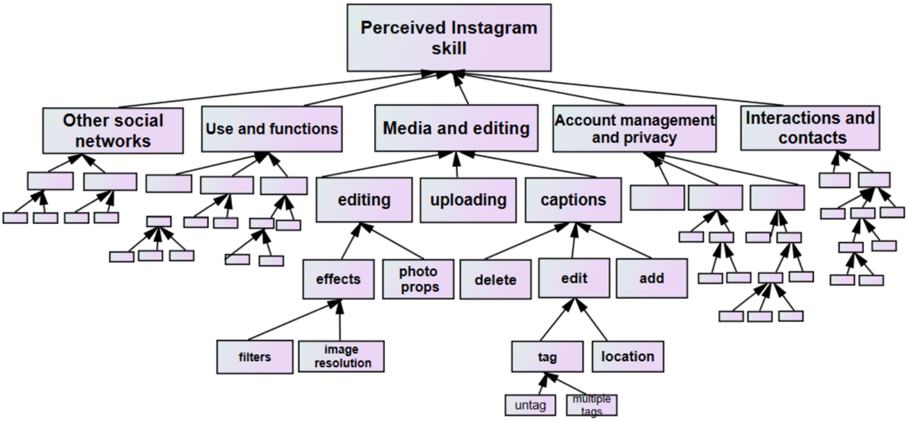


Fig. 1 Diagnostic tree example for perceived Instagram skill

encompass the different dimensions of Instagram skill. Each top-level node, in turn, links to more specific areas a user can experience problems in.

2.1 Comparison of trees

Diagnostic trees are typically built by experts and have certain properties. First, they can have hundreds of nodes, where nodes concerning higher-level concepts are mapped to nodes with greater precision. The nodes then have a parent–child relationship. Second, nodes higher up the tree are more important than those lower in the tree. Nodes lower in the tree are sub-nodes of those higher in the tree. This means that any errors or disagreement in the higher levels propagate to lower levels.

One way to validate diagnostic trees is to compare, the similarity of two diagnostic trees, created by two independent experts in the same domain. However, appropriate measures and "good enough" thresholds for demonstrating the similarity of two diagnostic trees are unknown. In addition, good measures for identifying problematic nodes in the tree are undiscovered. As an example, if two experts disagree on the mapping of two nodes, we would want to know whether the experts think that the nodes belong to different parents, or whether the experts disagree on the precision of the node in the hierarchy. In effect, measures akin to the modification indices of variance-based structural equation models need to be formulated (Gefen et al. 2000).

The remainder of this section reviews the principal existing methods of measuring tree similarity, which are edit-distance and statistics-based. We demonstrate the limitations of both methods and identify elements that can provide a foundation for creating a threshold for diagnostic trees.

2.1.1 Edit-distance based techniques

Edit-distance is a poor general comparator for diagnostic trees for several reasons. One is that existing algorithms do not take into account that nodes in the tree are not equally important (Jiang et al. 1995; Weinberg and Last 2017). To illustrate, consider Fig. 2, where

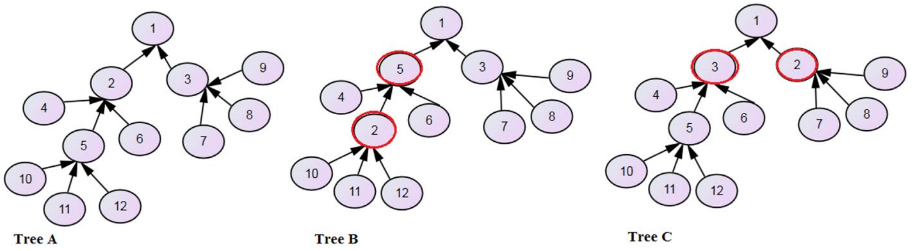


Fig. 2 Tree hierarchy limitation example

Trees B and C are each inconsistent with Tree A in exactly one way. Tree B swaps nodes 2 and 5, while Tree C swaps nodes 2 and 3. The typical edit-distance algorithm treats both inconsistencies equally. However, Tree B suggests the expert considered node 5 as a parent to node 2, while the expert for Tree B considered nodes 2 and 3 to be completely different nodes from the expert creating Tree A. The difference represented in Tree C is more serious than Tree B, as the expert (1) considered the nodes to effectively be two different nodes (as opposed to different levels in the same node), and (2) the issue occurred at a relatively high level in the tree, suggesting there are further problems lower in the tree that were undiscovered. In comparing trees, a technique that identifies such differences is necessary.

Also, edit-distance measures are often sample size sensitive. Clearly if there are two trees, each having 50 nodes, where 10 changes are required to transform one into the other, this is different from two trees, each having 500 nodes where only 10 changes are required. In statistical thinking, we want to compare the statistic to some probability distribution to standardize results according to “sample size”. We then calculate confidence intervals or *p* values of significance, where the threshold (typically 0.05 or 0.01) is sample size independent. The tree edit-distance literature has no equivalent analogue.

Finally, as a corollary to the above two points, we would like measures of tree similarity to systematically identify where the differences are between the trees. Edit-distance algorithms do this for individual nodes- they identify that to transform one tree into another, these are the nodes that must be changed and how (Grassi et al. 2015; Green and Ricca 2015). However, they do not, for example, tell us that most errors occur in the top of the tree (very bad) or at the bottom of the tree (not so serious)-or tell us that most of the errors are occurring in the children of node 1.

2.1.2 Statistics based techniques

Existing statistics-based methods are not suitable for several reasons. One is that existing measures and statistics are employed for generally “flat” question structures, and not the hierarchical structure of trees. For example, the traditional factor analytic concepts of convergent and divergent validity are assessed with correlations (Sartori 2006; Sartori and Pasini 2007; Hair et al. 1998). However, in diagnostic trees, nodes have a parent–child relationship. If the nodes behave correctly, the parent correlates highly with at least one child but is unlikely to correlate with all. For example, if a respondent answers that she is dissatisfied with food quality, then the respondent might be unhappy about the way the food was prepared but be satisfied with portion size. Factor loadings do not take this into account.

The inferential statistics tradition in several academic disciplines such as IS is to employ thresholds to evaluate whether two things are the same (Boudreau et al. 2001). For example, we regularly consider a *p* value under 0.05 to be "good enough". In cases where thresholds are unknown, research is done to identify them. As an example, Hu and Bentler (1999) examine the adequacy of the "rules of thumb" of conventional cutoff criteria and propose new alternatives for various fit indexes in structural equation models. However, such techniques have not been applied to trees.

In this study, we apply traditional statistical measures in a new way to measure tree similarity. Essentially, we map the tree into a contingency table and employ traditional contingency table statistics to evaluate similarity. The three measures used are Goodman and Lambda (λ), Cohen's Kappa (ƙ), and Goodman and Kruskal's Gamma (γ) (Goodman and Kruskal 1954). Goodman and Kruskal (1954, p. 749) interpret Lambda (λ) as "how much more probable it is to get like than unlike orders in the two classifications, when two individuals are chosen at random from the population." We chose Lambda (λ) because it has a meaning akin to r in a regression (Anderson and Gerbing 1988), i.e., Lambda (λ) is the measure of the strength of association in a contingency table (Everitt 1992; Goodman and Kruskal 1963). We chose Kappa (ƙ) because it is the observed proportion of agreement between the assigners after chance agreement is removed from consideration (Cohen 1968). Kappa (ƙ) is widely used as a measure of association for contingency tables (Hambleton and Zaal 2013; Rudick et al. 2013; Sengupta and Te'eni, D. 1993; You et al. 2012). In addition, Landis and Koch (1977) proposed the English-language meanings of Kappa (ƙ) thresholds featured in Table 1. We chose Gamma (γ) because it is explicitly designed for data with ordinal values (Higham and Higham 2019; Nelson 1984), and hierarchies are ordered data structures. Goodman and Kruskal (1954) interpret Gamma (γ) as how much more probable it is to get like than unlike orders in the two classifications, when two individuals are chosen at random from the population (Davis 1967; Göktaş and İşçi 2011; Goodman and Kruskal 1954). The value of the Gamma (γ) coefficient ranges from − 1 to + 1 where the latter value indicates perfect agreement between the two classifications (Baker 1974).

3 Foundation for threshold building

To build suitable thresholds for comparing and assessing diagnostic trees, we first generate a hypothetical "perfect" tree. We then make a copy of the tree and systematically change the tree and measure the statistic. We make a second change on the tree and measure the statistic again, repeating, the process many times to get a good appreciation

Table 1 Kappa (ƙ) interpretation

Kappa (ƙ) statistic	Strength of agreement
< 0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

for how statistics vary as two trees diverge. We then do the same with other "perfect" trees of various sizes.

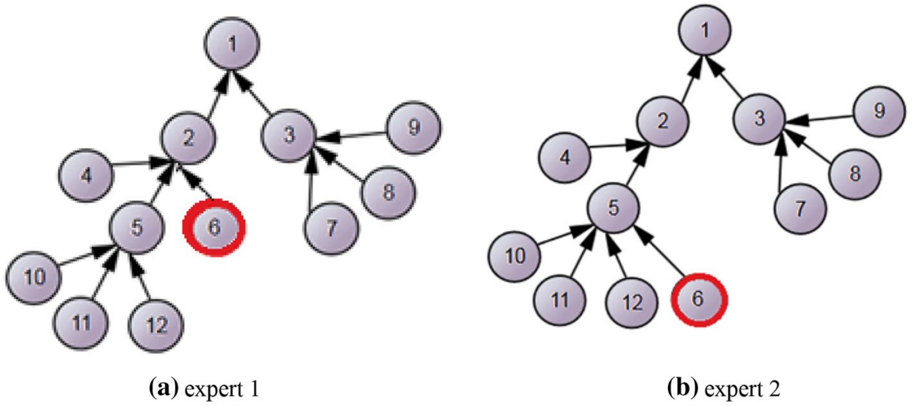
There is one constraint on modifications- each parent cannot have just one child node- with only one child node, there is no "branching". In the below, we formally define terms employed in the remainder of the paper.

Level is the distance of a node from the root. A node is on the $n+1$ level of its parent node. As an example, a node located on the 3rd level is three levels below the root node and its parent is on the 2nd level. Levels closer to the root are considered higher levels and levels further from the root are considered lower levels.
Root is the node with no parent. The root of the tree is on level 0.
Degree given two nodes a and b of level m and n such that a is the ancestor of b or b is the ancestor of a. The degree of the pair $d(a,b) = |m-n|$.
Descendant is the nth degree child of an ancestor node. As an example, *3rd* degree descendant of a node is located three levels below its ancestor node. A first-degree descendant of a node is also called a child node.
Top-level are first-degree descendants of the root. The top level of the tree has a level of 1.
Ancestor an ancestor is the nth degree parent of a descendant node where $n > 0$. As an example, the ancestor of *4*th degree is located four levels above its descendant node. A first-degree ancestor of a node is also called a parent node.
Relative Given two nodes, a and b, either a and b share an ancestor which is not the root, or a is an ancestor of b, or b is an ancestor of a.
Non-relative is any node whose only ancestor to another node is the root.
Modification given two trees, one of the differences between the two trees.
Movement Given a tree T with nodes labelled from *1* to n. A movement *M(a,b)* where a and b are nodes in T such that $0 \leq a \leq n$, $0 \leq b \leq n$, $a <> b$ and b has descendants and a is not a first-degree descendant of b, is defined as: T' such that a is the child of b, i.e., a is a first-degree descendant of b. There are three types of movements:

- *Type 1 movement* is a movement such that a in T is a childless node.
- *Type 2 movement* is a movement such that in T, a is a parent node. In T', all descendants of a in T become descendants of a's parent.
- *Type 3 movement with child(ren)* is a movement such that in T, a is a parent node. In T' all descendants of a in T have the same parents.

Direction of movement given two nodes a and b in tree T, a can move in three possible directions to become a child of b in T', (1) within relatives (up or down), (2) within its level (left or right), or (3) both within relatives and levels. Movements can occur with any kind of node (with or without descendant).

- *Hierarchy movement* is a direction movement *M(a,b)* where a and b are nodes in T and a is a relative of b. In T' a becomes a first-degree descendant of b. If b is descendant of a in T, then a hierarchy movement type 3 is not possible, because effectively, nothing happens to b. A hierarchy movement is effectively a movement up or down the tree. We care about hierarchy movements, because these suggest a certain type of error. In a diagnostic tree, the top-level nodes are unpacked and their descendants are mapped. This type of error indicates that experts disagree on the mapping of their direct relative nodes. As an example, consider Fig. 3 which presents two trees from experts 1 and 2. The experts disagree on the parent of node 6 as expert 1 has mapped node 6 to node 2, while expert 2 has mapped node 6 to node

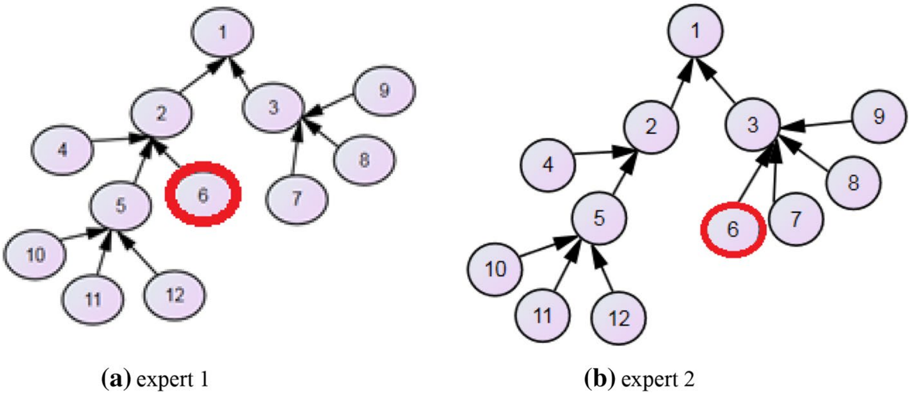


(a) expert 1 **(b)** expert 2

Fig. 3 Example of type 1 hierarchy movement

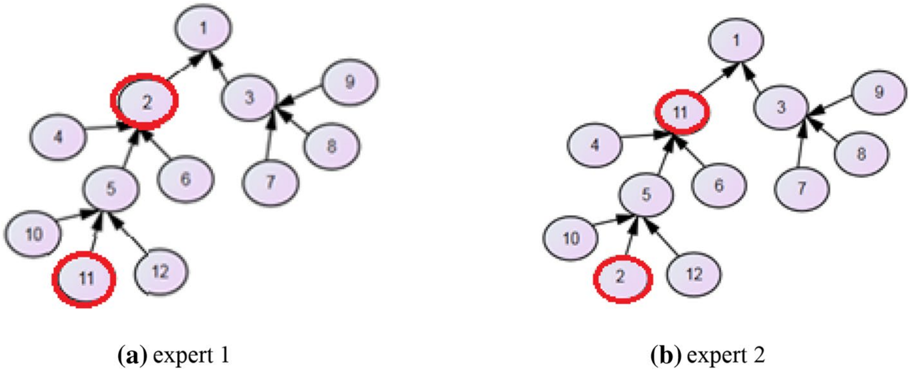
5. In addition, expert 2 sees node 6 on a lower level than expert 1, as expert 1 has mapped node 6 to node 2 which is on a higher level. In this example, as node 6 (*a*) moved to become a child of node 5 (*b*), and does not have any descendants, we call this a type 1 hierarchy movement.

- *Level movement* is a direction movement *M(a,b) a* is in the same level as a child of *b*, is defined as: *T*’ such that *a* is the child of *b*. We care about level movements, because these suggest that while experts agree on the level of the node, they disagree on the “family” of nodes the question relates to. As an example, consider Fig. 4a where in Tree 1, node 6 is on the same level as nodes 7, 8, and 9. Figure 4b presents level movement type 1 for node 6 (*a*) as it moved from node 2 to node 3 (*b*). The level of node 6 has not changed, however, experts disagree on the direct parent node. This indicates that the experts are confused between nodes 2 and 3.
- *Diagonal movement* is a direction movement *M(a,b*) that is both a hierarchy and level movement. Diagonal movements suggest two experts thought of a node in very different ways, as they disagree on both the level of the mapping and their direct relative nodes.



(a) expert 1 **(b)** expert 2

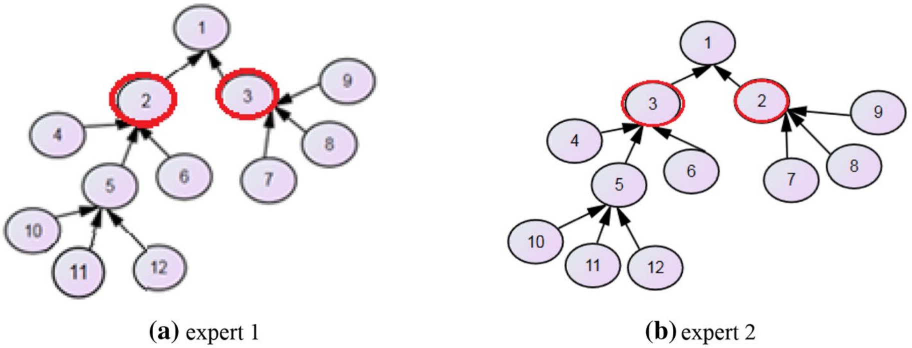
Fig. 4 Example of level movement type 1



(a) expert 1

(b) expert 2

Fig. 5 Example of hierarchy swap



(a) expert 1

(b) expert 2

Fig. 6 Example of level swap

- *Swap* a combination of two or more movements, which we treat as one. A swap is *S(a,b)* in *T*, where in *T*', *b* becomes the child of *a*'*s* parent and takes *a*'*s* children as descendants (if any), while *a* becomes the child of *b*'s parent and takes *b*'s children as descendants (if any). We consider swap distinctive from movements because this reflects a single cognitive difference between two experts rather than two or more cognitive differences. Similar to direction movements, there are three types of swaps, which are hierarchy, level, and diagonal.
- *Hierarchy swap* is where *a* is a relative of *b* in *T*. For example, consider Fig. 5, which presents a hierarchy swap between nodes 11 and 2. In Fig. 5b, node 11 is closer to the root (node 1), hence it becomes the ancestor of node 2. This shows that the experts disagree on the mapping of the direct relative nodes of nodes 2 and 11.
- *Level swap* is a swap where *a* and *b* in *T* are located in the same level and if both *a* and *b* do not have any descendants, they must not share a first-degree ancestor (direct parent) as tree T' will be the same as T. As an example, in Fig. 6, nodes 2 and 3 are on the same level and have been swapped. This shows that while the experts agree on the level of the nodes, they disagree on the mapping of the parent node.

- *Diagonal swap* is a swap where a and b in T are located on different levels and are not relatives. Diagonal swaps suggest the two experts are confused with two nodes in very different ways, as they disagree on both the level of the mapping and their direct relatives.

In addition, we want to perform analyses comparing the result of moving nodes at higher levels of the tree versus moving nodes at lower levels of the tree. Changing nodes at higher levels of the tree should have a greater impact, because this suggests problems with more important nodes. As an example, consider Fig. 7, which presents Trees A, B and C. In Trees A and B, experts disagree in the mapping of node 3 and in Trees A and C the experts disagree on the mapping of node 14. The disagreement between Trees A and B is more serious than the disagreement between Trees A and C.

To simulate these conditions, we perform analyses where we restrict the levels where movements and swaps occur. For every perfect tree with levels 0,…,n, we introduce a variable x where $2 < x < n$. Using the perfect tree as a base, we perform a set of swaps and movements between levels 1 and x. We then use the perfect tree as a base again, and perform a second set of swaps and movements between levels x and n and we compare the difference in scores. To distinguish the two, the swaps and movements performed between levels 1 and x are called movements and swaps on the "top" of the tree, and those between x and n as on the "bottom" of the tree.

3.1 Insertion and deletion

Finally, in some cases, one expert may not choose to map all pre-determined nodes and the two trees could have different numbers of nodes. Hence, we assess the impact of an insertion/deletion of a node in a tree. As deletion is the reverse of insertion, we only assess the impact of insertions. We consider two types of insertion as there are only two ways to insert a node to a tree, (1) insertion in levels where there is an increase in the number of branches per node and (2) insertion in a hierarchy where there is an increase in the number of levels.

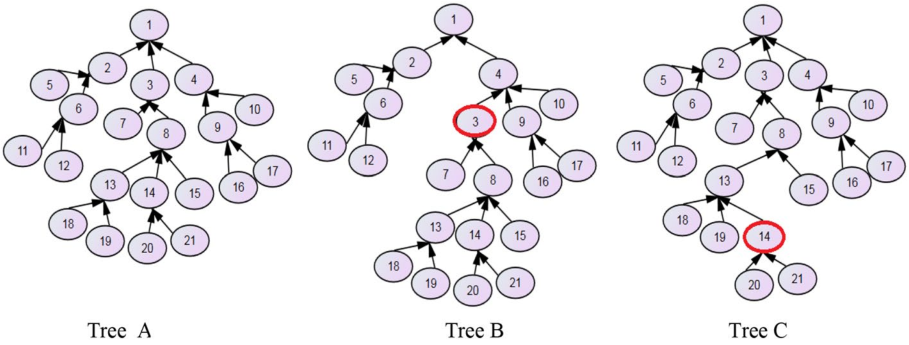


Fig. 7 Example of levels in diagnostic trees

4 Threshold building and diagnostic process

For the diagnosing process, we systematically identified all the ways a node can move in a tree hierarchy which has 1 to n- levels and 1 to m branches per node. We generate 12 (i.e., 3×4) perfect trees to test. Each tree has between three and five (i.e., three possibilities) branches per node and three to six levels (i.e., four possibilities) as tabulated in Table 3. We did not perform the simulation on trees with more than 200 nodes, because the computations required to simulate these trees become exponentially complex (Goldreich 2011). Our analysis of smaller trees suggests that statistics are similar regardless of the size of the tree. In addition, we drop the 3-branch 3-level tree as the number of nodes is too small to run a simulation for 100 rounds. Hence, six trees remain for the diagnosing process. These are identified as the bold cells in Table 2.

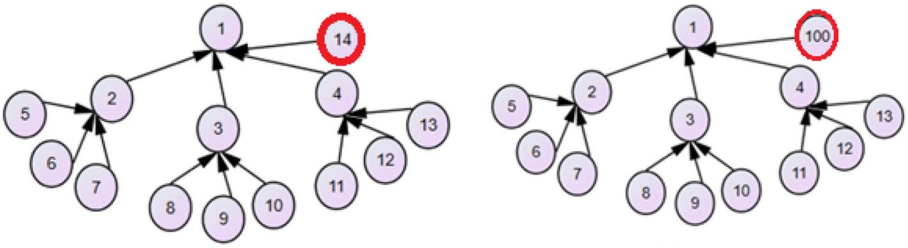
To diagnose each type of disagreement among the experts, each perfect tree is compared to a series of 27 possible modifications. Each modification is performed 100 times on each perfect tree. The total number of tests is therefore 16,200 (27×6×100). These modifications are:

- Nine possible direction movements comprising a combination of a movement type (Types 1–3) and direction (level, hierarchy, diagonal).
- Three possible movements where we keep the type constant, and allow random directions.
- Three possible movements where we keep the direction constant, and allow random types.
- Three possible swaps (level, hierarchy, diagonal).
- Eight top and bottom movements, where we restrict one half of a tree. Consider an example with tree *T* which has five levels. We first limit movements and swaps for only levels two and three and then for only levels four and five. It should be noted that by definition, the scores on the top half of the tree will change more than on the bottom half of the tree, given there are fewer nodes on the top half, and thus any change will have a greater effect. However, we wanted to know what the magnitude of the difference would be.
- One random movement/swap where a random change (either one of the 12 movements or 3 swaps) is performed. Each change is equally likely. The aim is to compare the results and evaluate how the statistics change and identify a suitable threshold.

Finally, for the insertion process, we assess our trees by first creating a perfect tree, *T*. Next, we make a copy of the tree, as *T*'. Then for each type of insertion, we randomly add one node to *T* and map it to a node. Contingency table analyses are unable to be performed to compare two different sample sizes. To address this, for every missing node in tree *T,* a node is represented in *T*' in the same location with a number not found in *T*'. We repeat this

Table 2 Number of nodes in each tree

Branches	Levels			
	3	4	5	6
3	13	**40**	**121**	364
4	**21**	**85**	339	1363
5	**30**	**156**	781	3906



(a) Tree T **(b)** Tree T'

Fig. 8 Level insertion for a 3-level 3-branch tree

20 times. As an example, consider a 3-level 3-branch tree as illustrated in Fig. 8, node 14 is in tree *T* and mapped to node 1. Node 14 does not exist in tree *T*', hence, for tree *T*', we insert the dummy node 100 to represent node 14 of tree *T*. As this increases the number of branches per node and not the number of levels, we consider this a level insertion.

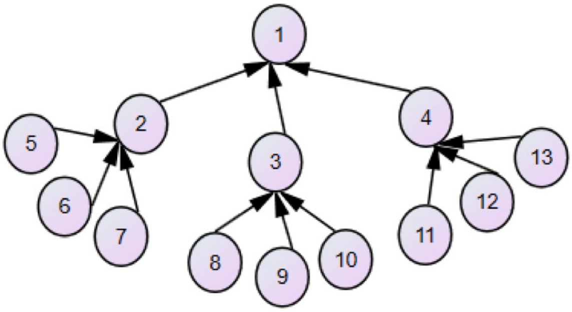
4.1 Data collection

Each variation of the tree is represented in a contingency table as follows. First, every node is given a number from *1* to *n*. 1 is the root node. The tree is then translated into a two-column table. The first column denotes the parent node, and the second denotes the child. Table 3 presents a 3-level 3-branch tree transformed into two columns. As seen, in Table 3, there are 12 child nodes and each parent node has 3 children. Each row represents a child and a parent. As an example, row 11 shows that child node 11 belongs to parent node 4.

4.2 Simulation analysis

The perfect tree is placed alongside the modified tree and a statistical comparison between the two is performed. Each pair of trees is compared on three statistics, Goodman and

Table 3 Transformed tree



Row	Child	Parent
1	1	0
2	2	1
3	3	1
4	4	1
5	5	2
6	6	2
7	7	2
8	8	3
9	9	3
10	10	3
11	11	4
12	12	4
13	13	4

Kruskal’s Lambda (λ), Cohen’s Kappa (ƙ), and Goodman and Kruskal’s Gamma (γ) (Goodman and Kruskal 1954). Recall that we analyse six possible perfect trees varying in number of levels and branches. Next, for the first 100 runs of each type of movement or swap, the six trees are transferred into a table, each child and parent is combined into an individual column and the means and standard deviations of the three statistics are calculated. In addition, the mean change (i.e., how much each statistic changes from one run to the next) and standard deviations of the mean change are calculated for the first 100 runs of each movement and swap (a total of 98 mean changes). Lambda (λ), Kappa (ƙ), and Gamma (γ) of the 100 rounds for six trees are recorded in each column and a paired sample t-test for each pair of the measures is calculated. Finally, for each type of insertion process, we calculate Lambda (λ), Kappa (ƙ), and Gamma (γ) of the 20 rounds.

5 Results

To build suitable thresholds for comparing and assessing diagnostic trees, we compare each of our hypothetical “perfect” trees to the modified tree and measured the statistic, repeating this process many times. Our results demonstrate Lambda (λ), Kappa (ƙ), and Gamma (γ) change at different rates depending on the kind of movement and swap performed. Table 4 presents a summary of these changes. There are several insights for each movement or swap, which we discuss below.

5.1 Movements

There are several insights for each direction or type of movement. Table 5 presents the means, standard deviations, and Cohen’s distance for the first 100 runs of each movement for the six trees. Cohen’s distance provides a measure of the strength of the difference in a t-test (Cohen 1988). In addition, Table 6 presents the mean changes in the measures for each directional movement.

Results indicate that for all hierarchy movement types, Gamma (γ) decreases more dramatically than the other two measures. In addition, the mean for Gamma (γ) is lower than the other two measures throughout all types of hierarchy movements. As an example, in a 4-level 3-branch diagnostic tree as illustrated in Fig. 9a, Gamma (γ) in run 20 drops from 0.978 to 0.683 in hierarchy movements while Kappa (ƙ) drops from 0.966 to 0.839. Paired sample t-tests between Gamma (γ) and Kappa (ƙ) (the next lowest measure) are all statistically significant.

In level movements, results indicate the mean for Kappa (ƙ) for the six trees is lower than the other two measures. In addition, Kappa (ƙ) decreases at the fastest rate of all three measures. All changes are statistically significant when Kappa (ƙ) is compared to Gamma (γ), the next lowest measure. As an example, in a 4-level 3-branch diagnostic tree as presented in Fig. 9b, the mean change for Kappa (ƙ) is 0.0036, while Gamma (γ) is only 0.0016 in level movement. In addition, in level movements, for a 4-level 3-branch diagnostic tree, Kappa (ƙ) in run 20, drops from 0.991 to 0.635, while Gamma (γ) drops from 0.999 to 0.761.

In diagonal movements, Lambda (λ) decreases at a faster rate than for any other movement as shown in Table 6. As an example, in a 4-level 3-branch diagnostic tree, Lambda (λ) in diagonal movements, in run 20, drops from 0.982 to 0.77, while in level movement it drops from 0.9871 to 0.8423 and in hierarchy movements it drops from 0.991 to 0.866.

Table 4 Summary of changes of Lambda (λ), Kappa (ƙ), and Gamma (γ) for different types of modifications

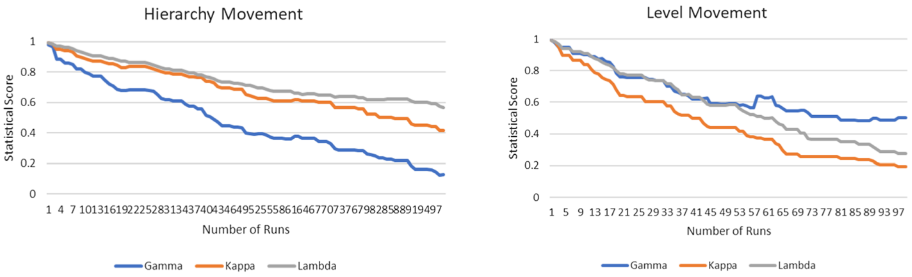
Type of modification	Description	Changes in Lambda (λ), Kappa (ƙ), and Gamma (γ)
Movement types	Modifications that impact descendants	Lambda (λ) decreases at a faster rate in type 2 than the other types
Swaps	Modifications that impact the mapping of two nodes	Lambda (λ) does not change
Hierarchy movements and swaps	Modifications that impact the number of levels in a tree	Gamma (γ) decreases the fastest
Level movements and swaps	Modifications that impact the number of branches in a tree	Kappa (ƙ) decreases the fastest
Diagonal movements	Modifications that impact the number of both levels and branches in a tree	Lambda (λ) decreases the fastest
Top and bottom movements and swaps	Modifications that impact higher levels versus lower levels	Lambda (λ), Kappa (ƙ), and Gamma (γ) decrease faster in top levels than bottom levels
Insertions	Modifications that increase the number of nodes in either branches per node or in the number of levels	Behaviour of Kappa (ƙ), and Gamma (γ) is similar to level and hierarchy movements

Table 5 Means, standard deviation and Cohen's distance for the different movements

Type of change	Lambda (λ) mean (SD)	Kappa (ꝁ) mean (SD)	Gamma (γ) mean (SD)	Cohen's distance for Kappa (ꝁ) and Gamma (γ)	Paired samples t-test for Kappa (ꝁ) and Gamma (γ) (*df* = 599)	Sig. (2-tailed)
Hierarchy movement	0.570 (0.125)	0.432 (0.294)	0.2005 (0.472)	0.590	− 27.947	< 0.001
Hierarchy movement Type 1	0.680 (0.014)	0.723 (0.192)	0.57 (0.189)	0.378	− 18.046	< 0.001
Hierarchy movement Type 2	0.555 (0.240)	0.411 (0.258)	0.108 (0.461)	0.810	− 11.985	< 0.001
Hierarchy movement Type 3	0.583 (0.250)	0.388 (0.25)	0.352 (0.242)	− 0.193	− 4.049	< 0.001
Level movement	0.479 (0.283)	0.354 (0.26)	0.522 (0.275)	− 0.615	18.97169	< 0.001
Level Movement Type 1	0.897 (0.388)	0.874 (0.059)	0.922 (0.0057)	0.826	35.404	< 0.001
Level movement Type 2	0.356 (0.307)	0.304 (0.306)	0.779 (0.2603)	1.674	29.416	< 0.001
Level movement Type 3	0.671 (0.165)	0.525 (0.25)	0.699 (0.17)	− 0.784	16.89	< 0.001
Diagonal movement	0.415 (0.251)	0.399 (0.267)	0.358 (0.248)	0.159	− 5.557	< 0.001
Diagonal movement Type 1	0.628 (0.112)	0.698 (0.136)	0.744 (0.118)	− 0.361	− 11.56	< 0.001
Diagonal movement Type 2	0.479 (0.250)	0.369 (0.272)	0.351 (0.234)	0.094	31.238	< 0.001
Diagonal movement Type 3	0.500 (0.234)	0.435 (0.293)	0.382 (0.263)	0.183	− 10.061	< 0.001
Movements	0.512 (0.228)	0.441 (0.268)	0.4125 (0.254)	0.109	− 2.639	0.009
Type 1 movement	0.743 (0.117)	0.428 (0.284)	0.414 (0.264)	0.051	− 8.230	< 0.001
Type 2 movement	0.448 (0.246)	0.371 (0.267)	0.337 (0.287)	0.123	− 5.867	< 0.001
Type 3 movement	0.568 (0.215)	0.529 (0.275)	0.553 (0.221)	− 0.099	5.172	< 0.001

Table 6 Mean changes in different directional movements

Movement/measure	Mean change for Lambda (λ)	Mean change for Kappa (ƙ)	Mean change for Gamma (γ)
Hierarchy movement	0.000647	0.00088	0.00138
Level movement	0.000660	0.000697	0.000067
Diagonal movement	0.001688	0.001667	0.00163



(a) Hierarchy Movement **(b)** Level Movement

Fig. 9 The difference between Kappa (ƙ) and Gamma (γ) in level and hierarchy movements for a 4-level 3-branch tree

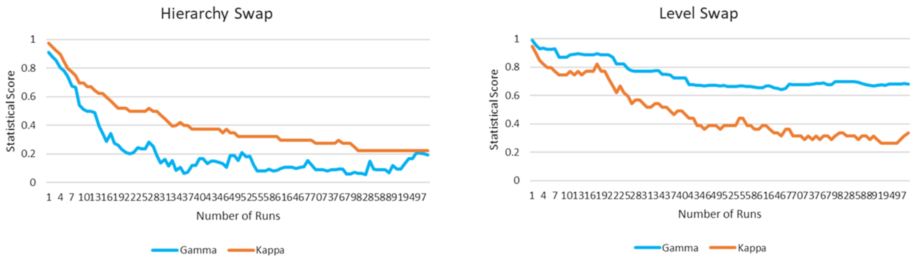
We ran a paired sample t-test on six different diagnostic trees to compare the raw scores of Lambda (λ) with the next lowest measure (Kappa (ƙ) or Gamma (γ)) for different diagnostic trees. The results for each pair was significant which indicates that the measures change at different rates.

Finally, as shown in Table 5 in type movements, Lambda (λ) is more sensitive to type 2 movements; the mean for Lambda (λ) is lower compared to other movement types (1 and 3). Type 2 movements consist of two steps, (1) a move of the parent node and, (2) a move of the child nodes to the former parent's parent node. These two steps have a bigger impact on Lambda (λ) than other measures, as more than one node is impacted.

5.2 Swaps

Our insights, which are shown in Tables 7 and 8 concerning swaps are as follows:

- Lambda (λ) does not change in swaps, as both the mean and mean change are zero.
- For hierarchy swaps, the mean of Gamma (γ) is lower than Kappa (ƙ), and mean changes for Gamma (γ) are higher than the mean difference for Kappa (ƙ), which indicates that Gamma (γ) drops faster than Kappa (ƙ). In the example shown in Fig. 10a, in a 4-level 3-branch tree, Gamma (γ) in run 20, for hierarchy swap drops from 0.9092 to 0.231, while Kappa (ƙ) drops from 0.974 to 0.520. The difference between Kappa (ƙ) and Gamma (γ) is statistically significant. This is consistent with Kappa (ƙ) and Gamma's (γ) behaviour for hierarchy movements.
- In level swaps, Kappa (ƙ) tends to decrease faster than Gamma (γ) as the mean change of Kappa (ƙ) is higher than Gamma (γ). As an example, as shown in Fig. 10b, in a 4-level 3-branch tree, Kappa (ƙ) drops from 0.949 to 0.72 while Gamma (γ) drops from 0.992



(a) Hierarchy Swap

(b) Level Swap

Fig. 10 The difference between Kappa (ƙ) and Gamma (γ) in level and hierarchy swaps for a 4-level 3-branch tree

to 0.889 in level swaps. The difference between Kappa (ƙ) and Gamma (γ) is statistically significant. This is consistent with Kappa (ƙ) and Gamma's (γ) behaviour for level movements.

5.3 Top and bottom movements and swaps

Given the sample size in a top movement/swap is always smaller than in the equivalent bottom movement/swap, our results unsurprisingly indicated that the measures decrease at a faster rate in top movements and swaps than bottom movements and swaps. As an example, Table 9 demonstrates the mean changes and standard deviations of the top, bottom, and general hierarchy and level movements and swaps for a 4-level 3-branch diagnostic tree. As presented in Table 9, Gamma (γ) decreases the fastest in top hierarchy movements and swaps, as the mean change is higher. In contrast, Kappa (ƙ) decreases the fastest in top-level movements and swaps.

5.4 Insertion process in diagnostic trees

Depending on the type of insertion, Lambda (λ), Kappa (ƙ), and Gamma (γ) change differently. Consider Table 10 which presents the results of level and hierarchy insertion for a 3-level tree with 3-5 branches. In total, 20 nodes were added to tree *T* and *T*'. In insertion to levels, Kappa (ƙ) is lower than Gamma (γ), while in insertion to hierarchy, Gamma (γ) is lower than Kappa (ƙ). Lambda (λ) drops faster in insertion to hierarchy than to level. The difference between Kappa (ƙ) and Gamma (γ) is statistically significant. This is consistent with Kappa (ƙ) and Gamma's (γ) behaviour for hierarchy and level movements and swaps. In all three cases, Gamma (γ) is lower than Kappa (ƙ) in hierarchy changes, and Kappa (ƙ) is lower than Gamma (γ) in level changes.

Table 7 Mean change and standard deviations for the first 100 runs of each swap for the 6 trees

Type of change	Lambda (λ) mean (SD)	Kappa (ƙ) mean (SD)	Gamma (γ) mean (SD)	Cohen's distance for Kappa (ƙ) and Gamma (γ)	Paired samples t-test for Kappa (ƙ) and Gamma (γ) (df = 599)	Sig. (2-tailed)
Hierarchy swap	1 (0)	0.280 (0.22)	0.213 (0.27)	0.191	− 5.423	< 0.001
Level swap	1 (0)	0.30 (0.23)	0.549 (0.21)	−0.75	35.661	< 0.001
Diagonal swap	1 (0)	0.184 (0.24)	0.138 (0.302)	0.124	− 7.193	< 0.001

Table 8 Mean and standard deviations for the first 100 runs of each swap for the six trees

Type of change	Mean difference and (SD) for Lambda (λ)	Mean difference and (SD) for Kappa (ꞵ)	Mean difference and (SD) for Gamma (γ)
Hierarchy swap	0 (0)	0.003 (0.023)	0.007 (0.035)
Level swap	0 (0)	0.004 (0.029)	0.002 (0.013)
Diagonal swap	0 (0)	0.014 (0.16)	0.015 (0.08)

Table 9 Mean difference and standard deviations for top, and bottom, and general hierarchy and level swaps and movement for a 4-level 3-branch diagnostic tree in 100 runs

Type of hierarchy swap/movement	Mean difference and (SD) for Lambda (λ)	Mean difference and (SD) for Kappa (ꞵ)	Mean difference and (SD) for Gamma (γ)
Top hierarchy swap	–	0.0072 (0.024)	0.0074 (0.043)
Bottom hierarchy swap	–	0.0050 (0.019)	0.0023 (0.02)
Hierarchy swap	–	0.0090 (0.012)	0.0072 (0.03)
Top hierarchy movement	0.0030 (0.005)	0.0050 (0.12)	0.0083 (0.016)
Bottom hierarchy movement	0.0020 (0.004)	0.00419 (0.03)	0.0001 (0.002)
Hierarchy movement	0.0040 (0.004)	0.0050 (0.09)	0.0080 (0.013)
Top level swap	–	0.0050 (0.019)	0.0020 (0.019)
Bottom level swap	–	0.0005 (0.0002)	0.0002 (0.00025)
Level swap	–	0.006 (0.025)	0.0030 (0.012)
Top level movement	0.0042 (0.005)	0.0066 (0.0087)	0.0030 (0.004)
Bottom level movement	0.0034 (0.004)	0.003 (0.004)	0.0012 (0.005)
Level movement	0.0053 (0.006)	0.006 (0.009)	0.0040 (0.012)

6 Threshold properties for empirical use

Many academic disciplines employ threshold values for “satisfactory” levels of inter-rater reliability. For example, the typical threshold for both Cronbach’s alpha and Cohen’s Kappa (ꞵ) is 0.7(Nunnally 1978; Watkins and Pacheco 2000). We believe suitable thresholds for comparing two diagnostic trees are when Lambda (λ) > 0.7, Kappa (ꞵ) > 0.4 and Gamma (γ) > 0.3. These thresholds are established for the following reasons:

- It is important to consider all three measures, because each measure signals different kinds of issues. Changes in Lambda (λ) signify movements are occurring, changes in Gamma (γ) suggest hierarchical inconsistencies, while changes in Kappa (ꞵ) suggest level inconsistencies.
- The three thresholds combined suggest that regardless of sample size, two trees that score above threshold differ in no more than 30% of their nodes. This has been assessed by testing the thresholds in random movements and swaps.

The question remains as to what happens and how efficient the measures are if, (1) only two of the thresholds were used in the comparison of all three thresholds and (2) small changes are made to the thresholds. Table 11 presents a comparison summary where only

Table 10 The results of level and hierarchy insertion for a 3-level 3–5 branch tree

Number of branches	Type of insertion	Lambda (λ) mean (SD)	Kappa (ƙ) mean (SD)	Gamma (γ) mean (SD)	Paired samples t-test for Kappa (ƙ) and Gamma (γ) (*df* = 599)	Sig. (2-tailed)
3-Branch	Insertion to hierarchy	0.746 (0.09)	0.534 (0.17)	0.419 (0.15)	− 15.13	< 0.001
3-Branch	Insertion to levels	0.975 (0.009)	0.56 (0.15)	0.630 (0.24)	3.2	0.00467
4-Branch	Insertion to hierarchy	0.813 (0.08)	0.650 (14)	0.440 (18)	− 25.7365	< 0.001
4-Branch	Insertion to levels	0.978 (0.006)	0.66 (0.14)	0.790 (0.15)	16.37	< 0.001
5-Branch	Insertion to hierarchy	0.873 (0.02)	0.883 (0.09)	0.753 (0.12)	− 12.84	< 0.001
5-Branch	Insertion to levels	0.980 (0.003)	0.740 (0.11)	0.860 (0.103)	18.30692	< 0.001

Table 11 A combination of thresholds for 3-branch trees of 3–6 levels

Number of levels	Number of nodes	$\lambda > 0.7$ and ƙ > 0.4	Percentage	$\lambda > 0.7$ and $\gamma > 0.3$	Percentage	ƙ > 0.4 and $\gamma > 0.3$	Percentage	$\lambda > 0.7$ and ƙ > 0.4 and $\gamma > 0.3$	Percentage
3 Levels	13	3	23.07	3	23.07	6	46.15	3	23.07
4 Levels	40	24	60	16	40	16	40	11	27.50
5 Levels	121	49	40.49	38	31.40	38	31.4	35	28.92
6 Levels	364	120	32.96	115	31.59	111	30.49	100	28.29
SD			13.50		5.90		6.4		2.29

Table 12 The impact of thresholds with small changes 3-branch trees of 3–6 levels

Thresholds	Number of levels	3 Levels	4 Levels	5 Levels	6 Levels	SD
	Number of nodes in tree	13	40	121	364	
$\lambda > 0.7$, $ƙ > 0.4$, $\gamma > 0.3$	Modifications	3	12	35	104	
	Percentage	23.07	30	28.92	28.5	2.6
$\lambda > 0.6$, $ƙ > 0.4$, $\gamma > 0.3$	Modifications	6	17	50	111	
	Percentage	46.15	42.5	41.32	30.49	5.83
$\lambda > 0.8$, $ƙ > 0.4$, $\gamma > 0.3$	Modifications	3	15	36	104	
	Percentage	23.07	37.5	29.75	28.57	5.14
$\lambda > 0.7$, $ƙ > 0.3$, $\gamma > 0.3$	Modifications	3	23	51	115	
	Percentage	23.07	57.5	42.14	31.59	12.84
$\lambda > 0.7$, $ƙ > 0.5$, $\gamma > 0.3$	Modifications	3	13	31	102	
	Percentage	23.076	32.5	25.61	28.02	3.47
$\lambda > 0.7$, $ƙ > 0.4$, $\gamma > 0.2$	Modifications	3	24	49	112	
	Percentage	23.07	60	40.49	30.76	13.82
$\lambda > 0.7$, $ƙ > 0.4$, $\gamma > 0.4$	Modifications	3	8	27	83	
	Percentage	23.07	20	22.31	22.8	1.21

two of the three thresholds are used. If only two thresholds are applied, it is possible for two trees to meet the thresholds when they have substantial differences from each other. For example, if only $\lambda > 0.7$ and $ƙ > 0.4$, then it is possible for our 4-level trees to differ by up to 60% of nodes.

Table 12 presents what other thresholds mean when comparing two trees. As an example, consider a 0.1 change of Lambda (λ) from 0.7 to either Lambda $(\lambda) > 0.6$ or Lambda $(\lambda) > 0.8$ while holding Kappa $(ƙ) > 0.4$ and Gamma $(\gamma) > 0.3$. We count the number of modifications for each tree and calculate the percentages. Table 12 presents the results of the impact of such 0.1 sized changes with each threshold and the standard deviation of the percentages demonstrates the accuracy of the thresholds for identifying the estimates.

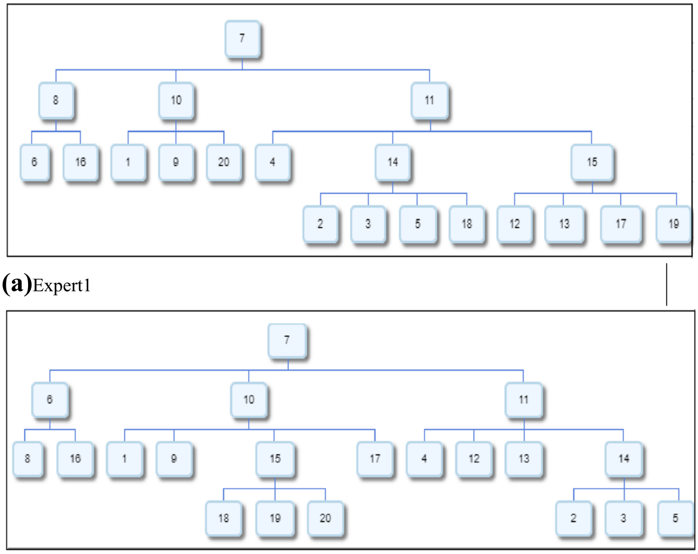
Results indicated that Kappa (ƙ) and Gamma (γ) are especially sensitive to changes to their threshold values. As an example, when only Gamma (γ) drops from 0.3 to 0.2, the standard deviation for the percentage of modifications is 13.82 while an increase from 0.3 to 0.4, the standard deviation is 1.21. However, when lambda (λ) drops from 0.7 to 0.6, the standard deviation of the percentage of modifications is 5.83 and with an increase from 0.7 to 0.8 the standard deviation is 5.14. There are several reasons. Firstly, the thresholds are tested in randomised movements and swaps, as Lambda (λ) does not change in swaps, hence small changes to Lambda (λ) would be less dramatic. Secondly, in random movements either or both levels and the hierarchy of nodes are affected, which makes each measure more sensitive to small changes, as each measure not only changes with both movements and swaps but changes more dramatically in swaps. Hence, Kappa (ƙ) and Gamma (γ) must be simultaneously adjusted to find suitable thresholds.

In addition, different combinations of measures can identify different levels of modification between two trees. Table 13 presents four thresholds for when the percentage of modifications are at 15, 20, 25, and 30% between two trees. As an example, a threshold of $\lambda > 0.75$, $ƙ > 0.5$, and $\gamma > 0.4$ can identify an estimate of 25% of modifications between two diagnostic trees, while a threshold of $\lambda > 0.85$, $ƙ > 0.7$, $\gamma > 0.5$ is suitable

Table 13 Thresholds according to different amounts of modifications for 3-branch trees of 3–6 levels

Thresholds	Number of levels	3 Levels	4 Levels	5 Levels	6 Levels	SD
	Number of nodes in tree	13	40	121	364	
$\lambda > 0.85$, ƙ > 0.7, $\gamma > 0.5$	Modifications	2	5	22	53	
15%	Percentages	15.38	12.5	18.1	14.5	2.03
$\lambda > 0.8$, ƙ > 0.65, $\gamma > 0.45$	Modifications	2	8	25	70	
20%	Percentages	15.38	20	20.6	19.2	2.04
$\lambda > 0.75$, ƙ > 0.5, $\gamma > 0.4$	Modifications	3	9	27	87	
25%	Percentage	23.07	22.5	22.3	23.9	0.61
$\lambda > 0.7$, ƙ > 0.4, $\gamma > 0.3$	Modifications	3	12	35	104	
30%	Percentage	23.07	30	28.92	28.5	2.6
$\lambda > 0.65$, ƙ > 0.4, $\gamma > 0.25$	Modifications	4	15	47	121	
35%	Percentage	30.76	37.5	38.84	33.241	3.23
$\lambda > 0.65$, ƙ > 0.35, $\gamma > 0.25$	Modifications	6	17	53	133	
40%	Percentage	46.15	42.5	43.8	36.53	3.54

Table 14 Trees and transformed trees of the node "Other Social Networks"

<table>
<tr><th>Nu
m</th><th>expert
1</th><th>expert
2</th><td rowspan="21">

(a)Expert1
(b)Expert2</td></tr>
<tr><td>1</td><td>10</td><td>10</td></tr>
<tr><td>2</td><td>14</td><td>14</td></tr>
<tr><td>3</td><td>14</td><td>14</td></tr>
<tr><td>4</td><td>11</td><td>11</td></tr>
<tr><td>5</td><td>14</td><td>14</td></tr>
<tr><td>6</td><td>8</td><td>7</td></tr>
<tr><td>7</td><td>0</td><td>0</td></tr>
<tr><td>8</td><td>7</td><td>6</td></tr>
<tr><td>9</td><td>10</td><td>10</td></tr>
<tr><td>10</td><td>7</td><td>7</td></tr>
<tr><td>11</td><td>7</td><td>7</td></tr>
<tr><td>12</td><td>15</td><td>11</td></tr>
<tr><td>13</td><td>15</td><td>11</td></tr>
<tr><td>14</td><td>11</td><td>11</td></tr>
<tr><td>15</td><td>11</td><td>10</td></tr>
<tr><td>16</td><td>8</td><td>6</td></tr>
<tr><td>17</td><td>15</td><td>10</td></tr>
<tr><td>18</td><td>14</td><td>15</td></tr>
<tr><td>19</td><td>15</td><td>15</td></tr>
<tr><td>20</td><td>10</td><td>15</td></tr>
</table>

Table 15 Results of the measures for the node "Other Social Networks"

Node	(λ)	(ꝁ)	(γ)
Social networks	0.531	0.474	0.673

for identifying an estimate of 15% of modifications of two trees. In addition, Table 13 presents less strict thresholds such as a λ > 0.65, ꝁ > 0.35, and γ > 0.25 which can identify an estimate of 40% of modifications between two diagnostic trees.

6.1 An example of assessing the similarity of diagnostic trees

As an example, consider a top-level node "Other social networks" from a perceived Instagram skill diagnostic tree presented in Fig. 1. Table 14 presents the two trees created by the experts (expert 1 and 2) and its transformation to tables and Table 15 presents the initial results of the measures. We have set the thresholds at 30% which suggests there is no more than 30% modification across the two trees. The actual scores for the measures are 0.531 for Lambda (λ), 0.474 for Kappa (ꝁ), and 0.673 for Gamma (γ). The measures provide several insights. Firstly, the trees are not similar enough, and the problematic nodes will need to be edited accordingly. Secondly, Kappa (ꝁ) being the lowest measure suggests the principal problem is the number of disagreements of mapping of nodes of the same level. Comparing across trees, we can see that the experts disagree on the parent nodes of nodes 12, 13, 14, 15, 17, 18, and 20 which are all located on the same level. Assume we correct this problem so that experts agree on the mapping of those nodes, the statistics become 0.85 for Lambda (λ), 0.76 for Kappa (ꝁ), and 0.76 for Gamma (γ) which indicates a strong inter-rater agreement.

At present, the only alternative to employing our measures is the use of edit-distance algorithms. As previously mentioned, these algorithms are neither sensitive to sample size nor to the various kinds of differences that can occur in two trees.

To illustrate, in our "Other social networks" Instagram efficacy instrument, the edit-distance of the two trees would have been 12 or 6% (i.e., edit-distance/total number of nodes). Observe that while this provides some measure of the non-correspondence between the two trees, it doesn't provide any useful diagnostic information. Furthermore, the reported level of difference- 6% does not appear too severe. In contrast, our statistical measures identified a systematic (level) difference across the two trees. If we were to correct the errors across nodes 12-18 and 20, the edit- distance jumps to 6 or 3%.

Contrast this example against another hypothetical one where we had the same number of nodes, but the problem was with the mapping of the nodes of the same level of the trees, as the experts disagree with the mapping of the parent/child nodes.

Edit-distance provides exactly the same statistics, but our measures provide additional information as each behaves differently based on the types of modifications occurred in the trees. Thus, Kappa (ꝁ) would decrease faster in changes within the same level, such as after 20 modifications Kappa (ꝁ) would drop from 0.991 to 0.635. Thus, as can be seen, our measures provide substantially more information than edit-distance and allow us to identify and target the principal problem first. Fixing the principal problem allowing us to quickly achieve satisfactory inter-rater agreement.

Table 16 Thresholds for different amounts of modifications for 3-level trees with 3–10 number of branches per node

Thresholds	Number of branches	B4	B5	B6	B7	B8	B9	B10	SD
	Number of nodes in tree	21	31	43	57	73	91	111	
λ > 0.85, ƙ > 0.7, γ > 0.5	Modifications	1	3	6	7	8	8	9	
15%	Percentages	4.7	9.6	13.9	12.2	10.95	8.79	8.1	2.77
λ > 0.8, ƙ > 0.65, γ > 0.45	Modifications	2	5	7	11	9	9	11	
20%	Percentages	9.5	16.1	16.2	19.3	12.32	9.89	9.9	3.6
λ > 0.75, ƙ > 0.5, γ > 0.4	Modifications	3	6	9	12	11	12	14	
25%	Percentage	14.2	19.3	20.9	21.05	15.06	13.18	12.61	3.41
λ > 0.7, ƙ > 0.4, γ > 0.3	Modifications	6	6	10	16	11	12	14	
30%	Percentage	28.5	19.3	23.2	28.07	15.06	13.18	12.61	6.26

7 Limitations and conclusion

Our analysis reveals several limitations with using Lambda (λ), Gamma (γ), and Kappa (ƙ) as measures of trees. First, the thresholds are inapplicable once the number of branches is greater than seven. To demonstrate this limitation, consider different thresholds for 3-level trees with 3-10 branches per node as presented in Table 16. Once there are eight or more branches, the measures are less effective for providing a threshold. As an example, in a 3-level 9-branch tree, the thresholds misrepresent the number of modifications, such as when the thresholds are set to identify 20% of the modifications, they only identify 10%, hence underestimating the number of modifications.

In addition, similar to other studies (van der Ark and van Aert 2015), we found Gamma (γ) too unstable to provide a reasonable threshold for small samples sizes such as trees with a total number of nodes below 25. However, for trees with a total number of nodes above 25, Gamma (γ) appears more stable. The point of trees is to facilitate choice between hundreds of options. Thus, for the purposes of assessing trees' similarity, Gamma (γ) remains a reasonable measure.

Furthermore, due to the exponentially complex computations required, we were unable to run simulations of trees with a total number of nodes above 200, hence could not make any conclusions. However, in our analysis, the measures have been fairly consistent as the growth of trees has been linear as the number of nodes per tree increased. Thus, the threshold results will most likely stay the same in trees with a total number of nodes above 200.

7.1 Conclusion

This study presents an analysis of the use Lambda (λ), Gamma (γ), and Kappa (ƙ) as measures of the similarity of diagnostic trees and tools for diagnosing their differences. To build suitable thresholds for comparing and assessing diagnostic trees, we first generated a hypothetical "perfect" tree. We then made a copy of the tree and systematically modified the tree. We created two general types of modifications, movements and swaps.

We repeated the modifications many times and did this for other "perfect" trees of various sizes. We found that:

- Gamma (γ) is useful for identifying disagreements with the hierarchy of the nodes.
- Kappa (ꝃ) is useful for identifying disagreements on the mapping of nodes of the same level.
- Lambda (λ) is useful for determining two things. The first is whether the principal problem is disagreement among single nodes (type 2 movements), which indicates that while experts agree on the grouping of child nodes, they disagree on the parent of the child nodes. The second is that a high Lambda (λ) concurrent with a low Kappa (ꝃ) or Gamma (γ) is useful to detect swaps.

We then proposed thresholds for various levels of inter-rater reliability, as an example, a threshold for when Lambda (λ) > 0.7, Kappa (ꝃ) > 0.4, and Gamma (γ) > 0.3, suggests there is no more than 30% modification between two trees.

This work is particularly useful for assessing the node and content validity of two diagnostic trees. As future research, we hope to explore and evaluate diagnostic trees in several areas. One, very little research has been done on measuring other types of validities for diagnostic trees. For example, we do not yet have clear techniques for assessing the nomological validity of diagnostic trees. Two, we intend to compare several popular measures used to compare trees with our statistical method to further demonstrate the use of this study's method.

References

Anderson, J., Gerbing, D.: Structural equation modeling in practice: a review and recommended two-step approach. Psychol. Bull. **103**(3), 411–423 (1988)

Baker, F.B.: Stability of two hierarchical grouping techniques case I: sensitivity to data errors. J. Am. Stat. Assoc. **69**(346), 440–445 (1974)

Boudreau, M.-C., Gefen, D., Straub, D.W.: Validation in information systems research. MIS Q. **25**(1), 1–16 (2001)

Clauset, A., Moore, C., Newman, M.: Hierarchical structure and the prediction of missing links in networks. Nature **453**(May), 98–101 (2008)

Cohen, J.: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. **70**(4), 213–220 (1968)

Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale (1988)

Davis, J.A.: A partial coefficient for Goodman and Kruskal's Gamma. J. Am. Stat. Assoc. **62**(317), 189–193 (1967)

Everitt, B.S.: The Analysis of Contingency Tables. Monographs on Statistics and Applied Probability, vol. 45. Chapman & Hall/CRC, Boca Raton (1992)

Gefen, D., Straub, D.W., Boudreau, M.C.: Structural equation modeling and regression: guidelines for research practice. Commun. Assoc. Inf. Syst. **4**(October), 7 (2000)

Geoffrion, A.M.: The formal aspects of structured modeling. Oper. Res. **1**, 30–51 (1989)

Göktaş, A., İşçi, Ö.: A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. Metodol. Zv. **8**(1), 17–37 (2011)

Goldreich, O.: Introduction to testing graph properties. Electron. Colloq. Comput. Complex. Rep. **82**(82), 470–506 (2011)

Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. J. Am. Stat. Assoc. **49**(268), 732–764 (1954)

Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications III: approximate sampling theory. J. Am. Stat. Assoc. **58**, 310–364 (1963). https://doi.org/10.1080/01621459.1963.10500850

Grassi, R., Fattore, M., Arcagni, A.: Structural and non-structural temporal evolution of socio-economic real networks. Qual. Quant. **49**(4), 1597–1608 (2015)

Green, K., Ricca, B.: Graph theoretic methods for the analysis of data in developing systems. Qual. Quant. **49**(5), 2037–2060 (2015)

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C.: Multivariate data analysis. In: Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C. (eds.) International Journal of Pharmaceutics, vol. 1. Prentice Hall, Upper Saddle River (1998)

Hambleton, R.K., Zaal, J.N.: Advances in Educational and Psychological Testing: Theory and Applications, vol. 28. Springer, Berlin (2013)

Higham, P.A., Higham, D.P.: New improved gamma: enhancing the accuracy of Goodman–Kruskal's gamma using ROC curves. Behav. Res. Methods **51**(1), 108–125 (2019)

Hopp, W.J., Iravani, S.M.R., Shou, B.: A diagnostic tree for improving production line performance. Prod. Oper. Manag. **16**(1), 77–92 (2007)

Hu, L., Bentler, P.M.: Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model. Multidiscip. J. **6**(1), 1–55 (1999)

Jiang, T., Wang, L., Zhang, K.: Alignment of trees—an alternative to tree edit. Theor. Comput. Sci. **143**(1), 137–148 (1995)

Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)

Nelson, T.O.: A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychol. Bull. **95**(1), 109 (1984)

Nunnally, J.: Psychometric Theory. McGraw-Hill, New York (1978)

Reiter, R.: A theory of diagnosis from first principles. Artif. Intell. **32**(1), 57–95 (1987)

Rooney, J.J., Van den Heuvel, L.N.: Root cause analysis for beginners. Q. Prog. **2004**, 45–53 (2004)

Rudick, M.M., Yam, W.H., Simms, L.J.: Comparing countdown- and IRT-based approaches to computerized adaptive personality testing. Psychol. Assess. **25**(3), 769–779 (2013)

Sartori, R.: The bell curve in psychological research and practice: myth or reality? Qual. Quant. **40**(3), 407–418 (2006)

Sartori, R., Pasini, M.: Quality and quantity in test validity: how can we be sure that psychological tests measure what they have to? Qual. Quant. **41**(3), 359–374 (2007)

Sengupta, K., Te'eni, D.: Cognitive feedback in GDSS: improving control and convergence. MIS Q. **17**(1), 87–113 (1993)

Shortliffe, E.: Computer-Based Medical Consultations: MYCIN, vol. 2. Elsevier, Amsterdam (2012)

van der Ark, L.A., van Aert, R.C.M.: Comparing confidence intervals for Goodman and Kruskal's gamma coefficient. J. Stat. Comput. Simul. **85**(12), 2491–2505 (2015)

Watkins, M.W., Pacheco, M.: Interobserver agreement in behavioral research: importance and calculation. J. Behav. Educ. **10**(4), 205–212 (2000)

Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: writing a review. MIS Q. **26**(2), 13–23 (2002)

Weinberg, A.I., Last, M.: Interpretable decision-tree induction in a big data parallel framework. Int. J. Appl. Math. Comput. Sci. **27**(4), 737–748 (2017)

You, W., Xia, M., Liu, L., Liu, D.: Customer knowledge discovery from online reviews. Electron. Mark. **22**(3), 131–142 (2012)