Inconsistency Indices in Pairwise Comparisons: An Improvement of the Consistency Index

Yuji Sato[*]**, Kim Hua Tan**

Nottingham University Business School
Jubilee Campus, Nottingham, NG8 1BB, United Kingdom

Abstract

The Consistency Index and the Consistency Ratio of the analytic hierarchy process (AHP) were designed to measure the ratio of inconsistent judgments among pairwise comparisons (PCs), which have been the principal indices for the past four decades. Definitions of inconsistency measures for PCs have yet to be established, however, because of the difficulty in quantifying subjectivity in judgments. Therefore, an empirical review that can take such subjective factors into account is essential. In this paper, the Consistency Ratio is thus reviewed using subjective data, and then a new inconsistency index for PCs is proposed based on the review. The review is based on subjective data obtained from two opinion surveys, which focuses on the relationship between the Consistency Ratio and two indicators: (1) the conformity of the results of the AHP and that of the ranking method, and (2) the goodness-of-fit of weight elicited by the AHP to human perception. A new inconsistency index is then proposed based on the mathematical property of a pairwise comparison matrix and further validated based on the conformity and the goodness-of-fit of weight. The results show that the proposed index detects inconsistency among real-world PCs more sensitively than could the Consistency Ratio; the index might suggest the reliability of the output of a pairwise comparison matrix.

Keywords Decision support systems · Pairwise comparison · Inconsistency · Human perception

Declarations

Funding: This research was supported in part by the Chukyo University Grant for Overseas Research Program, 2019.

[*] Corresponding author: ysatoh@1988.jukuin.keio.ac.jp
ORCID: 0000-0002-3317-7350

1 Introduction

When making decisions, a subject (individual or group) expresses preferences for alternatives in some way, for example, by rating, ranking, or multiple choice. Pairwise comparison (PC) has also been used, and presents some advantages compared to the other methods. For example, rating methods require that the calibration of preferences across subjects be synthesized, a very challenging task. The problem with the ranking method, whereby all alternatives must be ordered from the most preferred to the least, is that it requires more time than does PC (e.g., Inglehart and Abramson, 1993). The number of possible rankings also increases exponentially with any increase in the number of alternatives. Although the multiple-choice method allows decision makers to easily express their preferences, the precision of its output is not high, as no information for the not-selected alternatives can be obtained. In contrast, PC provides precise output and requires neither synthesis nor longer processing time. Nevertheless, the quadratic increase in number of comparisons and the inconsistency among redundant PCs remain to be resolved.

Ever since the pioneering work of Thurstone (1927), PC began being used primarily in the field of psychology. PC was then widely applied to decision-making methods in measuring human perception because of its flexibility and applicability to diverse situations. One significant application of PC in decision science has been to the analytic hierarchy process (AHP) developed by Saaty (1977). The AHP aggregates data obtained from a decision maker's PCs, and then quantifies his/her preference for alternatives. The quantification results not only in the identification of the most important alternative but also in the ranking of all alternatives. As such, the AHP has been applied to a wide variety of fields thanks to its user-friendly interface, simple architecture, and applicability to real problems. For example, it has been applied to the fields of manufacturing (e.g., Dey et al., 2015; Sato et al. 2017), logistics (e.g., Kengpol and Tuammee, 2016; Moktadir et al., 2018), and public administration (e.g., Dong et al., 2013; Ishizaka and Labib, 2014).

Nevertheless, because the AHP requires redundant PCs for decision makers, the transitivity of preferences is sometimes violated. Therefore, Saaty (1980) developed the Consistency Index (CI) to measure the ratio of inconsistent judgments among PCs. In the literature, the Consistency Ratio (CR) was also developed to take the size of a pairwise comparison matrix (PC-matrix) into account. Both indices have been considered crucial determinants for judging the degree of inconsistency of a PC-matrix, requiring that the CI (CR) remain in an appropriate range, viz. less than 0.1. However, criticisms arose (e.g., Lootsma, 1991; Murphy, 1993) focusing on how to interpret the CI and how to deal with PC-matrices with a large CI. Based on extensive discussions by Saaty responding to each of the criticisms (e.g., Saaty, 1997; 2001a; 2001b; 2005; 2006), the AHP evolved to include work on the interpretation of the CI. Later, Saaty (2010) noted that the CI should be used as an indicator to improve decision making as follows: as the CI indicates the ratio of inconsistent

judgments among PCs, a PC-matrix with a large CI needs to be reassessed by finding inconsistent judgments in the matrix. Decision makers are then encouraged to refine inconsistent judgments to plausible values. More recently, Yoon (2020) has shown that the inconsistency negatively affects the Compatibility Index that is useful for validating or grouping the differences of multiple pairwise comparison matrices.

In the reassessment, AHP software such as expertchoice® can reduce the burden for decision makers in finding inconsistent judgments. The AHP, however, may not always be implemented with such well-designed software, as in cases when Excel® may be used instead. An application of the AHP to opinion surveys is a case in point. Since weights elicited by the AHP can precisely describe each respondent's preference for alternatives given in a questionnaire, their application is of significant use (e.g., Sato et al., 2015; Moktadir et al., 2018). Because surveys are often conducted by mail, however, providing respondents with questionnaires equipped with pay software, e.g., expertchoice®, would not be feasible. As such, respondents would have no way to review their PCs to know whether their CIs would satisfy the threshold. Reassessments by respondents would require a great deal of time and effort, thus diminishing the applicability of the AHP.

In order to address inconsistency among PCs, a variety of indices have been proposed in addition to the CI (CR). Table 1 summarizes the major indices proposed to date. Based on an exhaustive review of these indices, Brunelli (2018) studied inconsistency indices for PCs. Szybowski et al. (2020) updated the work and have further addressed inconsistency by focusing on the differences between complete and incomplete PC-matrices. In other words, systematic reviews focusing on mathematical properties of the indices have been conducted. Nevertheless, one issue must be noted here: how to validate the indices. As can be seen in Table 1, most verifications of the indices use numerical examples and/or simulations. Shiraishi et al. (1998), Gass and Rapcsák (2004), Kou and Lin (2014), Kułakowski (2015), and Szybowski et al. (2020) validated their indices using numerical examples, while Crawford and Williams (1985), Stein and Mizzi (2007), Grzybowski (2016), and Fedrizzi and Ferrairi (2018) employed simulation in their verifications.

Table 1 Major indices and approaches to validation.

Index	Reference	Validation
CI, CR	Saaty (1977, p. 237)	Experiment
s^2	Crawford & Williams (1985, p. 395)	Simulation
G	Golden & Wang (1989, p. 71)	Experiment
MC(**A**)	Takeda (1993, p. 32)	* *mathematical property only*
CM(**A**)	Duszak & Koczkodaj (1994, p. 275)	* *mathematical property only*
AI(**A**)	Salo & Hämäläinen (1995, p. 465)	Experiment

RE(**A**)	Barzilai (1998, p. 125)	Numerical example
C_3	Shiraishi et al. (1998, p. 407)	Numerical example
IM	Gass & Rapcsák (2004, p. 581)	Numerical example
HCI	Stein & Mizzi (2007, p. 492)	Simulation
$I_{\mathscr{G}}$(**A**)	Cavallo & D'Apuzzo (2009, p. 395)	* *mathematical property only*
CCI	Kou & Lin (2014, p. 228)	Numerical example
$\mathscr{E}_{\max}$(**A**, *w*)	Kułakowski (2015, p. 334)	Numerical example
ATI	Grzybowski (2016, p. 202)	Simulation
$I_{\chi 2}$(**A**)	Fedrizzi & Ferrairi (2018, p. 1127)	Simulation
MII(**A**), *KII*(**A**)	Szybowski et al. (2020, p. 140)	Numerical example

An index can also be validated through experimentation. For example, Saaty (1977, p. 253) validated the efficacy of the CI based on the famous experiment on "*relative visual brightness*." Golden and Wang (1989, p. 77) compared their G-value with the CI through the case of "*Iran hostage rescue decision*." Salo and Hämäläinen (1995, p. 467) verified their interval judgment based on the case study of "*Finnish energy policy decision*." Of particular note are the validations by Golden and Wang (1989) and by Salo and Hämäläinen (1995). The experiments they used dealt with subjective judgments in the real world, meaning that those experiments did not have a "true" answer, which is known, as was the case with experiments in physics. Since it is often quite challenging to judge whether findings from such experiments are correct, far fewer verifications are conducted using experiments on subjective issues than those based on numerical examples or simulations. The fact remains, however, that given that PCs are used to describe subjective human perceptions, indices for PCs must be validated using subjective data. Brunelli (2018, p. 765) stressed the importance of empirically verifying indices using real data: "*considering the scarcity of empirical studies in the literature, it is foreseeable that new studies of this type could be valuable contributions*." In addition, the volume of data needs to be noted in the verifications to generalize the results of a validation. Data from not just one expert but from a large number of decision makers would be preferable.

In addition to determining how best to validate indices, another issue that needs to be addressed is determining the properties that ought to be satisfied by the indices. Since the AHP is a tool for decision making, weights obtained from a PC-matrix must contribute to supporting decision makers by accurately representing their perceptions. Indeed, from the earliest works of the AHP, Saaty (1977) regarded its usefulness to be in social applications. He thus emphasized the efficacy of the AHP and used the CI based on various experiments, which had a variety of objects and setups. Nevertheless, a great many studies have suggested that the degree of inconsistency among PCs affects the reliability of the output of the AHP: weights of alternatives. As shown in Table 1, although various indices have been proposed to measure inconsistency among PCs,

indices assessing the reliability of the weights have yet to be developed. Since a review of the relationship between input to PCs and its output would, of necessity, be retrospective by nature, decision makers cannot be aware how well the weights of alternatives represent their perceptions when implementing PCs. Therefore, an index not only detects inconsistencies among PCs but also suggests the reliability of weights of alternatives obtained from PCs would be of practical use. Hence, the objectives of this paper are as follows.

- To empirically review the CR using subjective data.
- To develop a new inconsistency index for PCs that measures inconsistency and suggests the reliability of weights of alternatives.

To address these research questions, an index is reviewed by using data obtained from survey questionnaires on subjective issues in this paper. The review focuses on the CR from among a variety of indices, as it has been the principal inconsistency index for PCs for the past four decades. By using a large amount of subjective data, the review is carried out from the following two perspectives: (1) the detectability of inconsistency among PCs and (2) the predictability of the reliability of weights obtained from a PC-matrix. A new inconsistency index is then proposed, which can not only detect inconsistency among PCs but also suggest the reliability of the weights. The index is then validated in the same way as the review of the CR. This paper is organized as follows. Section 2 summarizes the definitions and the interpretations of indices, and formulates two hypotheses examining the CR. The hypotheses are tested in Section 3 using data obtained from the applications of the AHP to the real world. Section 4 proposes a new inconsistency index followed by its validation. Section 5 concludes this paper and presents directions for future research.

2 Definitions and interpretations of indices

The CI was designed to measure the transitivity of preferences among PCs (Saaty, 1977). The CR is designed to be a similar index, taking the size of a PC-matrix into account. The definitions of the CI and the CR are as follows:

$$\mathrm{CI} := (\lambda_{max} - n)/(n - 1), \tag{1}$$

$$\mathrm{CR} := \mathrm{CI}/\mathrm{RI}, \tag{2}$$

where λ_{max} denotes the principal eigenvalue of a PC-matrix, n represents the matrix size, and RI (the Random Index) is the average of CIs based on PC-matrices whose elements are generated with uniform randomness. Both the CI and CR have been considered the principal criteria for judging the consistency of a PC-matrix. Since a PC-matrix is a non-negative reciprocally symmetric matrix, the following proposition holds: the matrix has one eigenvalue that equals n and the

remaining eigenvalues are all zero, if and only if the transitivity of preferences entirely holds in the matrix. Therefore, the more inconsistent the judgments included in PCs are, the larger the CI of a PC-matrix would be. All things being equal, the smaller the CI is, the more consistent the judgments made in PCs.

In addition to checking the consistency of a PC-matrix, we also need to determine how to measure inconsistency among PCs and how to deal with a PC-matrix with a large CI. As a weight with CR=0 is obtained from a consistent PC-matrix where the transitivity of preferences holds entirely within a PC-matrix, an inconsistent PC-matrix could thus be considered to induce "wrong" weights. Therefore, if decision makers are to trust the output of the AHP, the consistency of a PC-matrix must be determined. As discussed in Section 1, empirical verification would require a great deal of subjective data, which is what we have done here, by reviewing the CR using data obtained from the application of the AHP to subjective issues. In the case of a PC-matrix with a large CI, some papers necessitated the refinement of the CI by mathematically adjusting the elements of the PC-matrix (e.g., Alonso and Lamata, 2005; Zhang et al., 2014); other papers requiring the application of the AHP wound up discarding PC-matrices with a large CI. For example, Chang et al. (2007) discarded decision matrices if CR>0.15; Chen et al. (2010) did so when CR>0.1. While the reasonableness of accepting/rejecting a PC-matrix is not discussed in this paper, clarifying guidelines for inconsistent PC-matrices when using the AHP would be advisable from a practical point of view.

In the review of the CR using real-world PCs, this paper focuses on two functions: (1) the detectability of inconsistency among PCs, and (2) the predictability of the reliability of weights obtained from a PC-matrix. To verify detectability, we review the CR based on a criterion called CAR.

Definition 1 CAR (Conformity of rankings between the AHP and the ranking method)

The degree of conformity between two rankings: one based on weights elicited from the AHP, and the other obtained from the ranking method.

The CAR is measured by comparing two rankings of alternatives obtained using two different methods. The rankings employed in this paper were obtained from an opinion survey consisting of questions about an issue formatted in two ways. One is formatted by the AHP requiring that respondents implement PCs; the other uses the ranking method. The AHP provides the ranking of alternatives based on weights obtained from a PC-matrix; the ranking method directly outputs the ranking of alternatives. Since both rankings signify a respondent's preference for the set of alternatives, the same, or at least similar, rankings of alternatives are presumably derived. The high (low) correlation group can thus be considered to be composed of consistent (inconsistent) respondents. The degree of CAR is determined by a Spearman rank-correlation coefficient

between the two rankings. In this paper, the results are simply classified into two groups, a "high correlation group" and a "low correlation group," as the rankings are obtained from opinion surveys on subjective issues. According to the themes of the surveys (see Sections 3.1 and 3.2), more precise classification, such as that being done in physical experiments having theoretical thresholds, would not be appropriate in these cases. Indeed, a two-way ANOVA comparing factors of consistency (threshold value is Saaty's yardstick, CR=0.1) and conformity (threshold values are Spearman rank-correlation coefficients, r_S=0.7, 0.4, and 0.2. See Table 2 below), respectively, results in p-values of 0.31653 for consistency and 0.10593 for conformity. In short, null hypotheses examining the difference of means of CRs among the four groups cannot be rejected. Therefore, the differences in the distribution of CRs between the two above-mentioned groups are then verified by the Mann-Whitney U test. The correlation between the CR and the CAR could thus measure the CR's detectability of inconsistency among real-world PCs.

In addition to a review of the CR based on the CAR, the CR is reviewed based on an additional criterion, called GWP, to verify predictability.

Definition 2 GWP (Goodness-of-fit of weight to human perception)

The degree of goodness-of-fit of weight obtained from a PC-matrix to a decision-maker's perception assessed by him/herself.

Goodness of fit typically summarizes the discrepancy between observed values and the values expected under the model in question. The degree of GWP is measured through a panel survey in which participants, in the first round of a survey, implement PCs on a subjective issue. Once the results of PCs are fed back to participants, in the second round they assess how precisely the weights obtained from their PC-matrices represent their perceptions. The GWP thus summarizes the discrepancy between participants' perceptions of the input and the output—PCs and a set of weights of alternatives—to and from a PC-matrix. In other words, the GWP aims to assess the reliability of weights by reviewing how a decision maker implementing PCs retrospectively perceives the output from PCs. Based on the assessment, the degree of GWP is classified as "True," "Rather true," "Neutral," "Rather false" or "False." Since the CR is one of the components of this tool for decision making, ideally the index suggesting the reliability of weights should be obtained from a real-world PC-matrix. Even indices with good mathematical properties would be of little practical use if they did not contribute to supporting decision makers. The output must always be deemed trustworthy.

The above discussion leads us to propose the following hypotheses:

H_1: There is no significant difference in the distribution of CRs between the two groups classified by the CAR.

H_2: There is no significant difference in the size of CRs among the degrees of the GWP.

Hypothesis H_1 examines whether the distribution of CRs differs between the high and low correlation groups; hypothesis H_2 tests whether the CR correlates with decision makers' perceptions of the weights obtained from a PC-matrix. Since the verifications based on the CAR and the GWP use real-world PC-matrices, the empirical approach employed in this paper would prove the efficacy of the CR in more realistic terms than would the use of numerical examples or simulations.

3 Analyses

In this section, the two hypotheses, H_1 and H_2, are examined through opinion surveys. The use of subjective data obtained from the surveys derives from the extant criticisms to numerical examples and simulations—the lack of realistic conditions that real-world PCs must incorporate, such as biases. Two surveys (hereafter, Surveys 1 and 2) employed in the analyses were carried out in 2010, in which respondents were university students in Japan. Survey 1 was a one-shot survey designed for examining hypothesis H_1, while Survey 2 was a panel survey for testing hypothesis H_2. As Survey 2 required respondents in the first round to answer the questions in the second-round, participants in the surveys needed to be controllable. In addition, for the findings of reviews to be generalized, a certain number of respondents were needed in the surveys. Based on these requirements, students in a university were then chosen as respondents. Consequently, the number of respondents in the surveys were 370 (Survey 1) and 371 (Survey 2), respectively.

3.1 Verification of H_1

The test of Hypothesis H_1 uses data obtained from Survey 1 examining the correlation between the CR and the CAR. Participants in Survey 1 were university students in Japan who were asked to clarify their political opinions. In the questionnaire, two differently formatted questions on a particular issue were posed, each asking students to express their opinions on four controversial political issues in Japan at the time of the survey: (*i*) conservation of the social order, (*ii*) encouragement of political participation, (*iii*) reduction of the unemployment rate, (*iv*) protection of free speech. Students were asked which of the four issues they thought more important than the others. One format employed the ranking method, in which students were asked to rank the four issues given in the question from the most important to the least; the other was formatted using the AHP, which required the students to implement PCs across all possible combinations of the four issues.

The hypothesis test focuses on pairs of rankings of the same set of four alternatives. The ranking of alternatives obtained from the ranking method directly represents the respondent's preference order of alternatives; the ranking elicited from the AHP is based on the weight of each alternative. Thus, by comparing the CRs of PC-matrices between the high and low correlation groups, the CR's detectability of inconsistency among real-world PCs can be verified.

Table 2 summarizes the results of the analysis based on the Spearman rank-correlation coefficient, r_S. In this study, the CR of a set of rankings with $0.7 \leq r_S \leq 1$ is categorized as the high correlation group, while that with $-1 \leq r_S < 0.7$ is categorized as the low correlation group. As can be seen in the table, the rankings obtained from the differently formatted questions correlate to some degree (67.6% = 250/370). In addition, as Figure 1 illustrates, the distribution of CRs of the two groups correlate with each other, with the mean of CRs of the high (low) group being 0.20042 (0.21215), and their variance at 0.067154 (0.097263), respectively. Indeed, as shown in Table 3 summarizing the results of the Mann-Whitney U test between the two groups, Mann-Whitney U = 14548 and Z=-0.46933, which results in a p-value of 0.63836. These results thus indicate no significant difference exists in the distribution of CRs between the two groups. Hence, Hypothesis H_1 cannot be rejected. That is, the CR might not sensitively detect the inconsistency among real-world PCs based on the CAR.

Table 2 Results of the analysis based on Spearman rank-correlation coefficient.

r_S	# of obs.	Group	# of categorized obs.	Mean of CR (CR_β*)	Variance of CR (CR_β*)
[-1, 0.2)	25	low correlation	120	0.2234652 (0.3907898)	0.1275835 (0.2086084)
[0.2, 0.4)	58				
[0.4, 0.7)	37				
[0.7, 1]	250	high correlation	250	0.1949937 (0.2437422)	0.05307211 (0.08292517)

* See Section 4.1.

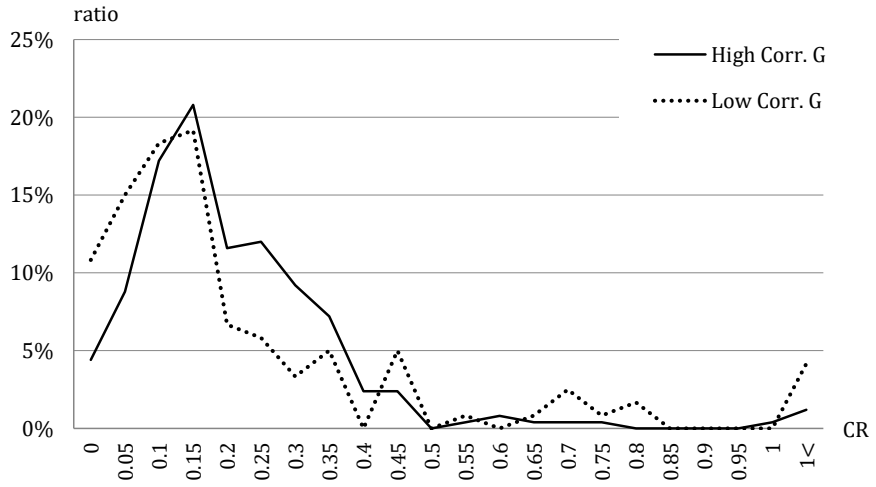

Fig. 1 Distributions of CRs.

Table 3 Results of the Mann-Whitney U test.

Mann-Whitney U	Mean	Variance	Z	p-value
14548	15000	927500	-0.46933	0.63836

3.2 Verification of H_2

The test of Hypothesis H_2 uses data obtained from Survey 2, which examines the correlation between the CR and the GWP. Participants in Survey 2 were university students in Japan searching for jobs who were asked to identify their preferences in their search. Survey 2 consisted of two rounds. In the first round, a question formatted using the AHP asked about the alternatives of: (*i*) salary, (*ii*) place of business, (*iii*) self-fulfillment, (*iv*) welfare system. Provided the feedback—a set of weights of four alternatives elicited from the AHP—from the first round, the students then assessed in the second round how precisely the weights represented their preferences.

The hypothesis test focuses on the degree of goodness of fit between the input and the output to and from a PC-matrix. The CR measures the ratio of inconsistent PCs within a PC-matrix; the index evaluates the input to the matrix. The GWP indicates a decision maker's assessment of a set of weights obtained from a PC-matrix; the index evaluates the output of the matrix. Since the AHP quantifies decision makers' perceptions through a PC-matrix, the compatibility between the input and the output of the matrix is crucial. Granted, the CR might not always be proportionate to the

degree of GWP. However, if the CR correlates with the degree of GWP, we could infer from the CR the reliability of weights obtained from a real-world PC-matrix.

In the second round, the students assess the GWP as follows:

- True: strongly agree with the weights obtained from the AHP
- Rather true: agree with the weights obtained from the AHP
- Neutral: neither agree nor disagree with the weights obtained from the AHP
- Rather false: disagree with the weights obtained from the AHP
- False: strongly disagree with the weights obtained from the AHP

Students were asked to choose one option from among the five; students were not notified of the CR of their PCs in the first round.

Figure 2 illustrates the distribution of CRs in the second round of Survey 2 classified by the degree of GWP. Table 4 shows the correlation coefficients for the five assessments and the means of their CRs; Table 5 summarizes the results of the assessments using ANOVA. As illustrated in Figure 2, the distributions of CRs correlate. Indeed, the correlation coefficients among the five assessments shown in Table 4 range between 0.92045 and 0.75238. Furthermore, Table 5 indicates that $F=0.85428<F_{crit}=2.4582$ and the p-value=0.49410. As a result, there is no significant difference in the CR among the assessments at p=0.05. Hence, Hypothesis H_2 cannot be rejected. Through the panel survey evaluating the correlation between the CR and the GWP, it is clear that the ability of the CR to predict the reliability of weights obtained from a real-world PC-matrix is not significant. That is, the CR might not contribute to implementing PCs from a practical perspective.

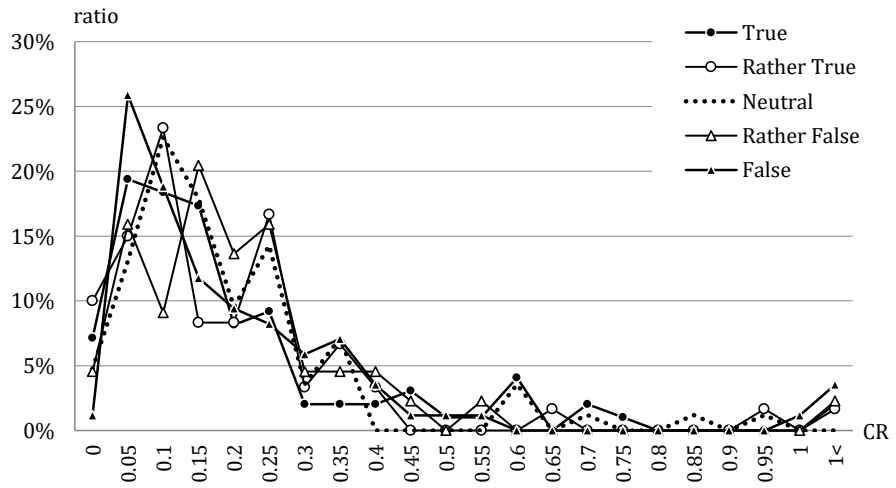

Fig. 2 Distributions of CRs.

Table 4 Correlation among the assessments, means of the CR.

GWP	True	Rather true	Neutral	Rather false	False	Mean of CR
True	1					0.2067549
Rather true	0.8539965	1				0.1943344
Neutral	0.9204539	0.9126712	1			0.1746002
Rather false	0.8583155	0.7523808	0.8455428	1		0.1964678
False	0.9020223	0.8300841	0.8354226	0.7969085	1	0.2091477

Table 5 Results of ANOVA.

Source of Variation	SS	df	MS	F	P-value	F_{crit}
Between Groups	86.03636	4	21.50909	0.8542838	0.4941003	2.458210
Within Groups	2643.682	105	25.17792			
Total	2729.718	109				

4 Proposal of a new inconsistency index

In this section, a new inconsistency index for PCs is proposed focusing on the missing information of eigenvalues of a PC-matrix. The index is then validated based on the two criteria of the hypothesis tests of the CR: the CAR and the GWP.

4.1 Definition of a new inconsistency index

This paper has uncovered the possibility of creating a new inconsistency index from the calculation processes of indices proposed to date. As illustrated by Brunelli (2018, p. 759, Fig. 4), some indices share many similarities, which might stem from the summation of eigenvalues of a PC-matrix in calculating indices, as the CI does. If the matrix has complex eigenvalues, then they occur in conjugate pairs by nature. For example, a PC-matrix **A** has the following eigenvalues:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1/2 & 1 & 2 & 3 & 4 \\ 1/3 & 1/2 & 1 & 2 & 3 \\ 1/4 & 1/3 & 1/2 & 1 & 2 \\ 1/5 & 1/4 & 1/3 & 1/2 & 1 \end{pmatrix}$$

λ_1=5.0681, λ_2=0.0049888+0.58276i, λ_3=0.0049888-0.58276i, λ_4=-0.039029+0.068286i, λ_5=-0.039029-0.068286i, which result in CI=0.017025 and CR=0.015201, where CI_β*=0.34255 and CR_β*=0.12408 (* see Definition 3 below). As such, the imaginary parts of the eigenvalues are cancelled when calculating indices. Based on this property of the matrix, we propose a new inconsistency index.

Since a PC-matrix is positive and reciprocally symmetric square, it has a unique real principal eigenvalue corresponding to the Frobenius root, whose corresponding eigenvector is always strictly positive, according to the Perron-Frobenius theorem (Perron, 1907; Frobenius, 1912). The principal eigenvalue equals the size of the matrix, n, and the remaining eigenvalues are all zeros, if and only if transitivity of preferences holds entirely within the matrix. If not, the remaining eigenvalues must occur in conjugate pairs because the roots of the characteristic polynomial of a PC-matrix with real elements occur in conjugate pairs, $\lambda=\alpha\pm\beta\cdot i$ where α and β respectively denote the real and the imaginary parts of an eigenvalue. Let $\mathbf{A}=(a_{kl})$, $(k, l=1, \ldots, n)$ be a PC-matrix. Further, let $\lambda_j=\alpha_j\pm\beta_j\cdot i$, $(j=1, \ldots, n)$ and λ_{max} respectively denote the eigenvalues and the principal eigenvalue of $\mathbf{A}$. Since diagonal elements a_{kk} equals 1 for any k, $\mathrm{tr}(\mathbf{A})=n$, then, $\sum_{j=1}^{n}\lambda_j=n$ holds, which results in

$$\mathrm{CI}=(\lambda_{max}-n)/(n-1)=(\lambda_{max}-\sum_{j=1}^{n}\lambda_j)/(n-1). \quad (3)$$

Eq. (3) means that the CI calculates the average of eigenvalues apart from the principal eigenvalue of $\mathbf{A}$. The CI is thus calculated based only on α_j; the imaginary part, β_j, is cancelled in its process.

Figure 3 illustrates the distributions of α and β of randomly generated PC-matrices coordinated by the CI, where $n=5$ and the sample size is 1000. As shown in the figure, the distribution of α seems to have a lower envelope with respect to the CI. On the other hand, β has a different distributional property from α, which likely has an upper envelope with respect to the CI. The relationship between the mathematical grounds of the distributions and human perception is not clear, though one implication can be derived as follows. According to the distribution of β, the CI is defined as the ignoring of a certain amount of information provided by the imaginary part of eigenvalues, which might affect the CI's detectability of inconsistency within a PC-matrix. Therefore, focusing on both the real and the imaginary parts of eigenvalues, a new inconsistency index, called CI_β, is defined in this paper.

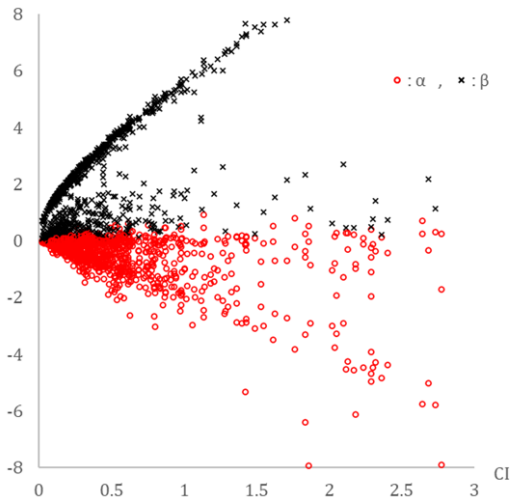

Fig. 3 Distributions of α and β.

Definition 3 CI_β and CR_β

$$CI_\beta := (\lambda_{max} - n + \sum_{j=1}^{n} |\beta_j|)/(n-1), \quad (4)$$

$$CR_\beta := CI_\beta / RI_\beta, \quad (5)$$

where RI_β is the average of CI_βs based on PC-matrices whose elements are generated with uniform randomness.

Table 6 summarizes the RI_βs for matrix sizes from 3 to 9 to obtain the threshold values based on randomly generated 1000 PC-matrices.

Table 6 RI_β (n=3, ..., 9).

n	3	4	5	6	7	8	9
RI_β	1.2471	2.2715	2.7608	3.0829	3.8272	4.1567	4.7004

4.2 Properties of the CI_β

Brunelli and Fedrizzi (2015) and Brunelli (2017) presented the following six properties that an inconsistency index of a PC-matrix would necessarily satisfy. Indices proposed in the literature were tested (e.g., Fedrizzi and Ferrari. (2018)), and some satisfied all of the properties and others were proven to violate several of the properties. In this paper, the CI_β is formally proved to satisfy

Properties 1, 2, 5, and 6, and verified through numerical simulations for Properties 3 and 4 as follows.

Property 1 *There exists a unique real number representing the situation of full consistency.*

$\lambda_{max}=n$ and $\beta_j=0$ for any j, if and only if transitivity of preferences in a PC-matrix **A** holds entirely within the matrix. Therefore, it directly follows that $CI_\beta=0$, if and only if **A** is consistent.

Property 2 *An inconsistency index is independent from the order of the alternatives.*

For any permutation matrix **P**, the characteristic equation of a matrix $\mathbf{PAP}^T$ is the same as that of **A**. CI_β obtained from $\mathbf{PAP}^T$ is thus equal to that of **A**.

Property 3 *An inconsistency index cannot decrease if inconsistent preferences are intensified.*

This property is verified through numerical simulation in this paper. For any PC-matrix $\mathbf{A}=(a_{kl})$ ($a_{kl}\in[1/9, 9]$, k, l=1, …, n), let $\mathbf{A}(b)=(a^b_{kl})$ be a PC-matrix whose preferences are modified by a parameter $b\in R$. Then the λ_{max} of $\mathbf{A}(b)$ is a non-decreasing function for $0\leq b$ and hence for $1<b$, which represents an intensification of preferences (see Brunelli and Fedrizzi (2015)). This intensification of inconsistent preferences thus results in $0\leq d\mathrm{CI}/db$. On the other hand, Figures A1-A7 illustrate comparisons between the CI and the CI_β for n=3, …, 9, where CI and CI_β are calculated based on randomly generated 1000 PC-matrices. As shown in the figures, indices are concavely distributed or have concave upper and lower envelopes, each of which suggests $0<d\mathrm{CI}_\beta/d\mathrm{CI}$. Therefore, $0\leq d\mathrm{CI}_\beta/db=d\mathrm{CI}_\beta/d\mathrm{CI}\cdot d\mathrm{CI}/db$ hold and CI_β satisfies this property.

Property 4 *An inconsistency index is non-decreasing with respect to an elementary modification of a consistent matrix.*

The same logic in the verification of Property 3 directly follows that CI_β satisfies this property.

Property 5 *An inconsistency index is continuous with respect to the entries of a matrix.*

The root of an algebraic equation is a continuous function of coefficients of the equation, and coefficients of a characteristic equation of a matrix is a continuous function of entries of the matrix. The eigenvalue of a matrix is thus a continuous function with respect to the entries of the matrix, which proves that CI_β satisfies this property.

Property 6 *An inconsistency index is invariant under inversion of preferences.*

Since the characteristic equation of a matrix $\mathbf{A}^T$ is the same as that of a matrix **A**, CI_β satisfies this property.

Therefore, the CR_β is an index that incorporates the missing information of β_j into the definition, which has similar properties to those of the CR. Figure 4 illustrates the scatter plot comparing the values of the CR and the CR_β whose data are obtained from Survey 2.

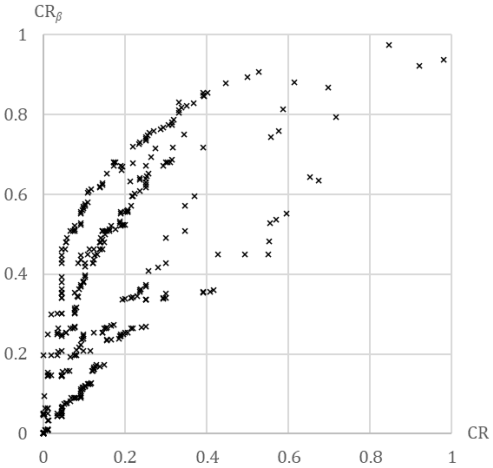

Fig. 4 Scatter plot of CR-CR_β (from Survey 2).

Based on the above discussion, the following hypotheses are formulated.

H_3: There is no significant difference in the distribution of CR_βs between the two groups classified by the CAR.

H_4: There is no significant difference in the size of CR_βs among the degrees of GWP.

4.3 Verifications of H_3 and H_4

The tests of Hypothesis H_3 and H_4 use data obtained from Surveys 1 and 2, respectively. As was done in Sections 3.1 (for H_1) and 3.2 (for H_2), the verifications focus on the relationship between the CR_β and the degrees of the CAR and the GWP in hypothesis tests of H_3 and H_4, respectively. CR_βs are categorized according to Table 2 summarizing the Spearman rank correlation coefficient, r_S, in Section 3.1.

To test Hypothesis H_3, Figure 5 illustrates the distributions of CR_βs in Survey 1, which shows apparent differences between the high and low correlation groups. Comparing Figures 5 to 1, the distributions in the two figures differ from each other. Indeed, as shown in parentheses (*) in Table 2, the mean of CR_βs of the high (low) group is 0.24374 (0.39079), and their variance is 0.082666 (0.20699), respectively. Furthermore, as shown in Table 7 summarizing the results of the Mann-Whitney U test between the two groups, Mann-Whitney U=11079 and Z=-4.0708, which results in a p-value of 0.000046739. These results indicate that there are significant differences in the distribution of CR_βs between the two groups. Hence, Hypothesis H_3 is rejected. Through empirical verification of the correlation between the CR_β and the CAR, the CR_β's detectability of

inconsistency among real-world PCs is confirmed. Compared to the results of the test of Hypothesis H_1 (a p-value of 0.63836 for the CR. See Table 3), the results of the test of Hypothesis H_3 show that the CR_β detects the inconsistency more sensitively than does the CR.

Fig. 5 Distributions of CR_βs.

Table 7 Results of the Mann-Whitney U test.

Mann-Whitney U	Mean	Variance	Z	p-value
11079	15000	927500	-4.0708	0.000046739

For the test of Hypothesis H_4, Figure 6 illustrates the distributions of CR_βs in the second round of Survey 2 classified by the degree of GWP. Table 8 shows the correlation coefficients for the five assessments and the means of their CR_βs; Table 9 summarizes the results of the assessments using ANOVA. As illustrated in Figure 6, the distribution of CR_βs differs among the five assessments. Indeed, the correlation coefficients among the assessments shown in Table 8 range between 0.63453 and -0.57673, which indicates they do not correlate with each other. Furthermore, the mean of the CR_βs gradually increases from "True" (0.20331) to "False" (0.58018), which might indicate an inverse relationship between the CR_β and the reliability of the output of a real-world PC-matrix. On the other hand, Table 9 indicates that $F=2.0135<F_{crit}=2.4582$ and the p-value= 0.097892. Consequently, there are differences among the five assessments at p=0.10, while no significant difference could be found at p=0.05. Thus, the ANOVA results are not statistically significant and Hypothesis H_4 cannot be rejected, despite the fact that the CR_β and the GWP correlate with each other. Hence, the results of the test of Hypothesis H_4 remain inconclusive.

Through the panel survey evaluating the correlation between the CR_β and the GWP, the ability of the CR_β to predict the reliability of weights obtained from a real-world PC-matrix might have been implied to some extent. That is, the CR_β might be seen as contributing to implementing PCs from a practical perspective.

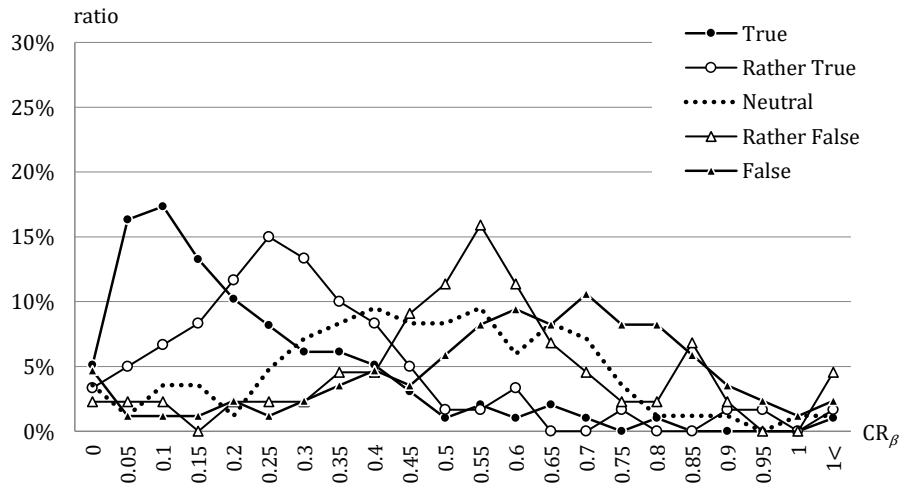

Fig. 6 Distributions of CR_βs.

Table 8 Correlation among the assessments, means of the CR_β.

GWP	True	Rather true	Neutral	Rather false	False	Mean of CR_β
True	1					0.2033068
Rather true	0.6116181	1				0.2952003
Neutral	-0.07264067	0.1910349	1			0.4407196
Rather false	-0.3210833	-0.2517166	0.6345343	1		0.5222685
False	-0.5767363	-0.5475067	0.3945195	0.5605225	1	0.5801768

Table 9 Results of ANOVA.

Source of Variation	SS	df	MS	F	P-value	F_{crit}
Between Groups	86.03636	4	21.50909	2.013454	0.09789164	2.458210
Within Groups	1121.682	105	10.68268			
Total	1207.718	109				

5 Concluding remarks and future research

This study empirically reviewed the CR from a practical perspective and proposed a new inconsistency index for PCs. The review explored the capabilities of the CR both in detecting inconsistency among PCs, and in suggesting the reliability of the output of a PC-matrix. In the review, the relationship between the CR and the degrees of the CAR and the GWP were verified using subjective data obtained from opinion surveys. The verification of the CR based on the CAR empirically compared representations of decision makers' perceptions obtained from the ranking method and the AHP. This approach derived from criticisms to "random culture" (Brunelli, 2018) in reviewing indices—employing numerical examples and/or data obtained from simulation—and incorporated practical conditions for real-world PCs. On the other hand, the verification of the CR based on the GWP went in a new direction to review indices for PCs. The use of a panel survey enabled us to conduct a retrospective evaluation of the compatibility of the input to PCs and the output of a PC-matrix. As noted in Section 1, the degree of inconsistency among PCs affects the reliability of the output of the AHP. The approach based on the GWP that evaluated reliability could thus provide new insights into the review of indices for PCs. The results showed that (*i*) based on the CAR, the CR might not sensitively detect inconsistency among real-world PCs, and (*ii*) based on the GWP, the CR did not correlate with decision makers' perceptions of the output of a PC-matrix.

In addition to the review of the CR, this study proposed a new inconsistency index, CR_β. The index refined the CR by incorporating complementary information provided by the imaginary part of eigenvalues of a PC-matrix that had been ignored by indices proposed to date. The CR_β was then validated based on the same criteria for the review of the CR in this paper: the CAR and the GWP. The results of reviews using subjective data showed that (*iii*) the CR_β detected inconsistency among real-world PCs more sensitively than could the CR, (*iv*) the CR_β might, but could not conclusively, suggest the reliability of the output of a PC-matrix. Since decision-makers agreeing with results of PCs do not consistently compare alternatives, nor do those disagreeing with results of PCs always make inconsistent judgments, result (*iv*) might not yet be solid. Kułakowski and Talaga (2020, p. 196) noted on the definition of indices: "*Finding a solution that combines robustness with the simplicity of implementation and calculation will still be a challenge for researchers*." Given such background in defining indices, the GWP could be a possible criterion for researchers of the AHP, and the CR_β could be an alternative inconsistency index for users of the AHP dealing with real-world PCs, respectively.

Nevertheless, this paper has some limitations. Unlike physical values, there are no "right" answers in subjective issues: determining definitive criteria for the review of indices is challenging. Rankings obtained from the ranking method and those elicited by the AHP do not necessarily

measure the same object in a narrow sense: the degree of CAR might just be a rough standard of inconsistency among PCs. Likewise, a high degree of GWP does not necessarily imply a right decision, nor would a low degree of GWP suggest a wrong decision: the degree of GWP might not be a solid criterion for reviewing indices. These criteria for the review of indices need to be explored in future research.

The new index, CR_{β}, also requires further verification. Properties 3 and 4 are verified through numerical simulation, though formal proofs will still be required. In addition, this paper validates the index through two opinion surveys, each of which has four alternatives in the questionnaires. We need to follow up with additional cases to investigate the index. In the explorations, it would be preferable that the number of alternatives in a case range from three to at least nine, based on Saaty (1980). Lastly, an important limitation of this paper is the lack of reviews of indices other than the CR. In this paper, the review focuses solely on the CR among indices proposed to date because the CR has been the principal index; other prominent indices have not yet been reviewed. For the findings of reviews to be generalized, these indices need to be verified as well based on the CAR and the GWP.

Acknowledgments The authors are grateful to anonymous reviewers for their valuable comments and suggestions. This research was supported in part by the Chukyo University Grant for Overseas Research Program, 2019.

Appendix: Scatter plots of CI-CI_β (n=3, …, 9).

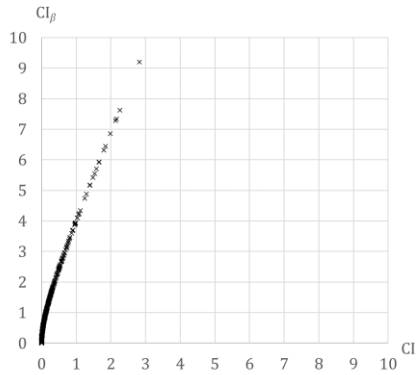

Fig. A1 n=3.

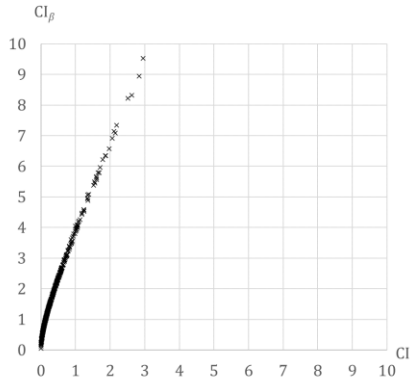

Fig. A2 n=4.

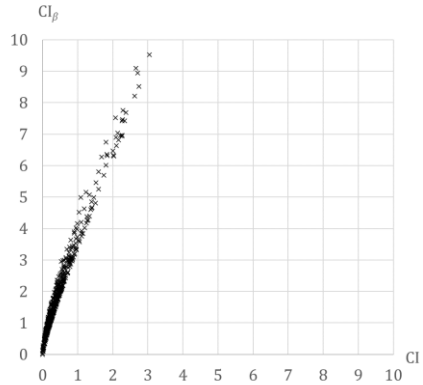

Fig. A3 n=5.

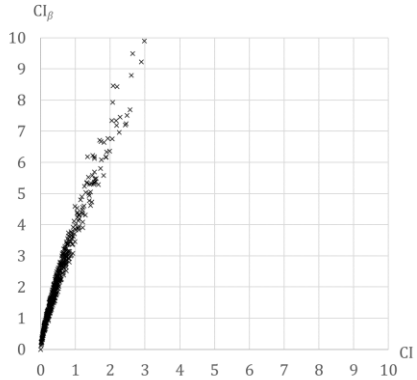

Fig. A4 n=6.

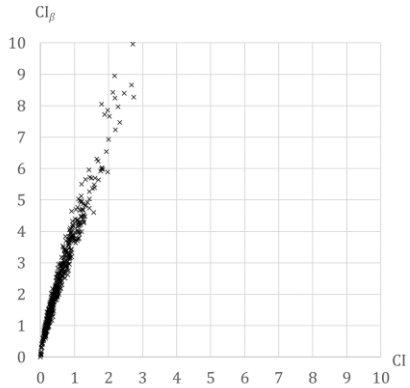

Fig. A5 n=7.

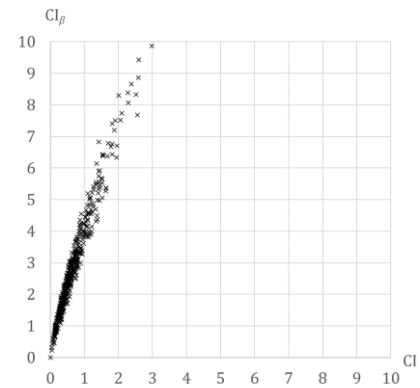

Fig. A6 n=8.

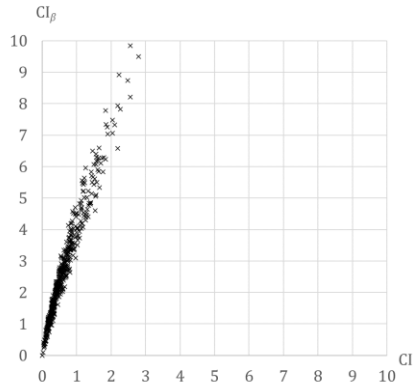

Fig. A7 n=9.

References

Alonso, J. A., & Lamata, M. T. (2005). A statistical criterion of consistency in the analytic hierarchy process. *Proceedings of the Second international conference on Modeling Decisions for Artificial Intelligence*, 67-76.

Barzilai, J. (1998). Consistency Measures for Pairwise Comparison Matrices. *Journal of Multi-Criteria Decision Analysis*, *7*(3), 123–132.

Brunelli, M. & Fedrizzi, M (2015). Axiomatic properties of inconsistency indices for pairwise comparisons. *Journal of the Operational Research Society*, *66*, 1–15.

Brunelli, M. (2017). Studying a Set of Properties of Inconsistency Indices for Pairwise Comparisons. *Annals of Operations Research*, *248*, 143-161.

Brunelli, M. (2018). A Survey of Inconsistency Indices for Pairwise Comparisons. *International Journal of General Systems*, *47*(8), 751-771.

Cavallo, B. & D'Apuzzo. L. (2009). A General Unified Framework for Pairwise Comparison Matrices in Multicriterial Methods. *International Journal of Intelligent Systems*, *24*(4), 377-398.

Chang, C. W., Wu, C. R., Lin, C. T., & Lin, H. L. (2007). Evaluating digital video recorder systems using analytic hierarchy and analytic network processes. *Information Science*, *177*, 3383-3396.

Chen, D. N., Hu, P. J. H., Kuo, Y. R., & Liang, T. P. (2010). A Web-based personalized recommendation system for mobile phone selection: design, implementation, and evaluation. *Expert System with Applications*, *37*(2), 8201-8210.

Crawford, G. & Williams, C. (1985). A Note on the Analysis of Subjective Judgment Matrices. *Journal of Mathematical Psychology*, *29*(4), 387–405.

Dey, P. K., Bhattacharya, A., & Ho, W. (2015). Strategic supplier performance evaluation: A case-based action research of a UK manufacturing organization. *International Journal of Production Economics*, *166*, 192-214.

Dong, Y., Hong, W. C., Xu, Y., & Yu, S. (2013). Numerical scales generated individually for analytic hierarchy process. *European Journal of Operational Research*, *229*(3), 654-662.

Duszak, Z. & Koczkodaj, W. W. (1994). Generalization of a New Definition of Consistency for Pairwise Comparisons. *Information Processing Letters*, *52*(5), 273–276.

expertchoice®. http://expertchoice.com/ Accessed 18 August 2020.

Fedrizzi, M. & Ferrari, F. (2018). A Chi-Square-Based Inconsistency Index for Pairwise Comparison Matrices. *Journal of the Operational Research Society*, *69*(7), 1125–1134.

Frobenius, G. (1912). Ueber Matrizen aus nicht negativen Elementen. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, 456-477.

Gass, S. I., & Rapcsák, T. (2004). Singular Value Decomposition in AHP. *European Journal of Operational Research*, *154*(3), 573–584.

Golden, B. L., & Wang, Q. (1989). An Alternate Measure of Consistency. In B. L. Golden, E. A. Wasil & P. T. Harker Eds.) *The Analytic Hierarchy Process, Applications and Studies*, (pp. 68–81). Springer-Verlag.

Grzybowski, A. Z. (2016). New Results on Inconsistency Indices and their Relationship with the Quality of Priority Vector Estimation. *Expert Systems with Applications*, *43*, 197–212.

Inglehart, R., & Abramson, P. (1993, September). Values and Value Change of Five Continents. Paper presented at the 1993 Annual Meeting of the American Political Science Association, Washington D.C.

Ishizaka, A., & Labib, A. (2014). A hybrid and integrated approach to evaluate and prevent disasters. *Journal of the Operational Research Society*, *65*(10), 1475-1489.

Kengpol, A., & Tuammee, S. (2016). The development of a decision support framework for a quantitative risk assessment in multimodal green logistics: An empirical study. *International Journal of Production Research*, *54*(4), 1020-1038.

Kou, G. & Lin, C. (2014). A Cosine Maximization Method for the Priority Vector Derivation in AHP. *European Journal of Operational Research*, *235*(1), 225–232.

Kułakowski, K. (2015). Notes on Order Preservation and Consistency in AHP. *European Journal of Operational Research*, *245*(1), 333–337.

Kułakowski, K. & Talaga, D. (2020). Inconsistency indices for incomplete pairwise comparisons matrices. *International Journal of General Systems*, *49*(2), 174-200.

Lootsma, F. A. (1991). *Scale sensitivity and rank preservation in a multiplicative variant of the Analytic Hierarchy Process*, Delft University of Technology, Report of the Faculty of Technical Mathematics, *91*(20).

Moktadir, M., Ali, S., Mangla, S., Sharmy, T., Luthra, S., Mishra, N. & Garza-Reyes, J. (2018). Decision modeling of risks in pharmaceutical supply chains. *Industrial Management & Data Systems*, *118*(7), 1388-1412.

Murphy, C. K. (1993). Limits on the analytic hierarchy process from its consistency index. *European Journal of Operational Research*, *65*, 138-139.

Perron, O. (1907). Zur Theorie der Matrices. *Mathematische Annalen*, *64*(2), 248-263.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, *15*, 234-281.

Saaty, T. L. (1980). *The Analytic Hierarchy Process*. New York, McGraw-Hill.

Saaty, T. L. (1997). That is not the analytic hierarchy process: what the AHP is and what it is not. *Journal of Multi-Criteria Decision Analysis*, *6*(6), 324-335.

Saaty, T. L. (2001a). *Fundamentals of Decision Making*. (2nd. Ed.). Pittsburgh, RWS Publications.

Saaty, T. L. (2001b). *The Analytic Network Process*. Pittsburgh, RWS Publications.

Saaty, T. L. (2005). *Theory and Applications of the Analytic Network Process*. Pittsburgh, RWS Publications.

Saaty, T. L. (2006). Rank from comparisons and from ratings in the analytic hierarchy/network processes. *European Journal of Operational Research*, *168*(2), 557-570.

Saaty, T. L. (2010). *Principia Mathematica Decernendi*. Pittsburgh, RWS Publications.

Salo, A. A., & Hämäläinen, R. (1995). Preference Programming through Approximate Ratio Comparisons. *European Journal of Operational Research*, *82*(3), 458–475.

Sato, Y., Tan, K. H. & Tse, Y. K. (2015). An Integrated Marginal Analysis Approach for Build-to-order Products. *International Journal of Production Economics*, *170*, 422-428.

Sato, Y., Tan, K. H. & Tse, Y. K. (2017). Investment Performance Analysis of Industrial Products: Case of an Effluent Processing Facility at a Chemical Company. *International Journal of Production Economics*, *194*, 52-58.

Shiraishi, S., Obata, T. & Daigo, M. (1998). Properties of a Positive Reciprocal Matrix and their Application to AHP. *Journal of the Operations Research Society of Japan*, *41*(3), 404–414.

Stein, W. E., & Mizzi, P. J. (2007). The harmonic consistency index for the analytic hierarchy process. *European Journal of Operational Research*, *177*(1), 488-497.

Szybowski, J., Kułakowski, K. & Prusak, A. (2020). New inconsistency indicators for incomplete pairwise comparisons matrices. *Mathematical Social Sciences*, *108*, 138-145.

Takeda, E. (1993). A Note on Consistent Adjustments of Pairwise Comparison Judgments. *Mathematical and Computer Modelling*, *17*(7), 29–35.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 278–286.

Yoon M.-S. (2020). Notes on the Use of Compatibility Index in the Analytic Hierarchy Process. *Proceedings of the International Symposium of the Analytic Hierarchy Process 2020*.

https://isahp2020.exordo.com/files/papers/53/final_draft/minsuk_yoon_submission_F.pdf/ Accessed 2 February 2021.

Zhang, H., Sekhari, A., Ouzrout, Y. & Bouras, A. (2014). Optimal Inconsistency Repairing of Pairwise Comparison Matrices Using Integrated Linear Programming and Eigenvector Methods. *Mathematical Problems in Engineering* doi.org/10.1155/2014/989726.

Formatted: Font: Italic

Formatted: Font: Not Italic

Formatted: Font: Italic