European Journal of Orthodontics 20 (1998) 569–578

The validation of an orthodontic expert system rule-base for fixed appliance treatment planning

Chris Stephens and Neil Mackin
Division of Child Dental Health, University of Bristol Dental School, UK

SUMMARY A peer review clinical trial was undertaken to assess the appropriateness of the advice produced by an expert system designed to plan orthodontic treatment in which the pre-adjusted bracket appliance was to be used. The results showed that the expert system's treatment plans were as reliable as those produced by a group of orthodontists. Two members of the panel actually ranked the expert system's plans more highly than their own.

Introduction

In Western countries, despite the increasing availability of fully trained specialists, approximately 50 per cent of all orthodontic treatment is carried out by general practitioners who lack any formal orthodontic qualification (Gorczyca *et al.,* 1989; Gravely, 1989). In the United Kingdom, where there are many fewer specialists than in neighbouring Northern Europe (Moss, 1993), much of this treatment, in the past, has been carried out using removable orthodontic appliances by those who have no more than a basic dental qualification.

Most UK general practitioners now recognize the limitations of removable orthodontic appliances (Gravely, 1989). However, at the present time there are insufficient numbers of fully trained orthodontists in the United Kingdom to provide fixed appliance therapy for all those children who need it, so many practitioners attempt to equip themselves with training in order to provide such treatment themselves. Usually, this training is technique-based, acquired on weekend courses, which in the view of the majority of teachers, is totally inadequate.

For this reason, over the past 10 years a number of 'integrated general practitioner training schemes' in orthodontics have been developed in the United Kingdom by groups of university teachers and National Health Service consultants modelled on that first described by Sullivan and DiBiase (1983) to provide small groups of practitioners with enhanced orthodontic skills which include the use of fixed appliances (Stephens *et al*., 1987). These are supported by the specialty (British Orthodontic Standards Working Party, 1986) and have proved popular (Power *et al*., 1995) and effective (Pender, 1988). However, experience has shown that although dentists who have completed these 2-year courses can provide an excellent standard of treatment, they still lack confidence in case selection and treatment planning. This restricts the number of patients they feel able to treat, and places an additional burden on the specialists who trained them and to whom they naturally turn for further advice.

The availability of a simple expert system designed to provide orthodontic advice to a general practitioner whose experience is limited to the use of removable orthodontic appliances has already been shown to be capable of guarding against many of the more fundamental errors which are made by general practitioners when planning treatment of Class II division 1 malocclusions (Sims-Williams *et al*., 1987; Brown *et al*., 1991; Stephens *et al*., 1996). However, to the present there has been no equivalent system to guard against similar errors when planning treatment for those cases where fixed appliance mechanics are intended.

To perform as an effective diagnostic aid for the clinician, a medical expert system must be easy to use and respond to the user quickly (Abbey, 1987). If the users are to be confident in and accept such a system then the ability to explain its reasoning is critical. Steethman (1995) argues convincingly that the precondition for the use of any clinical expert system is that it must have been thoroughly evaluated in a clinical setting. Furthermore, the knowledge encoded within the system should be generally accepted by experts within the field. It is therefore not surprising that expert systems often only relate to a relatively small field of knowledge.

Ortho Planner is a Windows-based program which was designed to support the planning of orthodontic treatment where a pre-adjusted fixed appliance was to be used. The development was undertaken jointly by the University of Bristol and a dental software company (Team Management Systems, Aylesbury, Buckinghamshire, HP22 5BL, UK). The program uses expert system techniques including rule-based reasoning (both forward and backward chaining) and fuzzy logic-based representations of orthodontic knowledge (Mackin, 1992). The advice which it generates consists of details of the treatment mechanics and extractions (referred to collectively, hereafter, as the 'treatment plan'), and associated clinical notes alerting the user to possible problems which might be encountered during treatment (referred to as 'notes of clinical guidance').

The purpose of this trial was to establish whether Ortho Planner was capable of providing advice which was acceptable to independent experts. The testing was divided into three parts:

1. *Verification that the rules within the system functioned as the designers intended*. This was tested by a controlled experiment in which the principal clinical expert involved in developing the software entered data from a large number of cases. The results were checked to confirm that the rules and algorithms devised for treatment planning had been applied correctly.
2. *Verification that the rule-base produced correct treatment plans* which is reported at length in this paper.
3. *Validation that a user could use the system to obtain acceptable advice*, which relates to a trial of software usability and will be the subject of a future paper.

Methods

There are two approaches which can be taken when assessing an expert system:

(1) the correct decision is evaluated in retrospect;
(2) the correct decision is defined as that which a human expert selects as the correct answer based upon the same information as the expert system.

Ethical problems (as well as time constraints) precluded the first approach whereby the system is tested on a sample of patients *in vivo* and the results assessed by examination of the sample at the end of treatment. Therefore, in order to use the second approach, it was necessary to define a 'gold standard' against which the system's output could be judged. The standard chosen was the collective opinion of recognized experts, the procedure adopted in the classic appraisal of MYCIN (Yu *et al*., 1979).

In the trial the following assumptions were made

1. It was considered more important to show that the system could be used successfully on routine cases which a general practitioner was likely to encounter rather than unusual and complex cases which would be likely to be referred to a fully trained orthodontist or hospital department.
2. In view of contemporary evidence regarding the long term effectiveness of deliberate labiolingual movement of the lower incisors (Mills, 1968; Houston and Edler, 1990), such treatment was regarded as 'complex' (as defined in 1 above).

An independent sample of 40 sets of case records was collected. These were of children judged to be at a stage of dental development where definitive treatment could be planned. These had

Table 1 The cases used in the study.

(a) Age pre treatment.					
10+	11+	12+	13+	14+	15+
2	8	11	12	5	2
(b) Incisor relationship.					
I	II.1	II.2	III		
10	24	1	5		
(c) Crowding.					
None	Mild	Moderate	Severe		
13	15	11	1		
(d) Subjective assessment of complexity.					
Low	Moderate	Severe			
11	19	10			

been drawn from a series of 60 consecutive cases referred for advice or treatment to the Orthodontic Department of Basingstoke District General Hospital, Hampshire. In essence this meant that the cases used were either in the late mixed or early permanent dentition stages. The records in each case comprised trimmed study models, panoramic tomographs, cephalometric radiographs/tracings, and a full face colour photograph.

The distribution of the ages of these patients, and the characteristics of the sample as judged by the system's main clinical expert (CS, the orthodontist who had worked most closely with the expert system rule-base development), are given in Table 1.

A panel consisting of 12 independent orthodontists was recruited. None were related to the project in any way and the group had been chosen to represent a reasonable spread of experience. All were drawn from Southern England to limit travel and accommodation costs. The panel was asked to attend on three occasions approximately 1 month apart for the purposes of:

(1) planning treatment for the 40 cases using a structured pro-forma (Figure 1);
(2) judging each other's treatment plans for the 40 cases;
(3) rejudging the treatment plans for a sub-sample of 10 cases which included a broad cross-section of treatment planning options in terms of extraction pattern and treatment mechanics.

The rejudging enabled the intra-examiner variability to be assessed. The whole panel of 12 orthodontists attended the first two sessions, but only nine were able to return for the third session.

In order to keep the trial as realistic as possible, the details of each session's work were only revealed to the participants at the start of each session.

Generating the computer-based treatment plans

One difficulty to be overcome when examining clinical expert systems relates to the data entry for the cases. The expert system encodes the knowledge of the system expert. It is quite possible that this system expert is biased in some of the clinical assessments, but has learned unconsciously to compensate for this by adjusting his/her treatment planning rules. This would mean that whilst the expert system would work well when the original expert entered data, it would probably provide inappropriate advice for other users. Therefore, in any validation trial, possible bias must be eliminated from the data entry. To achieve this data entry was prepared for every case by aggregating the clinical assessments made by panel members. For this reason each panel member was asked to complete a clinical assessment pro-forma for their sub-sample of cases in addition to planning treatment for them.

It became obvious to the authors that the clinical assessments made by panel members

Panel member [ ] **Case Indentifier** [ ]

Complexity of case

This case is suitable for treatment
- ○ within a hospital by a consultant
- ○ by a specialist practitioner
- ○ by a general practitioner with additional training from a hospital unit
- ○ by a general practitioner with only a basic undergraduate training

Treatment Method

This case should be treated using
- ○ Fixed appliances (inc RAs as needed)
- ○ Upper removable appliances only
- ○ No treatment required

Start with myofunctional ☐

Start with upper removable ☐
if so give reasons _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Extraction Pattern

Assuming all teeth are of good prognosis please indicate your choice of extractions

8 7 6 5 4 3 2 1	1 2 3 4 5 6 7 8
8 7 6 5 4 3 2 1	1 2 3 4 5 6 7 8

Lower Incisor Position

Do you intend to
- procline by > 5 degrees ○
- procline by < 5 degrees ○
- make no change ○
- retrocline < 5 degrees ○
- retrocline > 5 degrees ○

Anchorage requirement

Will you use a transpalatal arch N ○ Y ○

Will you use headgear N ○ Y ○ → If YES - How many hours
- < 8 hrs ○
- 8-10 hrs ○
- 10-12hrs ○
- 12-14hrs ○
- 14-16hrs ○
- > 16 hrs ○

Initial archwires

Upper ["] Partially engaged ☐

Lower ["] Partially engaged ☐

Which quadrants will you use canine lacebacks?

☐	☐
☐	☐

Elastics

What elastics do you intend to use?

☐	☐	Intra
☐	☐	Class II
☐	☐	Class III
☐	☐	Intra

Notes of Guidance

If this case had been referred to you for advice by a general dental practitioner, what cautions or advice would you have given? Note, these may well be a list of reasons as to why the case ought not to be treated by the practitioner.

Figure 1 The pro-forma used to record panel members' treatment planning decisions and their notes of guidance.

included two types of error. First, small differences in opinion and secondly large errors where, despite the accurate trimming of the models and the marking of the occlusion, it appeared that the observers had incorrectly articulated the models. An example of this second type of error would be where three examiners recorded a molar occlusion of Class I, but a fourth examiner entered an occlusion of a full-unit Class III. Merely averaging data such as these does not provide the best representation of the case. In order to eliminate such outliers for which there was a clear reason for suspicion, the panel's clinical assessments were aggregated as follows:

1. For parameters assessed as continuous attributes the distribution was plotted and any outliers were discounted by examination, the mean was then calculated.
2. For parameters collected as categorized attributes, the mode was used. Where the panel members' assessments were evenly split then the more severe category was taken as the datum input.

The aggregated data were then entered into the expert system and a further treatment plan proforma was completed for each of the cases using the advice generated by the expert system. Additionally, and quite independently, the system's main clinical expert also completed treatment plan pro-formas for each case.

Judging of treatment plans

All data entered onto treatment planning pro-formas were transcribed onto a computer database. For each case, the treatment plans and notes of clinical guidance were then printed off onto new standard pro-formas along with a unique reference number, omitting any reference to the plan's author. In other words, the identity of the originator of each plan was unknown to the participants and the organizers, but could be deduced subsequently by checking the reference number against the database.

The treatment plans in this form were then judged by the members of the panel during the second session, held 1 month later. At this the clinician was again presented with the pre-treatment records for each case as on the first occasion together with the anonymous proformas containing the transcribed details of the treatment plans and notes of guidance which had been devised for that case. In this assessment:

1. The treatment plans were scored according to the following:

(1) unacceptable—little chance of material benefit and/or the possibility of causing harm aesthetically or functionally;
(2) acceptable—a reasonable chance of conferring benefit and little risk of causing harm aesthetically or functionally;
(3) good—not necessarily the best treatment plan possible, but one with an excellent chance of conferring benefit both aesthetically or functionally;
(4) ideal—the best possible treatment for the case as it presented.

2. The notes of clinical guidance were judged by carrying out four separate assessments for 'relevance', 'correctness', 'clarity', and 'usefulness'. Each was scored according to the following:

(1) poor;
(2) mediocre;
(3) average;
(4) good;
(5) excellent.

3. The treatment plans for each case were then ranked from most appropriate to least appropriate, with equal scores being permitted.

Due to the constraints of time and the differences in the speed of working, it was not possible for all the orthodontists to plan treatment for every case, nor were all the treatment plans assessed by every orthodontist.

Results

In total, 533 orthodontist/case planning episodes were undertaken to produce treatment plans. These were subsequently assessed repeatedly

Table 2 The number of treatment plans and assessments undertaken per case.

Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Plans	14	14	13	13	15	14	13	14	12	13	12	14	14	13	14	13	12	14	13	12
Assessments	11	10	9	8	9	7	7	7	9	9	7	8	8	11	10	7	8	9	8	10
Case	21	22	23	24	25	26	27	28	29	30	31	33	33	34	35	36	37	38	39	40
Plans	14	12	14	13	13	13	14	13	13	13	13	13	14	14	13	13	15	13	14	13
Assessments	8	7	8	8	9	9	8	9	10	9	9	10	9	8	9	8	9	12	8	7

Table 3 The number of treatment plans and assessments undertaken by each orthodontist. CS refers to the system's main clinical expert whose plans were also included in the analysis, OP refers to Ortho Planner, the orthodontic expert system.

Orthodontist	A	B	C	D	E	F	G	H	I	J	K	L	CS	OP
Plans	37	39	38	36	38	33	36	35	36	39	39	37	40	40
Assessments	25	25	22	30	29	20	23	32	21	40	40	39	0	0

in 346 orthodontist/case assessments. Tables 2 and 3 show the distribution of treatment plans and assessments between the orthodontists. It is particularly striking that because the assessments were more time-consuming than the treatment planning there is a greater range in the number of assessments undertaken.

Table 4 The aggregated intra-examiner agreement between the first and second assessments of treatment mechanics made by each orthodontist on the treatment plans of a sub-sample of cases.

	Unacceptable	Acceptable	Good	Ideal
Unacceptable	137	71	20	3
Acceptable	79	238	113	25
Good	10	140	75	8
Ideal	4	51	34	6

Table 4 summarizes the repeated assessments aggregated over all nine orthodontists. Table 5 shows the weighted and unweighted Kappa statistic calculated to show the intra-examiner agreement for each orthodontist. These indicate that overall the orthodontists had a moderate level of agreement between their two assessments (Landis and Koch, 1977).

Table 6a shows all the assessments of treatment plans in terms of the categories awarded by the panel members averaged for each treatment plan author, be that a member of the panel, the system's main clinical expert or the expert system. This most basic analysis shows that the expert system was approximately mid-ranking amongst the peer group.

The trial had been carefully conducted during the assessment phase so as to minimize the possibility of an assessor recognizing that he/she was

Table 5 The unweighted and linearly weighted kappa statistics of intra-examiner agreement for each orthodontist over the sub-sample of cases.

Orthodontist	H	G	E	F	C	D	J	A	I
Unweighted	0.495	0.455	0.488	0.438	0.471	0.458	0.471	0.388	0.351
Weighted	0.801	0.791	0.785	0.782	0.782	0.773	0.771	0.769	0.721

assessing his/her own treatment plan. However, subsequently it came to light that some members of the panel believed that they had recognized their own work. For example, one clinician used an unusual choice of initial archwire in some of the cases which no other orthodontist had specified. For this reason it was decided to rework the analyses excluding the 318 assessments of plans by their own authors and using only the 4076 assessments made by orthodontists of plans they had not authored (Table 6b). Note that this process does not affect the average assessment of CS, the system's main clinical expert, at 2.33, nor that of OP, the expert system, at 2.17 since neither had taken part in the assessment process.

As explained above, due to time constraints, not all the participating orthodontists completed the intended number of treatment plans and assessments, and furthermore those that were completed were not randomly and evenly spread amongst the patients. Therefore, the trial failed to fulfil the requirements of being a balanced incomplete block design experiment with the patient as the blocking variable (Fleiss, 1986). In order to compensate for this shortcoming, normalization was undertaken. In this process the assessments made by each assessor were averaged. It is argued that the differences between support given by each assessor to a group of treatment plans could be accounted for in two ways:

1. A systematic difference in the level of criticism used by each assessor.
2. Underlying differences amongst the subgroups of plans assessed by each assessor.

It was assumed that since each assessor had considered the majority of the treatment plans, the difference in average assessments made related wholly to (1). Therefore the assessments of each orthodontist were normalized by weightings so that when the weighted assessments were averaged over each assessor they would equate. The derivation and values of these weightings are presented in the Appendices.

Table 6c shows all the normalized assessments of treatment mechanics, averaged for each

Table 6 The assessments of treatment mechanics.

(a) Averaged over the treatment plans' authors, and presented in descending order. CS refers to the system's main clinical expert whose plans were also included in the analysis, OP refers to Ortho Planner, the orthodontic expert system.

Orthodontist	F	A	CS	J	G	K	C	H	E	OP	L	C	I	B
Average score	2.36	2.33	2.33	2.31	2.22	2.21	2.19	2.18	2.18	2.17	2.16	2.12	2.09	1.80

(b) As above, excluding self-assessments.

Author	CS	F	A	J	K	OP	G	E	L	D	C	H	I	B
Average score	2.33	2.30	2.29	2.20	2.18	2.17	2.16	2.16	2.13	2.13	2.12	2.08	2.06	1.79

(c) Averaged normalized assessments excluding self-assessments.

Author	CS	A	F	J	OP	G	K	E	L	H	D	C	I	B
Average score	2.32	2.28	2.27	2.24	2.17	2.17	2.16	2.14	2.14	2.14	2.10	2.09	2.05	1.76

treatment plan author, be that a member of the orthodontist panel, the system's main clinical expert or the expert system. It can be seen that this normalization process made little difference to the resulting values.

The qualitative assessments of the notes of clinical guidance are presented collectively in Table 7. These were judged in terms of relevance, correctness, clarity and usefulness. Again the normalized assessments were averaged for each treatment plan author and self-assessments were excluded (cf. Table 6c).

Finally the assessors were asked to rank the treatment plans in order of preference for each case. These rankings were also averaged over all treatment plan authors, equal ranking being handled by taking the mean of the equal positions. Table 8 presents the orthodontists ordered by their averaged rankings (normalized as above), though no actual values are given (these, being purely relative, are of no intrinsic interest).

Discussion

The analysis of the assessments of treatment plans and overall rankings indicate that Ortho Planner performed in the top half of the peer group (Table 8).

The expert system also performed very well in terms of the notes of clinical guidance which it generated (Table 7). This may arise from the underlying thoroughness of using a rule-base approach to generating recommendations. The expert system would painstakingly consider all possible clinical difficulties or consequences, whereas the human clinician is more likely to focus upon one or two difficulties, and then fail to consider the remaining ones. This is highly supportive of the hypothesis that using expert systems is beneficial to user education (Abbey, 1987).

The overall level of approval given by peer group panel members to their colleagues' treatment decisions was surprisingly low. For example the average assessment for treatment mechanics as shown in Table 6a was 2.19 which if interpreted literally would imply that the peer group judged their own work to be barely acceptable. The authors are inclined to believe that the peer group of orthodontists were overly critical of their own work and that had the collection of treatment plans they were asked to judge been drawn from a broader cross-section of clinicians then the average assessments of their own work would have been higher.

The assessment of notes of clinical guidance (Table 7) shows a very high correlation between the judgments of relevance, correctness, clarity and usefulness; $r > 0.97$ (Pearsonian) for all combinations. In hindsight, the authors believe

Table 7 The averaged normalized assessments of notes of clinical guidance provided by each author to accompany their treatment plans, judged in terms of their relevance, correctness, clarity, and usefulness averaged over the treatment plans' authors. The ranking (averaged all over categories) is in descending order from left to right.

Author	G	E	OP	A	CS	H	C	I	D	F	L	B	J	K
Relevance	2.78	2.69	2.69	2.67	2.63	2.61	2.50	2.46	2.39	2.38	2.23	2.25	2.17	1.90
Correctness	2.77	2.72	2.64	2.65	2.66	2.59	2.51	2.45	2.38	2.38	2.31	2.25	2.21	1.91
Clarity	2.88	2.75	2.72	2.76	2.72	2.69	2.56	2.45	2.46	2.44	2.43	2.34	2.33	2.04
Usefulness	2.69	2.60	2.55	2.54	2.50	2.49	2.39	2.33	2.30	2.26	2.13	2.13	2.03	1.85

Table 8 Treatment plan authors ranked according to the preferences of their colleagues. Normalized rankings have been used and self-assessments excluded.

Position	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th	14th
Author	A	CS	F	G	H	J	OP	E	K	C	D	L	I	B

that the trial did not benefit from this high level of discrimination and that a broader overall consideration of this information would have been more time efficient.

In comparing the rankings given by each assessor of their own plans and those of the expert system on a case by case basis, it was noted that overall two orthodontists actually preferred the expert system's plans to their own.

The difference between the assessments of the system's main clinical expert and the expert system may be explained in two ways:

(1) the expert system is as yet a far from perfect expression of the knowledge of the system's main clinical expert. This may be overcome by extending the expert system's rule-base.
(2) the data input entered into the expert system forms an incomplete subset of the data used unconsciously by the system's main clinical expert when planning treatment. This may be addressed by enhancing the interface to the expert system, for example by using vision algorithms to assess automatically dental crowding from the study cast.

Conclusions

In this study, the treatment plans generated by an expert system were judged to be of similar quality to those of the orthodontist. The notes of clinical guidance provided by the expert system often surpassed those of its human counterparts. It remains to be seen whether general practitioners using the expert system will be able to generate treatment plans of the same level of acceptance as their orthodontic colleagues.

Address for correspondence

Professor C. D. Stephens
Division of Child Dental Health
University of Bristol Dental School
Lower Maudlin Street
Bristol BS1 2LY
UK

Acknowledgements

The authors would like to thank Team Management Systems Limited for their kind support in the trial and the following clinical experts who took part in the panel: Mr G. W. Butcher, Mr J. Catcheside, Mr C. R. Charles, Mrs M. R. Evans, Dr A. J. Ireland, Miss H. Knight, Mrs K. M. Postlethwaite, Mr R. T. Reed, Mr R. J. Robinson, Miss A. J. Smith, Mr S. Weerakone and Mr S. W. Wigglesworth.

References

Abbey L M 1987 An expert system for oral diagnosis. Journal of Dental Education 51: 475–480

British Orthodontic Standards Working Party 1986 2nd Report. British Journal of Orthodontics 13: 165–174

Brown I D, Adams S R, Stephens C D, Errit S J, Sims-Williams J H 1991 The initial use of a computer controlled expert system in the treatment planning of Class II division 1 malocclusion. British Journal of Orthodontics 18: 1–7

Fleiss J L 1986 The design and analysis of clinical experiments. Wiley, New York

Gorczyca A M, Jones J E, Douglass C W 1989 Orthodontic treatment provided by general practitioners and pedodontists in Massachusetts. Journal of Clinical Orthodontics 23: 346–352

Gravely J 1989 Who should practise orthodontics? British Journal of Orthodontics 16: 235–241

Houston W J B, Edler R 1990 Long term stability of the lower labial segment relative to APo. European Journal of Orthodontics 12: 302–310

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

Mackin N 1992 The development of an expert system for planning orthodontic treatment. PhD Thesis, University of Bristol, UK

Mills J R E 1968 The stability of the lower labial segment. Transactions of the British Society for the Study of Orthodontics, pp. 11–24

Moss J P 1993 Orthodontics in Europe. European Journal of Orthodontics 15: 393–401

Pender N 1988 The Four Thames project: development of training in orthodontics for dental practice. British Journal of Orthodontics 15: 269–275

Power S M, Hodgkins J F W, Stephens C D 1995 An evaluation of the views of general dental practitioners who have participated in an extended orthodontic training scheme. British Journal of Orthodontics 22: 159–166

Sims-Williams J H, Brown I D, Matthewman A, Stephens C D 1987 A computer-controlled expert system for orthodontic advice. British Dental Journal 163: 161–166

Steethman E S 1995 Formalizing diagnostic thinking in oral radiology. Vrije University Press, Amsterdam

Stephens C D, Hathorn I S, Brown I D, Harradine N H 1987 The SW/Wessex orthodontic training course for general practitioners. Dental Practice 25: 1–3

Stephens C D, Mackin N, Sims-Williams J H 1996 The development and validation of an orthodontic expert system. British Journal of Orthodontics 23: 1–7

Sullivan P, DiBiase D D 1983 The role of fixed appliances in general practice. British Dental Journal 155: 167–170

Yu V L *et al.* 1979 An analysis of MYCIN's advice. Journal of the American Medical Association 242: 1279–128

Appendix 1

Let the set of n orthodontists be S:{$S_1,\ldots,S_n$} who assessed the set of m cases C:{$C_1,\ldots,C_m$} such that A_{ij} was the assessment of orthodontist S_i upon case C_j.

Let the sub-group of cases assessed by S_i be denoted $C^S{}_i$ such that $C^S{}_i \subset C$.

Let the normalization weightings W:{$W1,\ldots,W_n$} be derived such that each orthodontist exhibits the same level of criticism over their respective sub-group of cases.

i.e. for orthodontist S_i,

$$\sum_{\forall C_j \in C_i^S} W_i A_{ij} = \frac{1}{n}\sum_{k=1}^{n} \sum_{\forall C_l \in C_k^S} A_{kl}$$

$$\therefore W_i = \frac{\sum_{k=1}^{n} \sum_{\forall C_l \in C_k^s} A_{kl}}{n \sum_{\forall C_j \in C_i^s} A_{ij}}$$

Where $\forall$ means *for all*, $\in$ means *is a member of*, $\subset$ means *is a sub-group of*.

Appendix 2

The weighting values derived from the formula in Appendix 1

Orthodontist	A	B	C	D	E	F	G	H	I	J	K	L
Mechanics	0.998	1.667	1.088	1.207	1.048	1.097	0.943	0.790	1.124	0.852	1.028	0.961
Relevance	1.029	1.099	1.147	1.020	1.004	0.994	0.934	0.719	0.955	0.910	0.921	1.545
Correctness	0.979	1.125	1.121	0.992	0.989	1.003	0.941	0.759	0.929	0.911	0.925	1.551
Clarity	1.013	1.178	1.130	0.998	0.977	0.934	0.954	0.782	0.785	0.946	0.936	1.595
Usefulness	1.045	1.131	1.117	0.946	0.983	0.980	0.905	0.791	1.077	0.876	0.923	1.475