Your order details
Our Order Ref: 00282545-013
Your Ref: 114923
Despatched on: 2/4/2013

Your shipping address
Kings College London
null

Your item details
UIN: BLL0109555698
Title: Journal of dental research.
Publisher: Baltimore, Md. : Journal of Dental Research, [1919-
ISSN: 0022-0345
Year: 1980 Volume: 59
Pages:
Author name(s):
Article title words: The inter-examiner variation in rating
dental formation form radiographs
Levesque

Comments
Visit our blog for all the latest news and find out more about our new service at http://britishlibrary.typepad.co.uk/bldss/

Thank you for using Document Supply Services!

Kings College London
null

S483

Copyright Statement
Unless out of copyright, the contents of the document(s) attached to or accompanying this page are protected by copyright. They are supplied on condition that, except to enable a single paper copy to be printed out by or for the individual who originally requested the document(s), you may not copy (even for internal purposes), store or retain any electronic medium, retransmit, resell, hire or dispose of for valuable consideration and of the contents (including the single paper copy referred to above). However these rules do not apply where:
1. you have written permission of the copyright owner to do otherwise; 2. you have the permission of The Copyright Licensing Agency Ltd, or similar licensing body; 3. the Copyright Act 1956 applies to the intended usage; 4. the intended usage is covered by statute.

Breach of the terms of this notice is enforceable against you by the copyright owner or their representative.

This document has been supplied under our Library Privilege service. You are therefore agreeing to the terms of supply for our Library Privilege service, available at:

The British Library, Document Supply, Boston Spa, Wetherby, United Kingdom, LS23 7BQ
www.bl.uk
The Inter-examiner Variation in Rating Dental Formation from Radiographs

GILLES-Y. LEVESQUE and ARTO DEMIRJIAN

Human Growth Research Centre, Université de Montréal, 2801 Edouard-Montpetit, Montréal, Québec Canada, H3T 1J6

Six examiners evaluated dental formation on 191 radiographs, using an eight-stage system. Discrepancies between these examiners are generally of the order of ± one stage, in about 20 to 25% of the cases. The authors thus suggest the use of reference radiographs, for the double purpose of ensuring a homogeneous rating by two examiners or more, as well as the assessment of eventual differences between populations.


Introduction.

Several systems have been proposed, using the radiographic approach for the assessment of dental development.\(^1\)\(^-\)\(^7\) These systems divide dental development into various numbers of stages, from three\(^8\) to 20.\(^4\) Although very diversified, they present one common difficulty in the precise reproduction of the results obtained. This brings up the problem of reliability in the rating of dental development, whether one person is responsible for two examinations or more (intra-examiner reliability), or whether the examinations were performed by two or more persons (inter-examiner reliability).

To this day, no results have been specifically published on the reliability of the rating of dental development from X-rays. Some information is provided by Fanning,\(^4\) Grøn,\(^9\) and Sapoka and Demirjian\(^10\) on the intra-examiner reliability, and by Demisch and Wartmann,\(^11\) Hotz,\(^12\) Grøn,\(^9\) and Demirjian et al.\(^7\) on the inter-examiner reliability. Furthermore, in all these studies, only the percentage of discrepancies between examinations or examiners is mentioned, along with the maximum difference in the number of stages.

Our purpose here is to study the inter-examiner reliability (with six examiners, in this case), since it shows more frequent differences than the intra-examiner approach, thus bringing out the most disadvantageous situation possible. We worked on several sources of variation, namely, the type of evaluation of the dental development (longitudinal or cross-sectional), four different age groups, and seven mandibular teeth.

Materials and methods.

The basic material used in this report consists of panoramic radiographs (Panorex) of school-age children who participated in a previous longitudinal study. The collection of the data was carried out between 1967 and 1976 at the Human Growth Research Centre of the Université de Montréal; 191 radiographs from 84 girls were selected in the following groups: 42 at six years, 50 at eight years, 49 at ten years, and 50 at 12 years. These ages covered the most active period of the dental formation. For 39 of the girls, the sample included three or four radiographs (70% of the material); these serial records were completed with one or two radiographs from the remaining 45 girls, mainly at age six. Six examiners evaluated the radiographs independently, according to the system devised by Demirjian et al.,\(^7\) which includes eight stages of dental formation for each of the seven left mandibular teeth. The stages generally encountered from six to 12 years are: C to G for M\(_2\) and PM\(_2\), D to H for PM\(_1\) and canine, E to H for I\(_2\), and F to H for M\(_1\) and I\(_1\). Four of the examiners (E3 to E6), who were not familiar with the technique, were trained by rating about 100 similar radiographs; the other two (E1 and E2) had already mastered the rating technique. A cross-sectional rating was done by five of the examiners (E1 and E3 to E6), that is, they evaluated only one age group at the time, while the E2 examiner pro-
ceed longitudinally. Because of the nature of the data, a non-parametric test based on ranks (Friedman’s ANOVA) was used for the simultaneous comparison of the ratings.

**Results.**

The two-by-two comparison of the examiners. — Since the E1 examiner was the only one who was experienced and had evaluated the X-rays cross-sectionally, we first compared his results with those of each other examiner by calculating the differences between the evaluations for each tooth. The mean differences and the proportion of zero, negative, and positive differences are presented in Table 1, where a positive difference indicates a higher rating by the E1 examiner than by the compared one.

The mean differences are generally of the order of ±0.1 stage or less, with the exception of those between the E1 and E2 examiners. In the latter case, for five out of seven teeth, the mean difference is of at least + 0.2 stages, reaching half a stage for the canine (Table 1). The percentage of discrepancies between examiners lies between 11 and 29%, but it is higher (38% and 48%) for the canine, as evaluated by the E3 and E2 examiners. When considering each tooth individually, we note that the proportion of discrepancies between examiners E1 and E2 is the highest, with the exception of the second premolar. Individual differences, whether positive or negative, are of one stage, except in 16 cases only, where such a difference reaches two stages. The positive and negative differences are usually of the same magnitude, but those between the E1 and E2 examiners are more often positive than negative (Table 1).

The longitudinal evaluation of the E2 examiner seems to differ, on the whole, from the cross-sectional ones performed by the other examiners.

The simultaneous comparison of the examiners. — The evaluations of six examiners for each tooth and each age group were compared according to Friedman’s test. They were found equivalent (P > .01) for the posterior teeth (molars and second premolar), with the exception of M1, at ten years of age, for which we note that \( \chi^2(5) = 30.8 \) and \( P < .001 \). However, for the our remaining teeth, the evaluations did not prove similar (P < .01) in most of the cases. For example, when considering the lateral incisor at eight years of age, we find \( \chi^2(5) = 22.5 \) and \( P < .001 \); thus, it seems that at least one of the examiners produced a different evaluation. Bearing in mind the results of the two-by-two comparisons between examiners, the E2 examiner would have a tendency to underestimate the stage reached within the longitudinal evaluation (Table 1). Therefore, we repeated the simultaneous comparisons, but with only the five examiners who proceeded cross-sectionally.

The differences between the cross-sectional evaluations of the five examiners are generally minor (Table 2) with, however, the exception of the second molar, first premolar, and central incisor at six years of age, and that of the first molar at ten years. These discrepancies and the corrections we provided will be discussed later.

**Discussion.**

The proportion of discrepancies between two examiners for molars and incisors is around 20% — the majority of values lying between 15 and 25%. This proportion is higher for premolars (25%) and canine (about 30%). Demisch and Wartmann obtained a close value of 22% with two examiners and a seven-stage system. On the other hand, Grøn and Demirjian et al. showed 8% (with the exception of the maxillary central incisor) and 16% of discrepancies, respectively, which is clearly below our observations, but may be explained through particular conditions leading to a higher reliability of the inter-examiner data. Grøn used only four stages of development, and Demirjian et al. produced daily comparisons of the evaluations of four examiners. Generally, the reliability of repeated evaluations by one examiner is greater than that between examiners; such repeated evaluations differ in only 10% of the cases, according to Grøn and Sapoka and Demirjian. Nevertheless, Fanning reports 27% of discrepancies for premolars and incisors. This low intra-examiner reliability value would be linked to the use of an evaluation system that includes a high number of developmental stages, 20 in this case.

The differences between examiners E1 and E2 are very unevenly distributed be-
TABLE 1
MEAN DIFFERENCES, AND DISTRIBUTION OF DIFFERENCES, WHEN COMPARING E2 TO E6 EXAMINERS TO E1

<table>
<thead>
<tr>
<th>Teeth</th>
<th>E1 and E2 Mean Diff. (1)</th>
<th>Diff. (%) (2)</th>
<th>E1 and E3 Mean Diff.</th>
<th>Diff. (%)</th>
<th>E1 and E4 Mean Diff.</th>
<th>Diff. (%)</th>
<th>E1 and E5 Mean Diff.</th>
<th>Diff. (%)</th>
<th>E1 and E6 Mean Diff.</th>
<th>Diff. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2</td>
<td>.12</td>
<td>77 6 17</td>
<td>.03 87 5</td>
<td>8</td>
<td>.00 79 11 11</td>
<td>.05 84</td>
<td>5 11</td>
<td>.02 81</td>
<td>9 11</td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>.22</td>
<td>76 1 23</td>
<td>-.02 87 8</td>
<td>6</td>
<td>-.05 87 9 4</td>
<td>.12 79</td>
<td>5 17</td>
<td>.10 81</td>
<td>5 15</td>
<td></td>
</tr>
<tr>
<td>PM2</td>
<td>.04</td>
<td>79 8 13</td>
<td>-.07 70 18</td>
<td>12</td>
<td>-.03 75 14 11</td>
<td>-.12 72</td>
<td>19 8</td>
<td>-.14 72</td>
<td>21 7</td>
<td></td>
</tr>
<tr>
<td>PM1</td>
<td>.24</td>
<td>72 2 26</td>
<td>.08 75 8</td>
<td>16</td>
<td>.04 80 8 12</td>
<td>-.09 77</td>
<td>16 7</td>
<td>-.07 80</td>
<td>14 6</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>.47</td>
<td>52 1 47</td>
<td>.08 63 15</td>
<td>23</td>
<td>-.07 78 15 8</td>
<td>.00 75</td>
<td>13 13</td>
<td>.01 80</td>
<td>9 11</td>
<td></td>
</tr>
<tr>
<td>I2</td>
<td>.23</td>
<td>71 3 26</td>
<td>-.03 86 8</td>
<td>6</td>
<td>-.08 89 9 2</td>
<td>-.06 83</td>
<td>12 6</td>
<td>-.05 78</td>
<td>13 9</td>
<td></td>
</tr>
<tr>
<td>I1</td>
<td>.20</td>
<td>79 1 21</td>
<td>-.02 87 7</td>
<td>5</td>
<td>-.12 85 13 2</td>
<td>-.03 86</td>
<td>8 5</td>
<td>.06 89</td>
<td>3 8</td>
<td></td>
</tr>
</tbody>
</table>

(1) Difference (E1 examiner - other examiner), expressed in number of stages.
(2) Zero, negative or positive differences (either of one or two stages)

Note: the ratings of examiner E2 are done longitudinally.

TABLE 2
SIMULTANEOUS COMPARISONS (FRIEDMAN’S TEST) OF FIVE CROSS-SECTIONAL EVALUATIONS

<table>
<thead>
<tr>
<th>AGE</th>
<th>M2 $x^2$ (4) P</th>
<th>M1 $x^2$ (4) P</th>
<th>PM2 $x^2$ (4) P</th>
<th>PM1 $x^2$ (4) P</th>
<th>C $x^2$ (4) P</th>
<th>I2 $x^2$ (4) P</th>
<th>I1 $x^2$ (4) P</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>13.3 .010</td>
<td>4.2 .38</td>
<td>6.0 .20</td>
<td>13.8 .008</td>
<td>7.4 .12</td>
<td>2.1 .72</td>
<td>16.7 .002</td>
</tr>
<tr>
<td>8</td>
<td>3.1 .53</td>
<td>0.37 .99</td>
<td>8.4 .079</td>
<td>2.5 .64</td>
<td>8.4 .079</td>
<td>10.0 .040</td>
<td>6.2 .19</td>
</tr>
<tr>
<td>10</td>
<td>3.3 .50</td>
<td>19.4 .001</td>
<td>3.9 .41</td>
<td>9.9 .042</td>
<td>1.5 .83</td>
<td>3.5 .47</td>
<td>0.41 .98</td>
</tr>
<tr>
<td>12</td>
<td>2.0 .74</td>
<td>0.05 ~1</td>
<td>2.9 .57</td>
<td>3.0 .55</td>
<td>11.1 .026</td>
<td>0.04 ~1</td>
<td>0 1</td>
</tr>
</tbody>
</table>

between positive and negative values, which indicates a strong tendency to underestimate the attained stage in the longitudinal approach of E2, as compared to the cross-sectional one of E1. The general bias thus introduced, the importance of which the simultaneous comparison of the six examiners confirms, is not present in cross-sectional evaluations, except in very few cases, mainly at age six. Because we had only one examiner who rated longitudinally, at this stage we cannot state the real reason for these differences in the discrepancies between the longitudinal and the cross-sectional ratings. They might be due to the examiners or the material (longitudinal and cross-sectional). The significant differences between the cross-sectional evaluations would be derived mainly from the fact that the examiners, including those without experience, began their evaluations with X-rays of six-year-old girls; moreover, the most difficult stages for evaluation are frequently reached at this age. Thus, delimitation between stages C and D requires the investigation of many radiographs, and furthermore, the examiners did not evaluate this latter stage during their training period, which might explain the difference obtained at six years of age for the second molar. On the other hand, the evaluations of the first premolar at the same age differ, according to the examiners, because of a partly imperfect formulation of the description of stage E (Ref. 7, p. 222). Consequently, it was rewritten as follows:

“b. The root length reaches at least 1/3 of the crown height.”

This definition applies to uniradicular teeth and to molars. It replaces the former one, that read:

“b. The root length is still less than the crown height.”

The other two important differences mentioned are more difficult to explain. For the first molar, at ten years of age, the
problem lies with the delimitation between stages G and H, while it involves stages F and G for the central incisor at six years. Since the latter stages were frequently attained for all teeth of the present sample, we cannot attribute the result obtained for the first molar at ten years to a definite factor. Furthermore, the general cause discussed above could explain the difference observed at six years for the central incisor.

All of the previously mentioned authors report a maximum difference of one stage between two evaluations for all teeth, whether evaluated by one or two examiners. Nearly all the discrepancies that we encountered are of one stage, but we have also observed a small number of cases (1.7%) in which the differences were of two stages; all the examiners noted their presence at all ages and for six of the seven teeth that were examined. A closer study of such cases reveals that they can be partly, if not totally, attributed to transcription errors.

Our results emphasize the necessity of verifying the evaluations of two or more examiners involved in a study. They also stress the need for such cross-checking when comparing populations evaluated by different persons using the same system. One suggested procedure would be the use of a sequence of reference radiographs. One could resolve the problems of rating in this manner before starting the evaluation of a given sample. Another method of reducing discrepancies between examiners would be the establishment of a systematic cross-checking procedure.

Conclusions.

The proportion of discrepancies between evaluations of molars and incisors is generally 15 to 25%, which leaves a percentage of inter-examiner reliability of 80%; the reliability is lower for premolars (75%) and canine (about 70%).

The use of a series of reference radiographs would be a sound way of detecting and correcting the differences between one or more examiners and a typical rating. If such a system is adopted, the eventual differences between populations could be easily achieved.

Acknowledgments.

The authors would like to thank the personnel of the Human Growth Research Centre for their co-operation, and especially Miss M. Boivin, Miss A. Charbonneau, Drs. L. Paras, D. Dilancea, C. Lamarche, and S. Toma, who were responsible for the radiographic evaluations; and finally, Miss L. Pinsonneault, who prepared the manuscript.

REFERENCES


*Duplicates of part of the radiographs used in this study, along with our typical rating, are available at a nominal cost through the Human Growth Research Centre.