Evaluation of Accuracy, Reliability, and Repeatability of Five Dental Pulp Tests

Eugene Chen, BDSc, BScDent(Hons), DClinDent, and Paul V. Abbott, BDS, MDS

Abstract

Introduction: The aim of this study was to compare the clinical accuracy, reliability, and repeatability of laser Doppler flowmetry (LDF), an electric pulp test (EPT), and various thermal pulp sensibility tests. Methods: Pulp tests were done on 121 teeth in 20 subjects by using LDF, EPT, and thermal pulp testing (CO₂, Endo Frost [EF], Ice) during 2 or 3 test sessions with at least 1-week intervals. The order of testing was reversed on the second visit. A laser Doppler flowmeter was used to measure mean pulp blood flow (Flux) calibrated against a brownian motion medium and zeroed against a static reflector. The laser source was 780 nm, with 0.5-mm fiber separation in the probe, 3.1 kHz as the primary bandwidth for filter set to 0.1-second time output constant. Customized polyvinylsiloxane splints were fabricated for each participant, and a minimum of 90-second recording time was used for each tooth. Raw data were analyzed by using repeated measure analysis of variance, pairwise comparisons, and inter-class correlations (ICC). Results: The accuracy of EPT, CO₂, and LDF tests was 97.7%, 97.0%, and 96.3%, respectively, without significant differences (P > .3). Accuracy of EF and Ice was 90.7% and 84.8%, respectively. EPT (P = .015) and CO₂ (P = .022) were significantly more accurate than EF. LDF was more accurate than EF, but this was not statistically significant (P = .063). Ice was significantly less accurate than EPT (P = .004), CO₂ (P = .005), LDF (P = .006), and EF (P = .019). With the exception of Ice (effect of visit: F₂,38 = 5.67, mean squared error = 0.01, P = .007, \( \eta^2_p = 0.23 \)), all tests were reliable. Ice (ICC = 0.677) and LDF (ICC = 0.654) were the most repeatable of the tests, whereas EPT (ICC = 0.434) and CO₂ (ICC = 0.432) were less repeatable. Conclusions: CO₂, EPT, and LDF were reliable and the most accurate tests, but CO₂ and EPT were less repeatable yet less time-consuming than LDF. EF was reliable but not as accurate as EPT and CO₂ and less repeatable than Ice and LDF. Ice was the most repeatable but the least accurate and least reliable test. (J Endod 2011;37:1619–1623)

Key Words

Electric pulp tests, laser Doppler flowmetry, pulp sensibility tests, thermal pulp tests, vitality tests

Cold tests and electric pulp tests (EPTs) are pulp sensibility tests that have been established as useful aids in the assessment of pulp vitality (1–3), despite being subjective and patient-dependent (4, 5). Nonetheless, the lack of pulp sensibility is often associated with advanced pulp necrosis (6).

Pulp sensibility tests have limitations, and false responses can occur (7). EPTs are known to be unreliable in immature teeth (3, 8) and in teeth undergoing orthodontic movement (3). Furthermore, false responses are known to occur when the electric current is conducted to adjacent periodontal tissues, adjacent teeth (9), and even to remnant inflamed pulp tissue with liquefactive pulp necrosis (2). In traumatized teeth, EPTs are less reliable than cold tests (7, 10).

Cold tests are known to elicit false-negative responses in elderly patients because of the amount of thermal insulation provided by secondary dentin (2). Commonly available agents for cold tests are CO₂ snow, ice, and refrigerant sprays (eg, tetrafluoroethane, butane, propane, isobutane, dichlorofluoromethane [DDM], and ethyl chloride).

Ethyl chloride was popular (11), but CO₂ snow and refrigerant such as DDM have been proved effective and even superior (6, 12) than ice and ethyl chloride. DDM has decreased in popularity because of environmental concerns (13), and it has been replaced by other gases such as the mixture of propane, butane, and isobutane (Endo Frost [EF]; Roeko, Langenburg, Germany). Concerns have been raised about possible damaging effects of cold testing agents to teeth, but studies have shown that they are safe (14–18).

Research into the application of laser Doppler flowmetry (LDF) in traumatized pulps has been extensive (19–22). The general aim has been to objectively measure the true vitality of the pulp, that is, the blood supply rather than the sensory function.

There have been few direct comparisons reported between LDF and other methods of pulp testing. Ingólfssson et al (23) compared LDF with EPT, and Evans et al (24) compared LDF with ethyl chloride and EPT. Both studies showed promising results for LDF.

The present study was designed to compare the clinical accuracy, reliability, and repeatability of LDF and various traditional pulp sensibility tests.

Materials and Methods

Twenty healthy patients (12 male and 8 female) participated in the study at the Oral Health Centre of Western Australia. The mean age was 30 years, with a range of 18–74 years. The Human Research Ethics Committee at the University of Western Australia approved the study protocol. All subjects were tested 3 times. An interval of at least 1
week was left between each test session. All examinations and tests were carried out by 1 investigator for standardization purposes.

In the first session, all teeth had provisional diagnosis of the pulp status made through clinical and radiographic examinations before pulp testing the teeth. Pulp diseases were confirmed by subsequent endodontic treatment carried out by the relevant practitioners chosen by the test subjects. Each subject had up to 11 teeth selected for pulp testing, with a total of 121 teeth tested. The teeth were selected on the basis of 1 of the following: suspected or known to have pulp pathosis (history of trauma, caries, extensive restorations, pain, or discomfort); previously received or currently undergoing endodontic treatment; or provisionally diagnosed as having a healthy pulp (having no caries and either minimally restored or with no restorations).

Six of the participants had no restorations in the teeth tested, and in 2 of the subjects all of the tested teeth were restored to a certain extent. The teeth morphotypes can be found in Table 1. The state of the pulps and the presence/absence of restorations are summarized in Table 2.

The order of pulp testing in the first and third visits was LDF, CO₂, EF, Ice, EPT. During the second visit, the order of testing was reversed.

### Thermal Pulp Tests

These were applied to the gingival third of the buccal surface of dried and isolated teeth for up to 10 seconds per tooth, with a 2-minute interval between different tests. Alternatively, when any part of the tooth had thin enamel or where there was exposed dentin or metal restorations, these surfaces were used as the test surface if deemed likely to give better thermal conduction. Subjects were instructed to indicate when they felt the cold sensation, which was then recorded. "No response" was recorded when no pulp response was noted after the digital display reached its maximum level of 80 twice consecutively in the 1 session.

For EPT (Electric Pulp Tester; Analytic Technology, Redmond, WA), in accordance with the manufacturer’s instructions, toothpaste was applied to the probe tip to assist electrical conductance. The rate of voltage increase was set between 3 and 4. The probe was placed on the incisal or occlusal third of the tooth where possible. Interproximal isolation and insulation with a plastic strip were done where current conduction to an adjacent tooth through metal was possible. "No response" was recorded when no pulp response was noted after the digital display reached its maximum level of 80 twice consecutively in the 1 session.

In LDF, pulp blood flow (Flux) was measured by a laser Doppler flowmeter (MoorLAB/HoLAB; Moor Instruments Ltd, Aximinster, UK) calibrated against a brownian motion medium and zeroed against a static reflector. The laser source was 780 nm, with 0.5-mm fiber separation by using the MoorLAB P13 and P5a probes, with 3.1 kHz as the primary bandwidth for filter set to 0.1-second time output constant. Customized polyvinylsiloxane (AFFINIS soft/fast putty, AFFINIS regular body; Coltène/Whaledent, Cuyahoga Falls, OH) splints, with holes drilled for metal sleeves to hold the probe tip in place 2–3 mm above the buccal gingival margin, were fabricated for each participant, and a 90-second recording time for each tooth was used where possible. Patients were rested in a supine position for 3–5 minutes before starting the LDF test. Signals measured from each tooth were recorded by using customized PC software (LABSOFT; Moor Instruments Ltd) and analyzed to give the raw data consisting of mean flux value for each recording.

To assess whether LDF produced a response, the tested tooth was paired with its contralateral tooth (except for 1 subject for whom no contralateral tooth was available) for flux comparison. This was based on the observation that a tooth with no pulp blood flow has approximately 40% less LDF output signal than a clinically normal pulp (23, 25). This can be expressed such that no response was recorded for a tooth if the suspected diseased pulp flux/known healthy pulp flux ratio was ≥0.6.

The raw data were processed by the software programs MoorSoft for Windows (Moor Instruments Ltd), Microsoft Excel, Microsoft Word (Microsoft Corporation, Redmond, WA), and SPSS 15.0 (SPSS Inc, Chicago, IL).

For comparison of accuracy, reliability, and repeatability, the average responses of the patients conducted during 3 visits were subjected to interclass correlation (ICC), repeated measures analysis of variance (RM ANOVA), and pairwise comparisons, with *P < .05* as the threshold of statistical significance.

In this analysis, accuracy was defined as a measure of test efficacy explored by RM ANOVA and pairwise comparisons examining between-test accuracy differences. Reliability was defined as a measure of test precision, ie, measurement of test efficacy/accuracy during the 3 visits. This was explored by RM ANOVA and pairwise comparisons examining between-test and within-test accuracy differences across visits. Repeatability was defined as a measure of whether a particular test will yield the same result, irrespective of its accuracy, during the 3-visit time frame. ICC was used to explore the repeatability of the

### Table 1. Morphotypes of the Teeth Tested

<table>
<thead>
<tr>
<th>Teeth morphotype</th>
<th>No. of teeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxillary central incisors</td>
<td>29</td>
</tr>
<tr>
<td>Maxillary lateral incisors</td>
<td>22</td>
</tr>
<tr>
<td>Maxillary canine</td>
<td>12</td>
</tr>
<tr>
<td>Maxillary premolars</td>
<td>24</td>
</tr>
<tr>
<td>Maxillary molars</td>
<td>12</td>
</tr>
<tr>
<td>Mandibular incisors</td>
<td>8</td>
</tr>
<tr>
<td>Mandibular canine and premolars</td>
<td>8</td>
</tr>
<tr>
<td>Mandibular molars</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>121</td>
</tr>
</tbody>
</table>

### Table 2. The Restorative and Pulp Status of the Tested Teeth

<table>
<thead>
<tr>
<th>Restorative status</th>
<th>Teeth with clinically normal pulps</th>
<th>Diseased pulp/RCT/RCF</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No restorations</td>
<td>82</td>
<td>1</td>
<td>83</td>
</tr>
<tr>
<td>Minimally restored</td>
<td>10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Moderately restored</td>
<td>7</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Extensively restored</td>
<td>13</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>Total</td>
<td>112</td>
<td>9</td>
<td>121</td>
</tr>
</tbody>
</table>

RCT, teeth had root canal treatment commenced but not completed; RCF, teeth had existing root canal filling.
tests. ICC has a value between 0 and 1, with 1 being most repeatable and 0 being not repeatable.

Results
The disease prevalence was approximately 7% (9 of 121 teeth). On average during the 3 visits, the EPT (0.98) and CO2 (0.97) tests had the highest accuracy, followed closely by the LDF (0.95) and EF (0.91) tests. The Ice test (0.85) had the lowest overall accuracy (Fig. 1). Summaries of the test accuracies of each visit are provided in Figure 2. RM ANOVA showed a significant difference between the accuracy of the tests ($F_{2,38} = 7.96$, mean squared error [MSE] = 0.23, $P < .001$, $\eta^2_p = 0.30$), i.e., some tests were more accurate than others. Pairwise comparisons indicated that there was no significant difference in accuracy between the CO2 test, LDF test, and EPT ($P > .3$). The CO2 test was significantly more accurate than the Ice test by 12.3% ($P = .005$) and EF by 6.3% ($P = .022$). EPT was significantly more accurate than Ice by 12.9% ($P = .004$) and EF by 7.0% ($P = .015$). The LDF test was significantly more accurate than Ice by 11.5% ($P = .006$); although the LDF test was 5.6% more accurate than the EF test, there was no statistical significance ($P = .065$). The EF test was significantly more accurate than the Ice test by 5.9% ($P = .019$) but less accurate than CO2 test and EPT, as described above. The Ice test was significantly less accurate than all other tests ($P < .019$).

RM ANOVA indicated that not all tests were reliable because of the significant differences found in between-test accuracy across visits ($F_{2,38} = 8.48$, MSE = 0.01, $P = .001$, $\eta^2_p = 0.31$). Pairwise comparisons indicated that test accuracies from visit 1 were significantly different to both visit 2 and visit 3 ($P = .013$ and $P = .004$, respectively), whereas no statistical difference was found between visit 2 and visit 3 ($P = .103$). Within-test accuracy across visits showed that results in all 3 visits significantly differed in accuracy for the Ice test ($F_{2,38} = 5.67$, MSE = 0.01, $P = .007$, $\eta^2_p = 0.23$), with a maximum difference of 10% across visits. Other tests did not have any significant differences in accuracy across visits ($P > .054$); therefore, the Ice test can be considered as being relatively unreliable. Pairwise comparisons indicated that the accuracy for the Ice test from visit 1 was statistically different to visit 2 and visit 3 ($P = .042$ and $P = .009$, respectively), whereas no statistical difference was found between visit 2 and visit 3 ($P = .32$). For repeatability with ICC, the order of the tests from the most to the least repeatable was Ice, LDF, EF, EPT, and CO2 (Table 3).

Discussion
When comparing the current study with other reports, the accuracy of pulp tests needs to be interpreted cautiously because other studies have reported higher disease prevalence. Petersson et al (26) reported 39% and Evans et al (24) reported 44%, whereas the current study had only 7%. The cohort of subjects and teeth in the study by Fuss et al (6) had the same disease prevalence as the current study at 7%, and their CO2 test correctly identified teeth with healthy pulps approximately 97% of the time. However, it should be noted that no diseased teeth were included in their study. In the current study, the CO2 test correctly identified teeth with no pulp disease 97% of the time, and EPT correctly identified teeth with no pulp disease 98% of the time. In the study by Fuss et al (6), EPT identified healthy pulps approximately 90% of the time.

This lower figure can be attributed to the inclusion of immature teeth in which EPT results have been found to be inconsistent (3, 8). In contrast to the current study and other studies (2, 23, 27), all 10 root-treated teeth tested in the study by Fuss et al (6) did not have false responses to EPT. The EF test identified teeth with healthy or no pulps 91% of the time in this study, which was less than DDM in the study by Fuss et al (6), in which healthy pulps were correctly identified 99% of the time. This can be attributed to the higher temperature of the EF test compared with DDM (28). In vitro studies (6, 12) have reported a faster rate of temperature change produced by refrigerant sprays than the rate with CO2, but without superior accuracy, and the current study had similar findings with regards to accuracy. Perhaps the warmer oral environment might play a more significant effect on test outcome than was originally suggested by Miller et al (13).

Interestingly, the greatest difference was found with the Ice test, where it correctly identified teeth with healthy or no pulps 85% of

<table>
<thead>
<tr>
<th>Test type</th>
<th>$\sigma^2_p$</th>
<th>$\sigma^2_w$</th>
<th>ICC</th>
<th>SE (ICC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO2</td>
<td>0.002171</td>
<td>0.00285</td>
<td>0.43233</td>
<td>0.26833</td>
</tr>
<tr>
<td>Ice</td>
<td>0.02533</td>
<td>0.0121</td>
<td>0.67668</td>
<td>0.22983</td>
</tr>
<tr>
<td>EF</td>
<td>0.01429</td>
<td>0.01036</td>
<td>0.57952</td>
<td>0.25908</td>
</tr>
<tr>
<td>EPT</td>
<td>0.002445</td>
<td>0.003189</td>
<td>0.43401</td>
<td>0.26846</td>
</tr>
<tr>
<td>LDF</td>
<td>0.005086</td>
<td>0.002697</td>
<td>0.65352</td>
<td>0.23847</td>
</tr>
</tbody>
</table>

$\sigma^2_p$, between-person variance component; $\sigma^2_w$, within-person variance component; SE, standard error.
Reliability analysis by using between-test accuracy across visits indicated that there was no effect as a result of altering the testing order. Given that the testing order was reversed in the second visit and if somehow had an effect on the test results, then the test results for the second visit should be statistically different to the first and third visits. Fuss et al (6) reported that order reversal of thermal testing and EPT did not affect their test results, and Pantera et al (35) reported that EPT reliability was not affected by the use of DDM during testing.

Although RMANOVA of between-test accuracy across visits (\(F_{2,38} = 8.48, \text{MSE} = 0.01, P = .001, \eta^2_p = 0.31\)) suggested that not all the tests were reliable, it should also be noted that there could have been a learning effect with 1 or all of the test modalities, because the first visit had lower accuracies compared with the other 2 visits. The examination of this possible effect on a test-by-test basis was shown by intratest reliability.

Only the Ice test showed significant within-test accuracy differences over visits (\(F_{2,38} = 5.67, \text{MSE} = 0.01, P = .007, \eta^2_p = 0.23\)), where visit 1 was significantly different to visit 2 (\(P = .042\)) and visit 3 (\(P = .009\)). This indicated that the Ice test was not reliable (that is, not consistent in accuracy over 3 visits). However, a learning effect of the operator or the participants could not be excluded. The other tests appeared to be relatively reliable (\(P > .054\)).

An attempt was made with this analysis to examine whether the tests would yield the same repeatable results irrespective of the accuracies (Table 3). On a relative scale, Ice and LDF appeared most repeatable, whereas EPT and CO2 were less repeatable. The ICC values of all the tests were between 0.4 and 0.7, indicating that no tests in this study were absolutely repeatable or unrepeatable.

Conclusions

The accuracy of EPT, CO2, and LDF was 97.7%, 97.0%, and 96.3%, respectively, with no significant differences in accuracy. The accuracy of EF and Ice was 90.7% and 84.8%, respectively. EPT and CO2 were significantly more accurate than EF. LDF was more accurate than EF, but this was not statistically significant. Ice was significantly less accurate than all the other tests. With the exception of Ice, all tests were reliable. Ice and LDF were the most repeatable of the tests, whereas EPT and CO2 were less repeatable. However, Ice was the least accurate and least reliable test. Within the limitations of this study, no test was superior in all accuracy, reliability, and repeatability scores compared with the other tests when analyzed by using different statistical methods. EF was reliable but not as accurate as EPT and CO2, and it was less repeatable than LDF and Ice. CO2, EPT, and LDF were reliable and the most accurate tests, but CO2 and EPT were less repeatable but less time-consuming than LDF.

Acknowledgments

The authors are very grateful to the Australian Dental Research Foundation for their support in this research through the Colin Corrie Scholarship.

The authors deny any conflicts of interest related to this study.

References


