Consistency of the Babinski reflex and its variants

J. Singerman and L. Lee
Division of Neurology, Sunnybrook Health Sciences Centre, University of Toronto, Toronto, ON, Canada

Keywords: neurological examination, reflex, Babinski, reflex, Chaddock, reflex, Gordon, reflex, Oppenheim, reproducibility of results

Received 17 February 2008
Accepted 27 May 2008

Background and purpose: The Babinski Reflex, first described in 1896, is still an integral part of the neurological examination. Many have studied the consistency of this reflex, but none have compared the inter- and intra-observer consistency of the Babinski reflex and its variants. Methods: Thirty-four subjects were examined by six neurologists. The Babinski, Gordon, Chaddock, and Oppenheim reflexes were tested, and each neurologist concluded if the plantar response was flexor or extensor. Six subjects were re-tested 1 week later to determine intra-observer consistency. Results: The Babinski reflex had the highest interobserver consistency with a kappa value of 0.5491. The Chaddock, Oppenheim, and Gordon reflexes had kappa values of 0.4065, 0.3739, and 0.3515, respectively. For intra-observer consistency, Gordon was the most consistent with a kappa value of 0.6731. When reflexes were combined in pairs, the Babinski and Chaddock reflexes together were the most reliable. Conclusions: The Babinski reflex was shown to be the most consistent between examiners. The Gordon reflex had the highest intra-observer consistency; however, the small sample size should limit conclusions drawn from this calculation. Clinicians often utilize more than one reflex to examine the plantar response; the combination of the Babinski and Chaddock reflexes was the most reliable.

Introduction

The plantar reflex was first described by Joseph Francois Felix Babinski in 1896. The initial stimulus consisted of a pinprick applied to the sole of the foot. Pricking the affected side in patients with hemiplegia because of a pyramidal tract lesion would cause extension of the hallux, even in cases where the patient could not extend the toes voluntarily [1]. In subsequent publications, Babinski noted that this sign was best elicited with a firm stroke on the lateral sole [2], and more refined descriptions of the response included the fanning of the toes as an adjunct to the extension of the great toe [3]. Babinski’s sign was used to help exclude hysteria as a cause of hemiparesis [4]. In 1904, Alfred Gordon, an American neurologist, described a similar response by firmly pressing the middle or lower portion of the calf muscles in patients with known pyramidal tract lesions, producing an extensor response even in situations when the Babinski reflex was absent [5]. Chaddock, who had worked under Babinski in France from 1897–1899, published his own version of the famous sign in 1911, in which the external inframalleolar skin, rather than the sole of the foot, was stroked [6–8]. Just as Babinski was the neurological prodigy of France, so was Hermann Oppenheim in Germany. Oppenheim had described a version of the great toe reflex himself, in which the extension of the great toe could be elicited by firmly stroking the medial tibia [9]. These modes of eliciting the great toe reflex have persisted into the twenty-first century. The Babinski reflex continues to be used routinely by clinicians as part of the examination to determine whether a lesion of the pyramidal tract exists and can be regarded as a tool to help determine lesion localization. However, extension of the great toe has also been reported in post-ictal states [10], with ingestion of certain drugs [11] and during normal sleep [12]. It is also well known that extension of the great toe can be elicited in infants, a normal response that disappears within the first year of life [13]. Although not specific, the Babinski reflex continues to be considered as an integral part of the neurological examination that continues to be taught in our medical schools. However, the Babinski reflex is not without its controversy. One of the major sources of debate focuses on its reliability amongst clinicians and its utility has been questioned [14]. Comparative studies between Babinski and the various techniques of eliciting the extensor response have been few, citing that plantar stimulation is the most effective [15]. In this
study, we compared four techniques for eliciting the plantar reflex in the same cohort of subjects to determine the intra- and inter-observer reliability of these reflexes.

Methods

Ethics approval for this project was obtained from both the Toronto Academic Health Sciences Council and from the Sunnybrook and Women’s College Health Sciences Centre Ethics Board. Neurologists from the Sunnybrook Health Sciences Centre, University of Toronto, Canada, performed the Babinski, Gordon, Chaddock, and Oppenheim reflexes on 34 subjects in this study that was conducted between November 2006 and March 2007. After obtaining informed consent, 23 subjects with varied neurological pathology were non-randomly selected from the inpatient medicine unit, of which stroke was the most common cause for admission. Eleven additional subjects who were admitted at the same time and did not have any neurological pathology were included, serving as controls. Examining neurologists did not know if they were examining subjects with neurological disease or controls, and they were blinded from any clues to underlying pathology (e.g. facial droop or hemiparesis) by pulling the curtain around the bed, masking the subject so that only their feet were visible.

The number of neurologists participating for a testing session ranged anywhere from two to six physicians. Physicians performed the Babinski, Gordon, Chaddock, and Oppenheim reflexes in a predetermined order that differed between trials. The neurologists were to interpret whether the reflex was flexor (downgoing) or extensor (upgoing) with each technique. ‘Equivocal’ responses were not accepted. Examiners were to score an upgoing toe if it was clearly upgoing only. They could observe for flexion synergy of other muscle groups or compare to the other foot to help in their determination. Each subject had all four reflexes tested on each foot, for a total of eight evoked reflexes per subject performed by each neurologist. The neurologists were allowed to repeat each reflex as often as they deemed necessary before concluding whether it was flexor or extensor, to mimic clinical practice.

The neurologists were given written and pictorial instructions of how to perform the reflexes. To accomplish the objective that the results of the study would most accurately reflect how these tests are used in practice, the neurologists were allowed to use their choice of tool to elicit the reflex, such as the end of a reflex hammer or a key to scratch the sole of the foot. A very small subgroup of subjects was re-tested 1 week later by the same group of physicians to examine the intra-observer consistency. Blinding again was preserved, in addition to the curtain, revealing only the subjects’ feet; many of the subjects were not located in the same room that they had been during the previous testing session. Therefore, the neurologists were unlikely to remember which subjects had flexor or extensor reflexes from the preceding week. Furthermore, the neurologists were not told which subjects they were testing for the first time and which they had already tested 1 week earlier.

The inter-observer agreement was assessed by means of the kappa statistic, first described by Landis and Koch [16]. The kappa statistic is measured in terms of strength of agreement (as opposed to a P-value for statistical significance), with a kappa value of 1 indicating near-perfect agreement and a kappa value of 0 indicating very poor agreement.

Results

This study was performed on 34 different subjects of whom six underwent repeat testing 1 week later. The trials were performed on subjects admitted to the Sunnybrook Health Sciences Centre medical clinical teaching unit on 14 separate sessions, and a total of 320 responses were obtained. Twenty-three subjects had a neurological diagnosis, of which stroke was the most common. Non-neurological causes for admission included subjects with congestive heart failure or liver abscesses. The number of neurologists who participated in each trial ranged from two to six.

Our data were analyzed and a value for kappa was given to each reflex. The kappa values for each reflex are shown in Table 1. The Babinski testing revealed the highest inter-observer consistency [0.5491, 95% confidence interval (CI) = 0.452–0.646], indicating moderate to substantial agreement. This was followed by the Chaddock technique (0.4065, 95% CI = 0.313–0.500), the Oppenheim technique (0.3739, 95% CI = 0.279–0.469), and the Gordon technique (0.3515, 95% CI = 0.255–0.448), all of whose kappa statistics ranged from fair to moderate agreement.

<table>
<thead>
<tr>
<th>Table 1 Interobserver and intra-observer consistency of plantar reflexesa</th>
<th>Interobserver consistency</th>
<th>Intra-observer consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babinski</td>
<td>0.5491</td>
<td>0.3489</td>
</tr>
<tr>
<td>Gordon</td>
<td>0.3515</td>
<td>0.6731</td>
</tr>
<tr>
<td>Chaddock</td>
<td>0.4065</td>
<td>0.5182</td>
</tr>
<tr>
<td>Oppenheim</td>
<td>0.3739</td>
<td>0.4868</td>
</tr>
</tbody>
</table>

aValues given as kappa statistics.
In the limited sample of subjects who underwent repeated testing (six subjects), the Gordon reflex had the highest intra-observer agreement (0.6731) followed by the Chaddock technique (0.5182), also shown in Table 1. In clinical practice, neurologists not infrequently, combine different techniques to assist in their interpretation. Table 2 shows the combination of testing techniques. The use of the Babinski and the Chaddock reflex together revealed the highest inter-observer reliability findings.

**Discussion**

Our results showed that the Babinski sign had the greatest agreement between observers, followed by the Chaddock reflex. This is in keeping with clinical practice, in that the Babinski sign appears to be the most agreed-upon sign amongst clinicians. Not surprisingly, the combination of the Babinski and Chaddock reflexes together yielded the most consistent conclusions amongst the neurologists. However, this higher agreement seen with the Babinski sign may only be a reflection of what is most frequently used by clinicians and the higher disagreement with the other techniques may be a result of lack of familiarity and/or comfort in performing these other tests. On the other hand, the Gordon reflex had the highest intra-observer agreement, followed by the Chaddock technique. These higher intra-observer agreements over the Babinski would thus suggest that the results might not purely be a result of familiarity of technique. At any rate, intra-observer interpretations were only based on six subjects; thus any conclusions derived should be limited. Certainly the numbers are too small to place any confidence that one reflex is more reliable within the individual neurologist.

Previous studies have focused on the reproducibility of the Babinski reflex. In a study by Maher et al. [17], complete agreement between examiners occurred only 50% of the time and the average agreement between examiners above chance was only 16.7% (kappa = 0.17). The intra-observer variation over 24 physicians averaged 59.6%. However, this earlier study only included 12 subjects. The physicians were divided into groups where each physician tested only two different subjects, unlike our study where the neurologists were involved in testing up to the 34 subjects enrolled. The poor reproducibility of the plantar response was similarly reported by McCance et al. [18]. However, this study consisted of subjects with psychogeriatric ailments rather than pyramidal tract lesions, and there were only two examining physicians involved in the testing.

A recently published study concluded that the Babinski reflex has poor reliability when used to identify upper motor neuron weakness, whilst other tests such as speed of foot tapping are more dependable (kappa values of 0.30 and 0.73, respectively) [14]. One of the criticisms of the design of that study was that the physicians were not permitted to elicit the Babinski response on both of the subjects’ feet in the same trial, and therefore were unable to compare the responses of the left and right foot, an arguably integral part of determining whether a response is flexor or extensor [19]. Alternatively, other studies have been reported to show much higher inter-observer reliability of the Babinski reflex, with a kappa values as high as 0.98 [20–22].

The results of our study are not unlike those reported by Dohrmann and Nowack [15], who concluded that the most sensitive method of eliciting extension of the great toe is stimulation of the lateral plantar surface and transverse arch in a single movement using a wooden stick (i.e. the Babinski method).

The authors recognize that there are inherent limitations in the conduct of this study. The main limitation in our study is the variable number of subjects and testing neurologists in any one testing session. Both recruitment of subjects and participation of a large number of neurologists were sometimes difficult, resulting in reduced power of the study. This was especially true of the intra-observer reliability, as many subjects were not available for a second trial. In those subjects who were available, there was a possibility of bias if the physicians were able to recall the results from the first trial. However, this was unlikely, as the physicians did not know which subjects were being seen for the second time, and in addition, they were permitted only to observe the subjects’ feet.

Another way bias could have been introduced into the study was by relying on the results of the Babinski reflex, the current ‘gold standard’ of eliciting an extensor response of the great toe, to interpret the remainder of the reflexes. We attempted to reduce this variable by having a pre-determined order of performing the reflexes that was changed with each new trial. However, this did not completely eliminate the possibility of biasing three of the reflexes based...
on the results of the first one performed, regardless of which reflex that was.

The assumption of this study is that the plantar response will remain the same throughout the time of the testing. The responses could conceivably change over the testing session between neurologists, even on the same day and certainly between the days of the repeat trials for the six subjects used for intra-observer consistency. An interesting report described how, over time, the threshold for producing an extensor response lowers, and the size of the receptive field for which such a response will occur enlarges [23].

The level of wakefulness may have changed during the time of testing, and as mentioned earlier, previous studies have cited the abnormal plantar response with sleep [12]. An attempt was made by the study authors to keep the subjects awake throughout the testing, but the level of the subject alertness was not objectively measured or controlled. All of these unavoidable limitations will impede on the interobserver and intra-observer agreement scores.

The lack of standardization in the execution of these reflexes in that the neurologists were permitted to use their choice of tool, may seem to impact the validity of this study. However, we allowed this to imitate how the physicians perform the reflexes in their daily practices, which may actually enhance the generalizability of this experiment.

An additional limitation was the hesitancy of the neurologists in performing some of the reflexes. The Oppenheim reflex, in particular, can be quite painful for patients, and thus the physicians may not have applied the amount of force required to elicit a response. It was also noted by one of the participating neurologists that many of the subjects had thin or broken skin in the test area; consequently, he was reluctant to perform the reflex. Contrary to a letter published in a popular medical journal [24], many physicians do try to minimize pain inflicted with these reflexes. Furthermore, repetitive testing may have become uncomfortable and subjects may have begun to withdraw unintentionally from repeated testing, making the examiner more likely to describe an extensor response.

Finally, the authors recognize that restricting interpretation of plantar responses to be extensor or flexor and not accepting an ‘equivocal’ answer was not a reflection of day-to-day clinical practice, and may have allowed an element of chance (or physician bias of what an extensor response truly is) to influence the results. This was chosen, however, to facilitate statistical calculations when we anticipated the limited number of study subjects and examining neurologists required to derive any conclusions. In addition, whilst the neurologists attempted to differentiate a true extensor plantar response from a withdrawal response or ‘pseudo-Babinski’, there is obviously no absolute confidence that the distinction was always made.

However, despite these limitations, our findings in this study suggest some interesting observations. The number of subjects examined and the number of neurologists involved in this study were larger than those previously reported. Our main conclusion is that the Babinski reflex had the highest interobserver reliability compared with the Gordon, Chaddock, and Oppenheim reflexes. There is a suggestion that the combination of the Babinski and Chaddock techniques would improve the agreement amongst clinicians as to whether the plantar response is extensor or flexor. One could arguably conclude that the interpretation of these reflexes, although with some variability, continues to have some validity and a role in the neurological examination, and a worthwhile endeavor to be taught to medical trainees.

**Acknowledgements**

The authors wish to thank Dr Donald Redelmeier, Dr Alex Kiss, Ms Kathleen Carr, Ms Melanie Winters, and Dr Gavin Ling for their assistance in conducting this study.

**References**


22. Landau WM, Clare MH. The plantar reflex in man, with special reference to some conditions where the extensor response is unexpectedly absent. *Brain* 1956; **79**: 529–556.

