The interrater reliability of a routine outcome measure for infants and pre-schoolers aged under 48 months: Health of the Nation Outcome Scales for Infants

Peter Brann, Gordana Culjak, Nick Kowalenko, Rosemary Dickson, Tim Coombs, Anne Sved Williams, Elisabeth Hoehn, Simon Davies, Margaret Hoyland and Philip Burgess

Background
A review of Australian mental health services identified a gap in routine outcome measures addressing social, emotional and behavioural domains for pre-schoolers and infants. The Child and Adolescent Mental Health Information Development Expert Advisory Panel Working Group developed the Health of the Nation Outcome Scales for Infants (HoNOSI), a clinician-reported routine outcome measure for use with those aged under 4 years. Prior psychometric testing showed that the HoNOSI was considered to show face validity, and that it met the standards for concurrent validity and internal consistency.

Aims
We aimed to investigate the interrater reliability of the HoNOSI.

Method
Forty-five infant mental health clinicians completed HoNOSI ratings on a set of five case vignettes.

Results
Quadratic weighted kappa interrater reliability estimates showed the HoNOSI to have Almost Perfect interrater reliability for the HoNOSI total score. Of the 15 scales, one had Moderate, seven had Substantial and seven had Almost Perfect interrater reliability. Ten of the fifteen scales and the total score exceeded the CONsensus-based Standards for the Selection of Health Measurement INstruments criteria for interrater reliability ($k_w \geq 0.7$).

Conclusions
There has been a clear need for a routine outcome measure for use with infants and pre-schoolers. This study provides evidence of interrater reliability. The current findings, combined with the face and concurrent validity studies, support further examination of HoNOSI in real-world settings.

Keywords
HoNOSI; mental health; outcome measure; infants; interrater reliability.

Copyright and usage
© The Author(s), 2021. Published by Cambridge University Press on behalf of the Royal College of Psychiatrists. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

In 1990, Jenkins identified an urgent need for a system of indicators to enable clinicians to monitor and evaluate mental healthcare.1 One reason identified for not routinely using standard outcome measures was the lack of appropriate instruments.2 In 1998, Wing et al2 developed the Health of the Nation Outcome Scales (HoNOS), an instrument covering symptoms, functioning, relationships and environmental issues,3,4 which could be used routinely in the UK National Health Service to measure progress toward the target set by the UK Department of Health ‘to improve significantly the health and social functioning of mentally ill people’.5 Since then, the HoNOS and its adaptations for children and adolescents (HoNOSCA) and for those aged >65 years (HoNOS65+), have been officially adopted in England, Australia, New Zealand and other European countries.7–10

Gowers et al11 developed the HoNOSCA as a set of scales to be used in Child and Adolescent Mental Health Services.12 The HoNOSCA has been widely used.13–19 It was designed to be brief, have a similar structure to the HoNOS and provide a broad, quantitative measure of severity to measure a range of behavioural, symptomatic, social and impairment domains in children and adolescents and it has been found to have sound psychometric properties.11,20,21

In Australia, the National Outcomes and Casemix Collection (NOCC) was introduced to ‘provide a suite of measures that support clinical practice and comparisons across services and different consumer populations’.19 This saw the implementation of a range of outcome measures to routine clinical practice across age groups. At the core of the NOCC was the family of clinician-rated HoNOS. A review of NOCC22–24 identified a measurement gap for infants and pre-schoolers.

The Australian Child and Adolescent Mental Health Information Development Expert Advisory Panel (CAMHIDEAP)25 provides advice to the Australian government on mental health information development in Australian child and adolescent mental health services. Several members of CAMHIDEAP participated in a collaborative international review of the interrater reliability of HoNOSCA.13 This study also involved participants from the UK, Denmark, Norway and New Zealand. CAMHIDEAP members, including Dr Peter Brann, Dr Tim Coombs and Dr Sally Merry as the New Zealand representative, commenced conversations with the other participants in the HoNOSCA interrater reliability study about the development of a measure for infants and pre-schoolers, i.e. the Health of the Nation Outcome Scales for Infants (HoNOSI).24 Australia and New Zealand, through the CAMHIDEAP, commenced the measure development work, with the initial mapping of domains undertaken by Dr Sally Merry. The CAMHIDEAP had formed a working group to look broadly at routine outcome measures in the very young and supported by the results of the review of the NOCC, took ownership of the HoNOSI initiative, developing and refining a measure suitable for sector consultation and field trialling.

Face validity testing25 showed that the HoNOSI was perceived to fill a much needed gap in infant mental health outcome...
measurement for the 0–47 month age group (infants and preschoolers). Following subsequent concurrent validity testing, the CAMHIDEAP Working Group identified the need to test the interrater reliability of the HoNOSI.

Aims

This HoNOSI field trial was designed to examine interrater reliability, exploring the degree of agreement among different raters when rating the same case vignettes with the HoNOSI.

Method

A pilot study with three vignettes was conducted by a panel of CAMHIDEAP Working Group members, along with individuals with expertise in infant mental health or mental health measurement. The pilot study identified that the vignettes were substantially more severe than the real cases in the concurrent validity study. The vignettes’ severity was adjusted to accord with the real cases and two more vignettes were added. The current study used the resulting five vignettes. The age and gender of the vignettes were 4 months (female), 9 months (male), 3 years (female), 3 years (female) and 4 years (male). The vignettes can be found within the full HoNOSI Field Trial Report.25 Presenting problems included perceived agitation, aggressiveness (9 months), a paediatric in-patient referral with feeding problems, sensory underresponsiveness and maternal depression (4 months).

For the current study, participants with expertise in infant and pre-school mental health were asked to rate these vignettes. Ethics and research governance approval was obtained from the relevant ethics committee for participants from the Centre for Perinatal and Infant Mental Health and Child and Youth Mental Health Service in Queensland (HREC Ref: HREC/16/QRCH/424) and for participants from the Eastern Health Child Youth Mental Health Service in Victoria (HREC Ref: LR19/040). A total of 45 people participated in the study; 26 from Queensland and 19 from Victoria. All participants received a brief online training in the rightmost column. Scale 2 (activity/attention) is only rated online. Those making the ratings were presented with the vignette and then the HoNOSI glossary, from which the rater selected their preferred rating. Results were analysed with Stata statistical software, version 16.1 for Windows.28 Interrater reliability was assessed with quadratic weighted kappa. A copy of the vignettes is available from the corresponding author, upon request.

The HoNOSI24 is a 15-item set of scales, rated from 0 to 9, with 0 indicating no problem, 1 indicating a minor problem requiring no formal action, 2 indicating a mild problem, 3 indicating a problem of moderate severity, 4 indicating a severe to very severe problem and 9 indicating not known or not applicable.

The total score is calculated by summing the first 13 scales, which relate to infant mental health status. Missing data is treated as zero in calculating totals. As with the HoNOS family of measures, clinicians typically classify ratings as ‘clinically significant’25 if a problem area is rated as either a Mild, Moderate or Severe to Very Severe problem (i.e. a rating of 2, 3 or 4). A list of HoNOSI scales can be found in Appendix 1.

The strength of agreement between raters was measured using the COmmonsense-based Standards for the Selection of Health Measurement Instruments (COSMIN).26 The COSMIN initiative26 was developed to provide guidance on the selection of outcome measures for clinical and research applications.31 The COSMIN includes a methodology for assessing the interrater reliability of measures,20,32 a taxonomy and definitions of measurement properties,26 checklists for assessing the methodological quality24 and criteria for good measurement properties.20 For measures constructed with ordinal ratings (i.e. the rating categories indicate the relative ordering of ‘clinical severity’), the gold standard for reporting is the weighted kappa,20 where reliability is sufficient if $\kappa_w \geq 0.7$. The standards outlined by Landis and Koch20 are commonly used in discussions of reliability and will also be reported to both facilitate comparison and support interpretation.

Results

Of the 45 raters, 39 completed all five vignettes with no missing values. Five raters had a missing rating for one scale with one vignette and one rater had a missing rating for one scale with two vignettes. No HoNOSI scale ratings were rated as 9 (not known/ not applicable). In accordance with the rating rules used by the HoNOSCA, all vignettes were able to have total scores calculated.20 Table 1 shows descriptive statistics and the HoNOSI total score distribution for each of the five case vignettes. The percentiles indicate the total score that occurred at that percentage of the distribution. For example, for Levi, a score of 17 was >10% of the total scores. A score of 18 was >25% of total scores. The 50th percentile is the median score. Percentiles aid in describing the distribution and spread of total scores for each vignette.

All of the possible HoNOSI ratings (0, 1, 2, 3, 4) were utilised for each of the 15 scales in at least one vignette. This demonstrates that the full range of possible scores were used in these vignettes and would contribute to the interrater reliability estimate.

Interrater reliability was assessed by quadratic weighted kappa ($\kappa_w$) estimates (Table 2). The COSMIN standard for acceptable interrater reliability ($\kappa_w \geq 0.7$)21 was exceeded by 10 of the 15 scales and by the total score. From Table 2, it can be seen that scales 7 and 15 were within 0.01 of the standard. Scales 11 and 14 were within 0.08 of the standard. The confidence intervals for those four scales are wide and overlap the standard. Scale 2 had a very wide confidence interval and was clearly lower than the standard.

The Landis and Koch Strength of Agreement criteria35 is shown in the rightmost column. Scale 2 (activity/attention) is only Moderate in agreement (0.41–0.60). However, seven of the scales and the total score have Almost Perfect agreement (0.81–1.00), whereas the remaining seven have Substantial agreement (0.61–0.80).

Discussion

This study was designed to test the interrater reliability of the HoNOSI. Results show that the level of interrater reliability for the total score was Almost Perfect when measured against the benchmarks outlined by Landis and Koch35 and that it also well exceeds the COSMIN standard for interrater reliability.32 The majority of individual scales clearly met the COSMIN criteria and applying the commonly used Landis and Koch35 descriptors, seven scales and the total score had Almost Perfect agreement and seven scales had Substantial agreement between raters.

COSMIN guidelines describe what constitutes a sufficient level of agreement. Using the COSMIN criteria for good measurement properties, the results of the present study well exceed the criteria for interrater reliability, where reliability is rated as either sufficient ($\kappa_w \geq 0.7$), insufficient ($\kappa_w < 0.7$) or indeterminate (where $\kappa_w$ is not reported). Scales 7 (problems associated with regulation and integration of sensory processing), 11 (problems with age-appropriate self-care and environmental exploration), 14 (problems with knowledge...
or understanding about the nature of the infant’s difficulties) and 15
(problems with lack of information, understanding about services or
managing the infant’s difficulties) were classed as insufficient in
terms of reliability (Table 2), although they are close to the arbitrary
cut-off ($\kappa_w \geq 0.7$).

It is only scale 2 (Problems with activity levels, joint and/or sus-
tained attention), with $\kappa_w < 0.5$, that was found to be insufficient
according to the COSMIN criteria, although assessed as a moderate
level of agreement according to the guidelines provided by Landis
and Koch.35 Aside from scale 2, the other 14 HoNOSI scales and
the total score showed a reasonable amount of variability between
vignettes. The study had sufficient power to test the null hypothesis
at the interrater reliability estimate of 0.5.

The HoNOSI has links with the HoNOS family of measures,
especially the HoNOSCA.11 The HoNOSI estimates are stronger
than those obtained with those measures.13,15 Although the
HoNOSCA covers different developmental stages in its use across
ages 4–18 years, the HoNOSI, with a much a narrower age range,
covers developmental shifts of arguably greater magnitude.
Although the interrater reliability was Substantial to Almost
Perfect for the majority of the scales and the total score, it would
be wise to establish whether this overall acceptable estimate hides
less acceptable interrater reliability for different age groups. The
lower estimate for attention may reflect a differential functioning
of the scale for different age groups. It is possible that the scale
can be applied very reliably for 9-month-olds and 3-year-olds, but
less so for 18-month-olds.

Vignettes have been found to be a valid tool when measuring the
quality of clinical practice.36 They remain a key method when
wanting to ensure that raters are exposed to exactly the same
stimuli. However, vignettes are not the same as clinical practice,
and the HoNOSI may perform better or worse with real cases.
Although the vignettes covered ages from 4 months to 4 years,
there were only five presentations assessed. The middle age group
was not covered. Evaluating the HoNOSI in routine clinical practice
would supplement the vignette approach by ensuring that a greater
range of presentations could be examined. Furthermore, the question
of whether the HoNOSI has differing reliability between, for example,
infants and pre-schoolers, could be examined with larger sampling.

It is anticipated that future research could explore HoNOSI psy-
chometric properties with respect to other domains and consumer
attributes, including the specific nature of presenting problems, diag-
nostic categories and developmental differences. With regard to the
rapid maturation that occurs in this age range, conceptualising and
assessing test–retest reliability will be an important task. A face validity
study35 and concurrent validity field trial36 have now been com-
pleted. There are other psychometric properties (e.g. sensitivity to
change) yet to be investigated. However, the interrater reliability find-
ings are sufficiently encouraging to support the further exploration of
the HoNOSI. The HoNOSI is promising, and the very young have
been excluded from the routine outcome measurement framework
for too long. Whether it proves to be useful in this endeavour requires
further exploration of reliability, validity and feasibility and this may
be best achieved by examination in routine clinical practice.

Table 1 Distribution of Health of the Nation Outcome Scales for Infants total score for the five vignettes

<table>
<thead>
<tr>
<th>Vignette</th>
<th>Raters, n</th>
<th>Mean</th>
<th>s.d.</th>
<th>Minimum</th>
<th>Maximum</th>
<th>10th</th>
<th>25th</th>
<th>50th</th>
<th>75th</th>
<th>90th</th>
<th>Percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – Levi (9 months)</td>
<td>45</td>
<td>20.7</td>
<td>3.9</td>
<td>14</td>
<td>31</td>
<td>17</td>
<td>18</td>
<td>20</td>
<td>22</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>2 – Lily (4 months)</td>
<td>45</td>
<td>17.8</td>
<td>4.4</td>
<td>12</td>
<td>31</td>
<td>13</td>
<td>15</td>
<td>17</td>
<td>20</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>3 – Chloe (3 years)</td>
<td>45</td>
<td>10.3</td>
<td>3.6</td>
<td>2</td>
<td>22</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>4 – Sandy (4 years)</td>
<td>45</td>
<td>31.1</td>
<td>4.0</td>
<td>26</td>
<td>42</td>
<td>26</td>
<td>28</td>
<td>31</td>
<td>33</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>5 – Sally (3 years)</td>
<td>45</td>
<td>37.6</td>
<td>4.3</td>
<td>27</td>
<td>46</td>
<td>31</td>
<td>36</td>
<td>37</td>
<td>40</td>
<td>43</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Interrater reliability quadratic weighted kappa estimates for vignettes

<table>
<thead>
<tr>
<th>HoNOSI scale</th>
<th>Raters, n</th>
<th>$\kappa_w$</th>
<th>95% CI lower</th>
<th>95% CI upper</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Problems with disruptive behaviour/irritability/undercontrolled emotional regulation</td>
<td>45</td>
<td>0.86</td>
<td>0.77</td>
<td>0.94</td>
</tr>
<tr>
<td>2</td>
<td>Problems with activity levels, joint and/or sustained attention</td>
<td>45</td>
<td>0.41</td>
<td>0.00</td>
<td>0.81</td>
</tr>
<tr>
<td>3</td>
<td>Non-accidental self-injury or lack of self-protective behaviours</td>
<td>44</td>
<td>0.85</td>
<td>0.79</td>
<td>0.92</td>
</tr>
<tr>
<td>4</td>
<td>Problems with feeding and eating behaviour</td>
<td>45</td>
<td>0.85</td>
<td>0.61</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>Problems with developmental delays</td>
<td>45</td>
<td>0.85</td>
<td>0.76</td>
<td>0.95</td>
</tr>
<tr>
<td>6</td>
<td>Problems with physical illness or disability</td>
<td>45</td>
<td>0.80</td>
<td>0.57</td>
<td>1.00</td>
</tr>
<tr>
<td>7</td>
<td>Problems associated with regulation and integration of sensory processing</td>
<td>43</td>
<td>0.69</td>
<td>0.35</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>Problems associated with sleep</td>
<td>45</td>
<td>0.85</td>
<td>0.57</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>Problems with emotional and related symptoms or overcontrolled emotional regulation</td>
<td>44</td>
<td>0.80</td>
<td>0.55</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>Problems with social reciprocity</td>
<td>45</td>
<td>0.87</td>
<td>0.76</td>
<td>0.98</td>
</tr>
<tr>
<td>11</td>
<td>Problems with age-appropriate self-care and environmental exploration</td>
<td>45</td>
<td>0.62</td>
<td>0.27</td>
<td>0.98</td>
</tr>
<tr>
<td>12</td>
<td>Problems with family life and relationships</td>
<td>44</td>
<td>0.81</td>
<td>0.64</td>
<td>0.98</td>
</tr>
<tr>
<td>13</td>
<td>Problems with attending care, education and socialisation settings</td>
<td>44</td>
<td>0.71</td>
<td>0.27</td>
<td>1.00</td>
</tr>
<tr>
<td>14</td>
<td>Problems with knowledge or understanding about the nature of the infant’s difficulties</td>
<td>45</td>
<td>0.66</td>
<td>0.34</td>
<td>0.97</td>
</tr>
<tr>
<td>15</td>
<td>Problems with lack of information, understanding about services or managing the infant’s difficulties</td>
<td>44</td>
<td>0.69</td>
<td>0.49</td>
<td>0.88</td>
</tr>
<tr>
<td>HoNOSI total score</td>
<td>45</td>
<td>0.85</td>
<td>0.70</td>
<td>1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>
A final caveat: if the HoNOSI continues to demonstrate good enough psychometric properties in future research, it will still share one common feature with other routine outcome measures. Measured change cannot specify the cause of the change. The HoNOSI, HoNOSCA and similar measures identify change but do not independently specify whether this is clearly maturational, treatment-initiated or a result of alteration in the child’s environment. Understanding the source of any change, positive or negative, remains the clinical task at an individual level and a research and policy task at the aggregate level. Hopefully, the HoNOSI will facilitate routine dialogue about mental health outcomes for our relatively-looked over infants and pre-schoolers.

Data availability
The data that support the findings of this study are available from the corresponding author, G.C., upon reasonable request.

Author contributions
All authors contributed to the conception and design of the study. P. Burgess conducted the analysis. G.C. prepared the final manuscript. All authors reviewed and approved the final version of the manuscript.

Funding
The study was undertaken by the Child and Adolescent Mental Health Information Development Expert Advisory Panel (CAMHIDEAP), supported by the Australian Mental Health Outcomes and Classification Network (AMHOCN). CAMHIDEAP and AMHOCN have been funded by the Australian Department of Health through a contract with the Health Education and Training Institute, Australia and the University of Sydney, Australia.

Declaration of interest
None.

Appendix 1: Health of the Nation Outcome Scales for Infants (HoNOSI) – list of scales

| Scale 1 | Problems with disruptive behaviour/irritability/undercontrolled emotional regulation |
| Scale 2 | Problems with activity levels, joint and/or sustained attention |
| Scale 3 | Non-accidental self-injury or lack of self-protective behaviours |
| Scale 4 | Problems with feeding and eating behaviour |
| Scale 5 | Problems with development delays |
| Scale 6 | Problems with physical illness or disability |
| Scale 7 | Problems associated with regulation and integration of sensory processing |
| Scale 8 | Problems associated with sleep |
| Scale 9 | Problems with emotional and related symptoms or overcontrolled emotional regulation |

References


24 Commonwealth of Australia for and on behalf of the Australian Mental Health Outcomes and Classification Network. HoNOSI Health of the Nation Outcome Scales for Infants (0-47 months). Glossary. Ver 0.93. Developed under the guidance of the Child and Adolescent Mental Health Information Development Expert Advisory Panel, 2016.


