Author’s response to reviews

Title: TRACE (Trunk Aesthetic Clinical Evaluation): a routine clinical tool to evaluate aesthetics in scoliosis patients. Development from the Aesthetic Index (AI) and repeatability

Authors:

Fabio Zaina (fabio.zaina@isico.it)
Stefano Negrini (stefano.negrini@isico.it)
Salvatore Atanasio (salvatore.atanasio@isico.it)

Version: 2 Date: 3 January 2009

Author’s response to reviews: see over
Referee 1

Thank you for your kind suggestions.

Frequent errors in grammar and syntax.
The manuscript has been sent to an US English professional translator for correction before submission. Nevertheless, for further corrections, we accepted the kind offer of Dr Hawes to check again the paper.

While a tool to assess aesthetic results of scoliosis treatment would certainly be useful, this study has demonstrated the failure of the TRACE tool to do so with acceptable reliability. Therefore I don't think the results support the conclusion that this tool should be put into use.

Sorry, but we don’t agree with the opinion of the referee. Like all measurement tools also TRACE has some limitations; nevertheless, knowing these limitations, provided that TRACE allows a better understanding of aesthetics than nothing (that is the only alternative in this case for everyday use), we can use it for clinical and research purpose. Moreover, this scale can be a starting point for the development of other instruments, like AI was for TRACE.

INTRO: authors should cite references or examples of existing tools for assessing aesthetic results. References for existing tools were


Authors should state how many years they mean by "some years." METHODS: 
Specify what "many years" is.
Ok. We specified the number of years: “..Our institute has for five years evaluated the…”
BACKGROUND: I would strongly disagree that aesthetics is the main goal of surgical treatment. We agree with you, but there are many studies about the interest of orthopaedic surgeons for aesthetics, and we reported what published by a group of them.


Which "high cost" tools do the authors mean?
High cost is related to the instruments defined in the previous sentence, where references are given, and include rasterstereography, the Quantec system and other similar tools.

RESULTS:
Figures should be clarified.
We gave more details in the captions as follows:
“Figure 1. Shoulder asymmetry, as evaluated in TRACE, ranges from 0-3. For the shoulders it is easy to detect some intermediate values, so we defined asymmetry (from the top) slight (1), moderate (2) and important (3).

Figure 2. Hemi-thorax asymmetry as evaluated in TRACE: This item was created as a complement of the scapulae, since we noted that occasionally there is an evident prominence of the last ribs posteriorly even when there is no real asymmetry in the scapulae. From the left: slight (1) and important (2) asymmetry.

Figure 3. Scapulae asymmetry as evaluated in TRACE: (from the left) slight (1) and important (2).

Figure 4. Waist asymmetry as evaluated in TRACE: it was quite easy to define a total asymmetry (a score of 4) when one flank was straight or when there was a lateral decompensation of the trunk. It was easy as well to define a very slight (a score of 1) and an important but not complete (a score of 3) asymmetry; between these points we defined a mild asymmetry (a score of 2). In the figure, (from the top) slight (1), mild (2), moderate (3) and important (4) asymmetry.”
Referee 2
Thank you for your kind suggestions

In Discussion section the authors could address the question whether the evaluation of AI and TRACE done after photographs gives the same results as the evaluation done on live exam. The photographs are static while the exam reveals more three-dimensional images which are consistently being recorded by the physician.

The issues proposed by the referee have been added:
“Nevertheless, evaluation through pictures has been the standard applied in previous studies, and we can presume the repeatability to be even greater during everyday clinical practice due to the opportunity for a three-dimensional evaluation of the patient. In fact, pictures are static while a three dimensional clinical assessment presumably can be more consistently recorded by the physician.”

Another issue which could be discussed is the eventual correlation of TRACE values with POTS index values, as both indexes can be measured on photographs.

We agree with the referee that pictures evaluation could allow to compare the TRACE with the POTS index, and that this could be an interesting topic, but we think this to go beyond the aim of the present study, since we wanted to define the repeatability of TRACE.
Referee 3
Thank you for your kind suggestions

The reference has been inserted as suggested (see comment to referee 4)

3. Are the data sound and well controlled?
Sensibility and specificity of raters are not evident
To evaluate sensitivity and specificity of raters a gold standard reference should be needed, and this case we lack such a gold standard. In some studies one rater is considered as the gold standard and the others are compared to him: but in this case we preferred to gather a more clinically useful tool from the 95% confidence interval, so to know which values should be used in everyday clinics.

6. Do the title and abstract accurately convey what has been found?
No I think that the authors must add “versus AI (Aesthetic Index)”
We changed the title as suggested
Thank you for your kind suggestions.
Referee 4

the 'AIS' abbreviation needs to be introduced after the first use of adolescent idiopathic scoliosis, and then used consistently throughout. We introduced AIS at the beginning both of title, abstract and article and then used the abbreviation.

split infinitives and other grammatical issues in the abstract are highlighted for correction (below). I can do this for the whole manuscript, if needed.

Thank you for your kind suggestions. We would really appreciate if you would check the entire manuscript as you proposed, principally to improve it editorially.

A reference needs to be provided for Kappa statistics. Given the controversies associated with this method, a parenthetical explanation together with additional references might be worthwhile (i.e. 'We used Kappa statistics, a method that allows xxx with xxx samples, based on a sample of n..' and acknowledge the published controversies (e.g. a-c, below)


We added this paragraph:

We used the Kappa Statistics, that is and index of the observer disagreement, which compares the agreement against that which might be expected by chance. Kappa can be thought of as the chance-corrected proportional agreement, and possible values range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to -1 (complete disagreement). We are aware of the published controversies, nevertheless Kappa statistics is still widely used. In particular, we are aware that Kappa may be low even though there are high levels of agreement and even though individual ratings are accurate. Whether a given kappa value implies a good or a bad rating system or diagnostic method depends on what model one assumes about the decision making of raters. Accordingly we also propose here a clinically oriented measurement, that is the 95% level of agreement.

There are a few places in the abstract that I have provided suggested changes in wording, for consideration by the authors.

p. 1, Abstract.

para. 1. I would recommend for sentence 3: "The aim of the present study is to develop a clinical tool for the assessment of aesthetics in adolescent idiopathic scoliosis (AIS) patients, and to verify its reliability and repeatability."

para 2. sentence 2: "We built on this experience to develop a Trunk Aesthetic Clinical Evaluation (TRACE), that is based on four sub-scales—shoulders (0-3), scapulae (0-2), hemi-thorax (0-2) and flanks (0-4). Each point is fully described, to provide an ordinal scale for increase of asymmetry."

next to last sentence--'12-point scale' not '12-points scale'.

Instead of 'A hundred and sixty PA photographs....', I would recommend 'One hundred-sixty posterior-anterior (PA) photographs of (How many?) AIS patients were scored in two independent tests, by each of four observers, using AI and TRACE. Kappa statistical analysis was used to identify the minimum clinically significant change (95% confidence level)."

Para 4 (RESULTS): 'We found the intra- and inter-rater reliability of AI to be fair.
Para. 5 (conclusions): I recommend 'TRACE is a low-cost tool that provides a direct measure of aesthetic appearance that can be used to educate patients and to monitor changes in clinical signs over time. Its routine clinical use by physicians is advised.'

The abstract was modified as suggested.

*Background:* Aesthetic appearance is of primary importance in the treatment of scoliosis, but to date tools for everyday clinical practice have not become available. The existing tools suffer from high costs that limit their use to surgical and/or research settings. The aim of the present study is to develop a tool for the assessment of aesthetics in adolescent idiopathic scoliosis (AIS) patients and to verify its reliability and repeatability.

**Methods:**

Instrumentation: At first we developed the Aesthetic Index (AI), based on a three-point scale for asymmetry of the shoulders, scapulae and waist that we tested for 5 years. We built on this experience to develop a Trunk Aesthetic Clinical Evaluation (TRACE), that is based on four sub-scales, shoulders (0-3), scapulae (0-2), hemi-thorax (0-2) and waist (0-4). Each point is fully described and gives an ordinal scale for increase of asymmetry. Accordingly, TRACE is indicated by the sum of the sub-score plus one to give a 12-point scale.

Procedures: A hundred-sixty posterior-anterior (PA) photographs of a hundred-sixty AIS patients were scored in two independent tests by four observers using AI, and TRACE.

Data analysis: Kappa statistical analysis was used to identify the minimum clinically significant change (95% confidence level).

**Results:** We found the intra- and inter-raters repeatability of AI to be fair. Three points out of six is the minimum change to be considered significant between two different evaluations both for the same and different raters. For TRACE, however, we found intra-rater repeatability to be fair and inter-raters to be poor. Three points out of twelve is the minimum change to be considered significant between two different evaluations for the same rater, while four out of twelve is the minimum for different raters.

**Conclusions:** Even if it has only a fair repeatability, TRACE is a sensible, low-cost tool for everyday clinical practice in scoliosis patients able to detect to patient changes due to scoliosis progression or treatment. It’s everyday clinical use by physicians is advised..”
Referee 5
Thank you for your useful suggestions.

1. Title: I would suggest using the word ‘routine’ clinical tool rather than ‘everyday’ clinical tool. Also it is not clear from the title that the AI is being compared to the TRACE (MER)
The title has been changed as suggested (see comment to referee 3)

2. I am not sure that I would agree completely with the statement “tools for everyday clinical practice have not become available” though I do agree that they are very limited (such as using the Bunnell Scoliometer and the plumb-line). I would also suggest that the whole abstract and paper should be written in the past tense. I am not quite sure what the scoliosis journal recommends but we tell all students that their work should be written in the 3rd person past tense. On the 5th line of the abstract could the authors clarify what precisely is the difference between reliability and repeatability? My understanding of these is that they are fairly similar concepts. [MER] ATR can be considered as a parameter partially linked to aesthetics but not completely. In fact, for the same Bunnell degrees measured 2 patients can have completely different aesthetic appearance and even some patients can have a worst appearance with less degrees with respect to another. The prominence can be seen during the forward bending and usually girl don’t perform this movement frequently during their everyday life. That’s way we consider not so relevant for an aesthetic evaluation ATR. About the decompensation of the trunk evaluated with a plumbline, in our opinion this parameter is not relevant at all in an aesthetic evaluation but it’s relevant a visual evaluation of it. Nevertheless, these are really useful tools that we use at every clinical evaluation, to be more accurate during the treatment. About the use of the 3rd person and past tense, we will follow the general guidelines of the journal, in case nothing is stated about, we don’t think it’s relevant to change it. We changed the methods section of the abstract according to the suggestions. We didn’t create a “population” paragraph since we didn’t evaluate patients but photographs.

Instrumentation: At first we developed the Aesthetic Index (AI), based on a three-point scale for asymmetry of the shoulders, scapulae and waist that we tested for 5 some years. We built on this experience to develop a Trunk Aesthetic Clinical Evaluation Based on this first experience we evolved it developing (TRACE) (Trunk Aesthetic Clinical Evaluation), that is based on four subscales, --shoulders (0-3), scapulae (0-2), hemi-thorax (0-2) and waist (0-4). Each point is fully described and gives an ordinal scale for the increase of asymmetry. Accordingly, TRACE is indicated by the sum of the sub-score plus one to give a 12-points scale.

Procedures: A hundred-sixty posterior-anterior (PA) photographs of a hundred-sixty AIS patients were scored in two independent tests by four observers using AI, and TRACE. A hundred and sixty PA photographs of the trunks of adolescent idiopathic scoliosis patients were scored two times independently by four observers using AI, and TRACE.

Data analysis: We used the Kappa statistical analysis was used to identify the minimum clinically significant change (95% confidence level).s; moreover, we looked at the 95% level of agreement in order to identify the minimum clinically significant change to be considered. We cut the word realiability and used only repeatability.

3. I would also strongly recommend that the authors divide the methods section in the abstract into further sub-sections these being: subjects, instrumentation, procedures, data-analysis and then proceed to discuss them in that order. [MER] ok, we did.

4. It would be helpful when reporting the results section that the authors also include the actual Kappa value together with the level of significance (P value). [MCR]
We changed according to the referees suggestions: Regarding the AI, we found the repeatability of both intra- and inter-raters to be fair (Kappa value 0.28-0.41; 0.17-0.28). For TRACE, we found intra-rater repeatability to be fair, while inter-raters to be poor (Kappa value: 0.16-0.24; 0.09-0.14;

5. The fact that the authors have included 2 tools (the AI and the Trace) make this article a little difficult to read. I would suggest that they just describe the TRACE tool and provide the results for that and discuss the AI somewhere else (possibly another paper).[MER]

We think it’s relevant to tell the reader how we developed the TRACE from the AI. That’s why we put the data together. Moreover, a relevant part of this article consists in the comparison of these two scales and in the conversion values we calculated. If we cut all the part about the AI we would lose some interesting data. Another aspect, is the consideration that after publishing data about TRACE it won’t be useful for the readers to publish data about another scale that is less sensitive.

6. pg 2: regarding the conclusions I do agree that it is a sensible low cost tool and better than using nothing at all. However the fact that the repeatability is only ‘fair’ makes me question whether or not we can recommend routine clinical use.? Wouldn’t it be possible to improve the method for instance by improving the scales (i.e. stating the side of the problem and actually having a form and figure on paper on which to document this-Like the one found in Posture Score sheet [Reedco 1974]). I have a copy of this and would be very happy to send it to the authors.[MER]

“Our point of view is that, at this stage of research, we must state the importance of more then only Xrays as outcomes and use the instruments for measurement that we have today, even if what we have is not the best. We are just trying to get out from 50 years of clinical research based only on Cobb angles, considered the gold standard and forgetting all its limitations. If we continue to fight for the best measurement we loose the good measurements, that nevertheless do tell something.

The distances from the plumpline tell something about kyphosis, even if they are based only on a three point measurement: nevertheless this is not different from the Cobb angle, that is even worse because it is based on two points (end-vertebrae only), and x-rays require to move the arms so changing the sagittal curve. As stated in the paper we know the measurement error of the distances from the plumpline, so that we can tell clinically if a patient did change or not; moreover, if we reach a statistical significance in a population study, we can tell that there was a difference on this measurement in the two populations. In any case, with its advantages and disadvantages, this tell something on kyphosis: a general variation, not particulars on the eventually existing intermediate rigid regions that are not on the point measured (rare but still possible), but it is a change!

This situation is true for any kind of measurement system. Just to go in another field: VAS for pain is not a very satisfactory measurement, but tells something, it’s easy to use and it is widely used. To measure pain e.g. the McGill Pain questionnaire could be considered better in some respects, but its difficult to use and it’s rarely used. We hope in the future we will have something better, but now we all use VAS for pain. It could be the same for kyphosis: three (distances from the plumpline, Arcometer) and two points (Cobb degrees, inclinometer) measurements are easy to use and are widely used, while other tools could in the future show to be more accurate, but as always happen they will have some other limitations for sure.... In any case, we do not have any other measurement now, so what we have to do ? Not to measure ?”. This was our answer to Dr Rigo about plumpline measurement of sagittal profile included in a paper recently published on “Scoliosis”(Negrini S, Atanasio S, Negrini F, Zaina F, Marchini G: The Sforzesco brace

We think this answer pertinent to the doubts of the referee. We hope to have soon a new tool, and we will work on it in the next future.

I am still unclear how the assessment of the patient is recorded in the authors actual clinical practice? Do they record this on the patient records on a form? IF SO IT SHOULD BE INCLUDED [MER]
Further if one sub-section say the scapulae are seen to improve and not the other sub-scales how is this recorded?[MER]

We use a software to record our patients data. When assessing TRACE, we sign a number close to each item for the value we assigned. We have no pictures nor drawings in our software. When 1 item changes from one assessment to another, we have the new value on the left of the monitor and all the previous ones on the right, in order to make immediately a comparison.

Background
I thought that the background was well discussed but a little brief. There are also some other scales that have been used to assess back shape and posture in the Scoliosis AND physiotherapy literature like for instance the REEDCO one. It would be helpful I think if the authors could discuss the pros and cons of the different methods they mention in some further detail. [MER]
In the background it would be worth including an operational definition for ‘repeatability’ and ‘reliability’ and describing what exactly is the difference between the two?[mer]
Ok
We changed the text to make readers understand better what we have done, avoiding the use of the term reliability and using only repeatability.

Methods
Again I think it would be helpful in the METHODS section to have the same subsections as recommended in the abstract ie start with the subjects, instrumentation, procedures and data analysis, in that order.[MCR] I would also just include only the TRACE and not the AI as otherwise the reader gets mixed up between the two (well I did )[MER]
We divided the sections as suggested.
About population, we must highlight that we didn’t evaluate directly patients but only photographs of the back of AIS patients. We think it’s useless to know about weight, height and BMI. If the referee think it relevant to have these information we can provide them. Which relevant information can we drive from this?

Was consent to be included on this study obtained from the patients?[MER]
This is not an interventional study, so no consent is needed. Nevertheless, all our patients sign a consent for blind data treatment both for administrative and research purpose.

Instrumentation
The TRACE has been appropriately described (except for the comments referred to above) but it would be helpful to place it under this sub-heading. I am not sure what Adding +1 does to the validity of the scale? [MER]
We made the required changes
In any case is a 12 point scale: you can start from 0 to 11, or from 1 to 12; it was made only to make it more “sound”. Also AI is a 7 point scale, going from 0 to 6 (even if in this case we did not feel the need to add one point)

Procedure
Again this has been well described though again better just stick to the TRACE tool only. [MER]
I could not locate the section within the text where the photos of the patients were discussed. [me] It needs signposting. I am assuming the best place for them would be in the procedures section where the authors describe how they calculated the TRACE values or possibly the results section. [MER]
It would also be helpful to provide some discussion of the figures and what values each one was given by each of the observers. [MER]
A signpost for figures was added in the procedures as suggested.

Data Analysis
The methods of data-analysis need to go here. Although mentioned in the abstract it is not included here. Can the authors give some more detail on how the levels of agreement were reached? [MCR]
Data analysis:
Kappa statistical analysis and 95% level of agreement were used; we also identified the minimum significant change (95% confidence level).
The Percent of agreement is the percentage of the answers that were equal in the two repeated measures. The 95% level of agreement means how many points of difference are needed to reach an agreement of 95%. To give an example: for the rater who obtained the worst results the percent of agreement for TRACE was 28.8% (Table 2), but if you consider in the repeated measurements also one point above or below it’s possible to reach a 99.4% of agreement, that corresponds to a 95% of agreement of 2 points out of 12; this result has an high clinical significance, because it means that, in everyday practice, considering two evaluations made by the same rater, a real change occurs only if the variation is over the 95% level of agreement, that in the example given corresponds to 3 points out of 12.

Tables
Please include P values (levels of significance) [MES]
Unclear what is meant by conversion value in tables 3 and 4, however if you are only including TRACE they would not be relevant here.
We included P values.