J Orofac Orthop
https://doi.org/10.1007/s00056-022-00376-9

ORIGINAL ARTICLE

Comparison between stone and digital cast measurements in mixed dentition

Validity, reliability, reproducibility, and objectivity

Lisa Schieffer[1] · Lukas Latzko[2] · Hanno Ulmer[3] · Natalie Schenz-Spisic[1] · Ulrike Lepperdinger[4] · Magdalena Paulus[1] · Adriano G. Crismani[1]

Received: 1 May 2021 / Accepted: 29 December 2021

Abstract

Purpose To assess the validity, reliability, reproducibility, and objectivity of measurements on stone casts of patients with mixed dentitions compared to measurements on three-dimensional (3D) digital models derived from surface scans of the stone casts.

Methods Pairs of stone casts of 30 young patients in their mixed dentition stage were included and processed into 3D digital models using an intraoral scanner (iTero Element 2; Align Technology, San Jose, CA, USA). Then an experienced and an inexperienced examiner independently performed measurements of five defined parameters, each in triplicate, both on the digital models with analysis software (OnyxCeph3™; Image Instruments, Chemnitz, Germany) and on the original casts with a vernier calliper. Paired *t*-tests were used for validity and interexaminer objectivity, Pearson correlation coefficients for intermethod reliability, and intraclass correlation coefficients (ICCs) for reproducibility testing.

Results Significant ($p<0.05$) intermethod differences were identified for four parameters, but only the differences for overbite and intermolar distance exceeded the threshold of clinical relevance (≥ 0.5 mm). Intermethod reliability was high and method error invariably lower for the digital measurements and for the experienced examiner. Both examiners achieved ICCs > 0.907 with both methods. Interexaminer variation involved significant differences for all parameters but one (intermolar distance) on the stone casts and for three parameters on the digital models.

Conclusion Measurements performed on digital models of mixed dentitions can yield clinically acceptable outcomes with OnyxCeph3™ software. Both the digital and the analogue measurements were highly reproducible and reliable. Objectivity of the measurements could not be confirmed, as operator experience did make a difference.

Keywords Orthodontics · Dental casting technique · Virtual arch models · Three-dimensional imaging · Intraoral scanner

Availability of data and material The datasets analysed for this study are not shared with the public due to data privacy considerations. All relevant data are included in the article.

✉ Lisa Schieffer
lisa.schieffer@i-med.ac.at

[1] University Hospital for Orthodontics, Department of Dental and Oral Medicine and Cranio-Maxillofacial and Oral Surgery, Medical University of Innsbruck, Anichstr. 35, 6020 Innsbruck, Austria

[2] University Hospital for Craniomaxillofacial and Oral Surgery, Department of Dental and Oral Medicine and Cranio-Maxillofacial and Oral Surgery, Medical University of Innsbruck, Anichstr. 35, 6020 Innsbruck, Austria

[3] Department of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Schöpfstr. 41/1, 6020 Innsbruck, Austria

[4] University Hospital for Dental Prosthetics and Restorative Dentistry, Department of Dental and Oral Medicine and Cranio-Maxillofacial and Oral Surgery, Medical University of Innsbruck, Anichstr. 35, 6020 Innsbruck, Austria

Published online: 03 March 2022

Der Vergleich manueller versus digitaler Modellvermessung im Wechselgebiss

Validität, Reliabilität, Wiederholbarkeit und Objektivität

Zusammenfassung

Ziel Ziel der Studie war es, einen Methodenvergleich zwischen manueller und digitaler Vermessung an Modellen des Wechselgebisses vorzunehmen. Bewertet werden sollten die Gütekriterien der Validität, Reliabilität, Wiederholbarkeit und Objektivität.

Material und Methode In die Studie eingeschlossen wurden 30 Gipsmodelle von Patient:innen mit Wechselgebiss. Die Gipsmodelle wurden mit Hilfe eines Intraoralscanners (iTero Element 2; Align Technologiy, San Jose, CA, USA) digitalisiert. Zwei unabhängige Untersuchende führten die Modellvermessungen (5 Parameter) jeweils 3-mal an jedem Modell durch. Dabei verwendete man eine Schiebelehre für die Vermessung der Gipsmodelle und eine Computersoftware (OnyxCeph3™; Image Instruments, Chemnitz, Deutschland) für die Vermessung der virtuellen Modelle. Gepaarte t-Tests wurden für die Bewertung von Validität und Objektivität herangezogen, der Pearson-Korrelationskoeffizient für die Reliabilität und der Intraklassen-Korrelationskoeffizient (ICC) für die Wiederholbarkeit.

Ergebnisse Signifikante ($p<0{,}05$) Methodenunterschiede wurden bei 4 gemessenen Parametern beobachtet, jedoch überstieg lediglich die Differenz für Überbiss und Intermolarenabstand die klinische Relevanz (≥0,5 mm). Die Reliabilität beider Methoden war gegeben und der Methodenfehler zeigte sich für die digitale Messmethode und den erfahrenen Befundenden stets niedriger. Beide Befundenden erreichten ICCs >0,907 für beide Methoden. Signifikante Unterschiede zeigten sich zwischen den beiden Befundenden in allen Parametern außer der Intermolarendistanz bei der Vermessung der Gipsmodelle und in 3 Parametern nach der digitalen Auswertung.

Conclusio Die digitale Vermessung von Wechselgebissen erzielt klinisch verwertbare Ergebnisse, wenn sie mit Hilfe von OnyxCeph3™ durchgeführt wird. Beide Messmethoden zeigten eine hohe Reliabilität und Wiederholbarkeit. Jedoch spielt die Erfahrung des Befundenden eine wichtige Rolle.

Schlüsselwörter Zahnärztliche Modelltechniken · Virtuelle Bogenmodelle · Dreidimensionale Bildgebung · Intraoraler Scanner

Introduction

Orthodontics is going digital. In recent years, digitization has become an indispensable part of orthodontists' workflows. Paperless patient information systems, including digital photographs and radiographs, are today standard [33].

In addition, it is essential for comprehensive orthodontic diagnostics and treatment planning to analyse study casts that represent the patient's dentition. Mixed dentitions require high accuracy in evaluating arch length discrepancies (ALD) for successful orthodontic treatment, and space analysis in these situations is traditionally performed by contrasting the existing mesiodistal width of the supporting area with a nominal value [27].

The gold standard for diagnostic measurements is to use a calliper on a stone cast [3]. Study casts made of dental stone are, however, susceptible to damage and loss of accuracy by external forces [10]. Following research and development to address this issue, Cadent (now Align Technology, San Jose, CA, USA) introduced the first generation of OrthoCad™ software for "digital casts" in 1999 [15]. In 2006, this was followed by the iTero Element (Align Technology, San Jose, CA, USA) intraoral scanner (IOS), which uses parallel confocal imaging and point-and-stitch reconstruction to generate three-dimensional (3D) computerized images. Current software applications (e.g., OnyxCeph3™; Image Instruments, Chemnitz, Germany) are capable of managing, analysing, and generating virtual arch models.

Reported benefits of using an IOS include better patient comfort [43], reduced storage requirements for casts [29], ability to share data easily anywhere in the world [39], extensive possibilities of analysing digital models [39], and time efficiency [14, 26]. As a result, 3D imaging is today routinely used in orthodontic diagnostics and treatment planning. Young patients, in particular, tend to require maximum comfort and hence prefer the digital impression technique [5]. Digital study models can be generated in three different ways: directly by intraoral scanning, indirectly by surface scanning of stone casts, or based on a cone-beam computed tomography (CBCT) scan [40].

Previous studies have confirmed the utility of intraoral scanning in orthodontic diagnostics and planning [1, 2, 19]. Digital models obtained by extraoral scanning of stone casts have also been shown to offer both high accuracy compared to stone casts per se [10, 31, 42] and adequate precision for orthodontic applications [15, 39]. Numerous studies have confirmed the validity [2, 4, 14, 21, 28, 29, 42, 45], the reliability [4, 14, 15, 20, 21, 24, 34, 42], and the reproducibility [36, 42, 45] of measurements performed on dig-

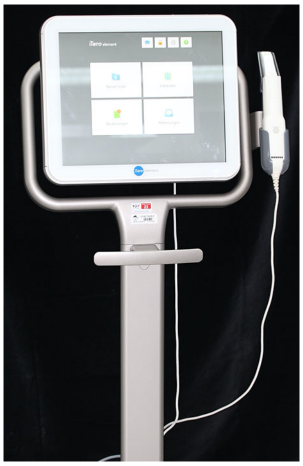


Fig. 1 Intraoral scanner (iTero Element 2; Align Technologies)
Abb. 1 Intraoralscanner (iTero Element 2; Align Technologies)

ital models versus on stone casts in permanent dentitions. Despite some findings of statistically significant differences between the methods, these measurement differences were clinically not relevant [7, 8, 20, 23, 24, 30, 34, 36]. Studies have also verified that digital measurements are clinically acceptable and are not inferior for treatment planning [30, 36, 41].

However, while the digital method is a clinically acceptable alternative to the analogue gold standard in analysing permanent dentitions, no comparative studies have been available for digital versus analogue measurements in children with mixed dentitions. These latter situations are different in that, rather than measuring all teeth, a limited number of permanent teeth need to be analysed along with the supporting area. Hence, longer distances must be measured when examining the supporting area, which might be more difficult to achieve accurately.

Evidence-based orthodontics relies on analysis methods known to offer *validity, reliability, reproducibility*, and *objectivity* for treatment planning to remain unaffected. The aim of validation is to prove that these requirements are met by an intended analysis technique in daily practice [39]. In this context, *validity* is defined as the truth value of evidence and whether what is measured equals what was intended to be measured, equating accuracy [35], *intermethod reliability* as the degree to which test scores are consistent when there is variation in the methods or instruments used [13] and *reproducibility*, being concerned with the consistency of evidence [12], as the closeness of agreement between the results of successive measurements of the same measure carried out under the same conditions [44].

Against this background, we designed the present study to assess measurements obtained on digital models as compared to the gold standard. Given the variance in how well a method performs depending on the experience of the examiner, we incorporated this factor into the study design to assess *objectivity*, defined as neutrality of evidence [12].

The null hypothesis of the study was that measurements on digital models would not show significant differences from that on conventional stone casts so that the digital method will be considered as accurate, reliable, reproducible, and objective.

Methods

Approval of the study design, which was in accordance with the 1975 Declaration of Helsinki as revised in 2013, was obtained by the Ethics Committee of the Medical University Innsbruck (Innsbruck, Austria; study ID: 1124/2021). Using G*power software (version 3.0.1; University of Düsseldorf, Germany), a sample size estimation was conducted as described elsewhere [16, 23, 25] to compare differences between the dependent groups by paired *t*-tests. Assuming mean differences of 0.5 mm with standard deviations < 1 mm as clinically relevant, a sample size of 30 pairs of stone casts was found to be required to achieve a power of 80%, with α set to 0.05 and β to 0.2.

Hence, we randomly retrieved 30 pre-existing pairs (upper and lower arch) of stone casts that represented mixed dentitions from the records of our Department of Orthodontics (Medical University of Innsbruck, Austria). All had been fabricated from ISO type 4 stone (Silky-Rock yellow; Whip Mix, Louisville, KY, USA). Based on dental age, we estimated that these fully anonymized orthodontic patients had been 7–13 years old.

We gathered this sample based on the criteria of mixed dentition (permanent incisors and first molars fully erupted), intact stone casts, and no previous orthodontic treatment. To ensure homogeneity of the sample, care was taken choosing only casts with fully erupted teeth. Exclusion criteria were syndromes.

Baseline measurements on the casts were performed using a vernier calliper (Zurich model; Dentaurum, Ispringen, Germany) to the nearest 0.1 mm. Then, the casts were mounted in a stand and digitized by one experienced operator using an IOS (iTero Element 2, version 1.12.0.990; Align Technology) as recommended by the manufacturer (Fig. 1). All scans were visually checked on screen, followed by re-scanning whenever a flaw was identified. Upper and lower casts were scanned separately, the scans saved as standard tessellation language (STL) files and uploaded to the internet for further transformation into digital casts.

All of the returned digital casts, accessible for download within 48 h, were measured applying the OnyxCeph3™ Pro analysis software (version 3.2.147; Image Instruments, Chemnitz, Germany) on the same computer (Fig. 2). This environment supports an effective workflow for orthodontic archiving, diagnostics, planning, and counselling, thus assisting orthodontists in optimizing administrative tasks and

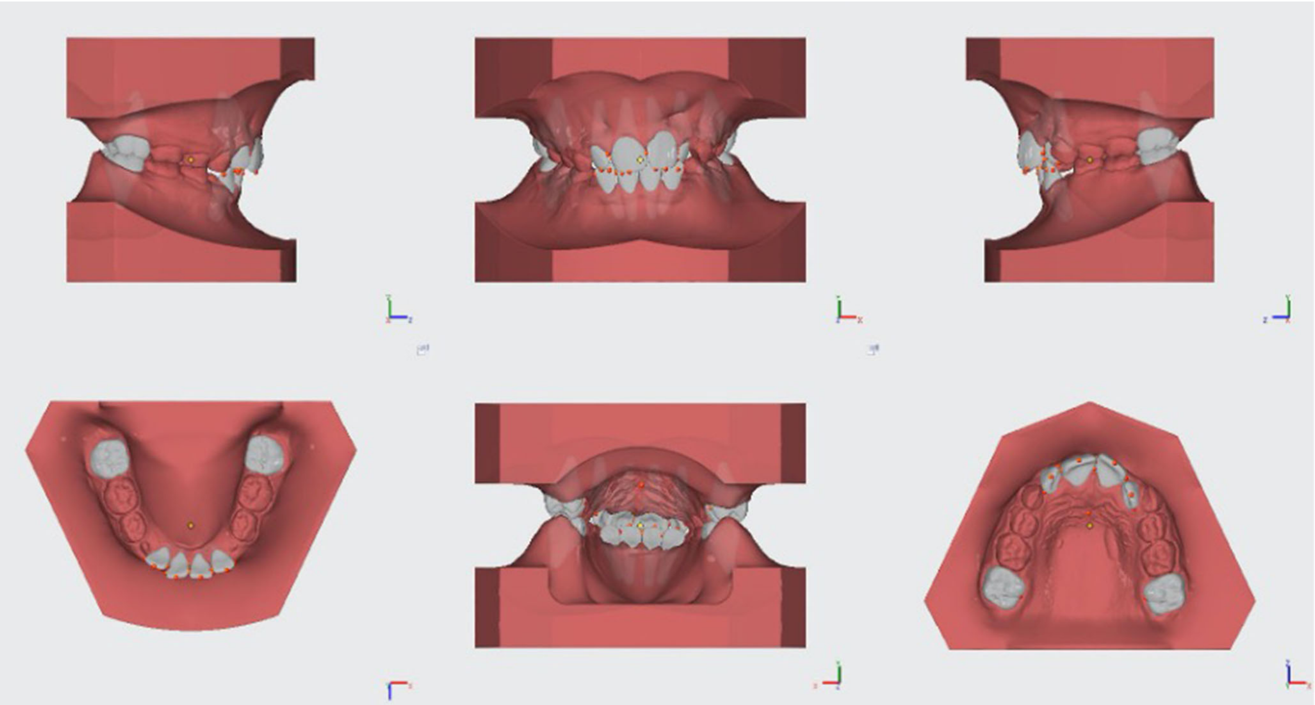
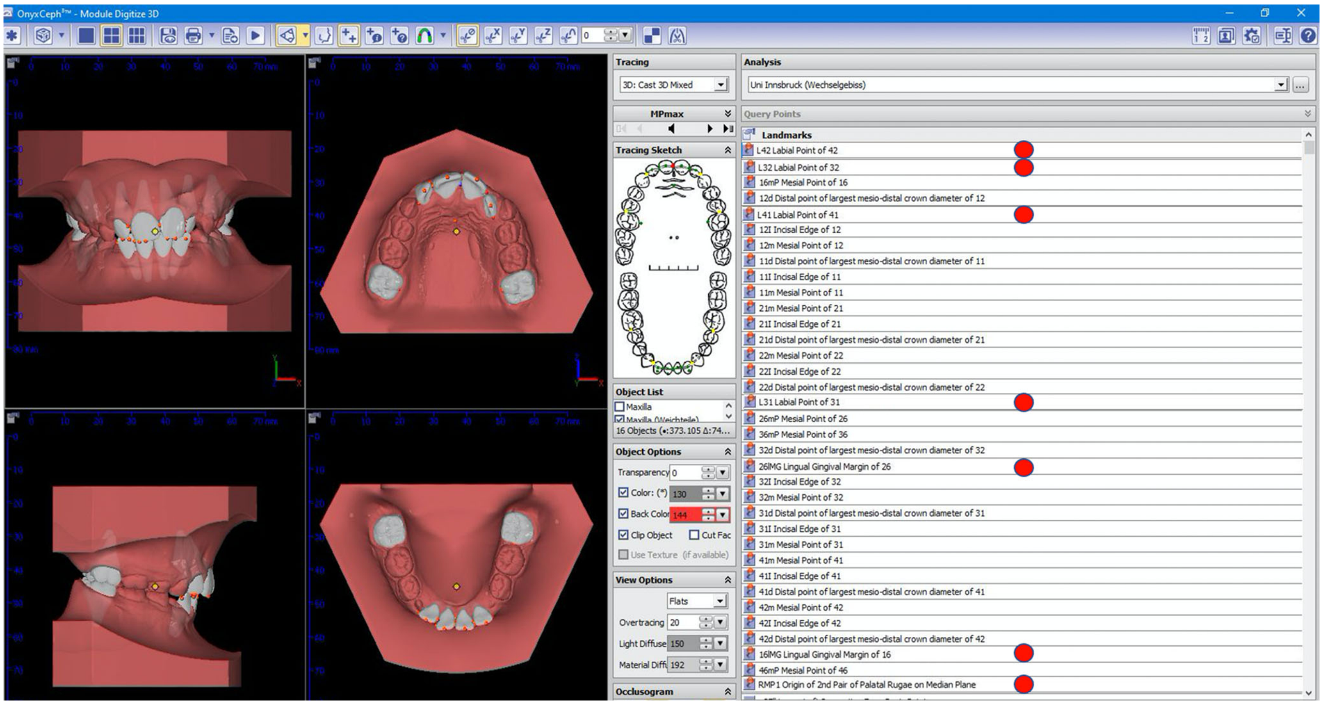


Fig. 2 Screenshots of a digital model stored and measured in OnyxCeph3^{TM} (Image Instruments) software

Abb. 2 Screenshots eines digital vermessenen Modells, Software OnyxCeph3^{TM} (Image Instruments)



Tracing: 3D: Cast 3D Mixed

Analysis: Uni Innsbruck (Wechselgebiss)

Query Points

Object List
☐ Maxilla
☑ Maxilla (Weichteile)
16 Objects (•:373.105 Δ:74...

Object Options
Transparency 0
☑ Color: (*) 130
☑ Back Color 144
☑ Clip Object ☐ Cut Fac
☐ Use Texture (if available)

View Options
Flats
Overtracing 20
Light Diffuse 150
Material Diffu 192

Landmarks
L42 Labial Point of 42 ●
L32 Labial Point of 32 ●
16mP Mesial Point of 16
12d Distal point of largest mesio-distal crown diameter of 12
L41 Labial Point of 41 ●
12I Incisal Edge of 12
12m Mesial Point of 12
11d Distal point of largest mesio-distal crown diameter of 11
11I Incisal Edge of 11
11m Mesial Point of 11
21m Mesial Point of 21
21I Incisal Edge of 21
21d Distal point of largest mesio-distal crown diameter of 21
22m Mesial Point of 22
22I Incisal Edge of 22
22d Distal point of largest mesio-distal crown diameter of 22
L31 Labial Point of 31 ●
26mP Mesial Point of 26
36mP Mesial Point of 36
32d Distal point of largest mesio-distal crown diameter of 32
26lMG Lingual Gingival Margin of 26 ●
32I Incisal Edge of 32
32m Mesial Point of 32
31d Distal point of largest mesio-distal crown diameter of 31
31I Incisal Edge of 31
31m Mesial Point of 31
41m Mesial Point of 41
41I Incisal Edge of 41
41d Distal point of largest mesio-distal crown diameter of 41
42m Mesial Point of 42
42I Incisal Edge of 42
42d Distal point of largest mesio-distal crown diameter of 42
16lMG Lingual Gingival Margin of 16 ●
46mP Mesial Point of 46
RMP1 Origin of 2nd Pair of Palatal Rugae on Median Plane ●

Fig. 3 A digital study cast portrayed in OnyxCeph3^{TM} (Image Instruments) dental software with measuring points set. The *red dots* indicate the landmarks which must be set by the operator

Abb. 3 Ein digitales Modell im OnyxCeph3^{TM} (Image Instruments) Bearbeitungsprogramm. In *rot* jene Punkte, die vom Befundenden selbst gesetzt werden müssen

implementing treatment. For the segmentation process, the middle of each tooth must be selected by the examiner to simulate the shape of each tooth. Then, the software offers automated identification of 20 landmarks (12 mesiodistal widths of incisors and molars, 8 upper and lower incisal edges) and requires input for seven more points, including one point on the palatal aspect of each upper molar, one on the labial aspect of each lower incisor, and the origin of the second pair of palatal rugae on the median plane (Fig. 3).

To assess objectivity, two examiners with different levels of experience performed the same measurements independently: an orthodontist with four years of practice and experienced in scanning (examiner 1), and a dental student in his last year of training without any experience in orthodontics or scanning (examiner 2). To calibrate reproducibility and minimize interexaminer variability, all measurements on stone casts and digital models were performed in triplicate by both examiners under identical conditions with intervals of at least one week.

To evaluate the validity, measurements for maxillary and mandibular space analysis were performed to the nearest 0.1 mm, recording and calculating for each model the amounts of available space, required space, and arch length discrepancy (ALD). Overjet, overbite, and upper intermolar distance were measured to the nearest 0.5 mm: overjet from the incisal edge of the most anterior upper incisor to the labial surface of the most anterior lower incisor; overbite as the longest vertical overlap between upper and lower incisors; maxillary intermolar distance (McNamara, [22]) between the lingual surfaces of both upper first molars. Clinical relevance was assessed by subtracting the mean differences of repeated measurements obtained on the stone casts from those on the digital models. Positive values of mean differences (Δ) indicated that the measurements on the stone casts were smaller than those on the digital models. In accordance with the literature, differences ≥ 0.5 mm were regarded as clinically relevant [34].

For validity assessment, only measurements of examiner 1 were taken into consideration, in order to avoid distorted results due to inexperience.

Statistics

Measurements were entered in a spreadsheet (Microsoft Excel, Redmond, WA, USA) and analysed in SPSS Statistics (v. 26; IBM SPSS Inc., Armonk, NY, USA). Displaying the variables along with their mean values (trueness) and standard deviations (precision), a Shapiro–Wilk test was used to check for normal distribution and Levene's test for homogeneity of variance. Additionally, 95% confidence intervals (CIs) were calculated. Mean repeated measurements for each examiner and method were compared using paired *t*-tests for validity and objectivity. Method errors were evaluated based on Dahlberg's formula and intermethod reliability by Pearson correlation coefficients (r). Reproducibility was assessed as intra-examiner repeatability of measurements, using intraclass correlation coefficients (ICCs), > 0.9 indicating excellent, 0.75 to < 0.9 good, 0.5 to < 0.75, moderate, and < 0.5 poor reproducibility. Differences emerging from any of these statistical tests were regarded as significant at $p < 0.05$ or as highly significant at $p < 0.01$.

Results

Validity

Table 1 summarizes the mean values, standard deviations, and the corresponding 95% confidence intervals (CI) for all measurements performed by the experienced examiner. Comparing the mean values obtained for each of the five parameters measured (each mean value representing three repeated measurements performed on either the stone casts or the digital models), we observed differences of +0.49 mm for ALD upper jaw (UJ), –0.44 mm for ALD lower jaw (LJ), +1.04 mm for overbite, –0.45 mm for overjet, and +1.18 mm for McNamara. With the exception of ALD UJ, all these differences between both methods were significant ($p < 0.05$) or highly significant ($p < 0.01$). However, only the parameters of overbite and McNamara exceeded the threshold of clinical significance (≥ 0.5 mm) with differences of 1.04

Table 1 Validity expressed as between-method differences in the hands of the experienced examiner
Tab. 1 Validität ausgedrückt als Unterschiede zwischen den Methoden, durchgeführt vom erfahrenen Untersuchenden

	Stone casts (*n* = 30)			Digital models (*n* = 30)				Paired *t*-test (*p*)
Parameter	Mean	SD	CI	Mean	SD	CI	Δ	
ALD UJ (mm)	0.82	2.13	0.02–1.62	0.33	2.85	–0.73–1.39	+0.49	0.28
ALD LJ (mm)	0.47	3.17	–0.71–1.65	0.91	2.98	–0.20–2.02	–0.44	< 0.05*
Overbite (mm)	3.43	1.81	2.75–4.11	2.39	1.82	1.71–3.07	+1.04	< 0.01**
Overjet (mm)	4.69	2.49	3.76–5.62	5.24	2.33	4.37–6.11	–0.45	< 0.01**
McNamara (mm)	33.52	1.75	32.87–34.17	32.34	2.05	31.57–33.11	+1.18	< 0.01**

ALD arch length discrepancy, *UJ* upper jaw, *LJ* lower jaw, *SD* standard deviation, *CI* confidence intervals, Δ difference of means from deviations
*Difference statistically significant at $p < 0.05$, **difference highly significant at $p < 0.01$

Table 2 Intermethod reliability expressed as measurement errors calculated by Dahlberg's formula for each examiner and as between-method correlations

Tab. 2 Inter-Methoden-Reliabilität ausgedrückt als Messfehler, berechnet nach der Dahlberg-Formel für jeden Untersuchenden und als Korrelationen zwischen den Methoden

	Stone casts ($n=30$)		Digital models ($n=30$)		PCC (r)
Parameter	Examiner 1 Dahlberg	Examiner 2 Dahlberg	Examiner 1 Dahlberg	Examiner 2 Dahlberg	
ALD UJ (mm)	0.53	0.56	0.07	0.25	0.499**
ALD LJ (mm)	0.25	0.74	0.13	0.19	0.918**
Overbite (mm)	0.03	0.17	0.02	0.06	0.836**
Overjet (mm)	0.06	0.1	0.04	0.07	0.826**
McNamara (mm)	0.18	0.23	0.08	0.14	0.806**

ALD arch length discrepancy, *UJ* upper jaw, *LJ* lower jaw, *PCC* Pearson correlation coefficient
**Difference highly significant at $p<0.01$

and 1.18 mm, respectively (Table 1). The differences for ALD UJ, ALD LJ, and overjet were not clinically relevant (<0.5 mm).

Thus, the null hypothesis regarding the accuracy of measurements obtained from digital casts had to be rejected, at least partly.

Intermethod reliability

Table 2 lists both examiners' random errors of repeated measurements. On the stone casts, the experienced examiner 1 produced errors within a range of 0.03 to 0.53 mm, while the inexperienced examiner 2 had a considerably harder time, with errors ranging from 0.1 to 0.75 mm. While both did invariably better on the digital models, examiner 2 again incurred higher errors for all five parameters. Given r-values of 0.499–0.918, a positive correlation between both measuring methods could be established for all parameters (Table 2).

Based on these results, the null hypothesis had to be accepted.

Reproducibility

Table 3 gives an overview of intraclass correlation coefficients, which were calculated as a measure of consistency which the experienced and the inexperienced examiner achieved in reproducing their results over repeated measurements. By and large, both examiners achieved high ICCs on the stone casts and on the digital models. The lowest ICCs associated with any of the parameters measured, regardless of the measuring methods, were 0.918 for examiner 1 and 0.666 for examiner 2. Overall, therefore, 0.666 was the lowest ICC. Examiner 1 was most consistent with the transversal measurement according to McNamara on the stone casts (ICC: 0.998) and with overbite on the digital models (ICC: 0.999) versus least consistent with ALD UJ (ICC: 0.918) and McNamara (ICC: 0.968), respectively. Examiner 2 was most consistent with overjet on the stone casts (ICC: 0.975) and ALD LJ on the digital models (ICC: 0.953) versus least consistent with overbite (ICC 0.931) and McNamara (ICC 0.907), respectively (Table 3).

Overall, the null hypothesis regarding the reproducibility of measurements for both methods had to be verified.

Objectivity

Table 4 summarizes all mean values of measurements and the calculated interexaminer differences, evaluated both by paired *t*-tests to determine statistical significance and by absolute differences to determine clinical relevance. The method on stone casts was associated with significant

Table 3 Reproducibility expressed as within-examiner agreement of repeated measurements

Tab. 3 Reproduzierbarkeit ausgedrückt als prüferinterne Übereinstimmung wiederholter Messungen

	Stone casts ($n=30$)	Digital models ($n=30$)	Stone casts ($n=30$)	Digital models ($n=30$)
Parameter	Examiner 1, ICC		Examiner 2, ICC	
ALD UJ	0.918	0.986	0.666	0.875
ALD LJ	0.952	0.987	0.934	0.953
Overbite	0.999	0.969	0.931	0.929
Overjet	0.982	0.979	0.975	0.944
McNamara	0.998	0.968	0.952	0.907

ALD arch length discrepancy, *UJ* upper jaw, *LJ* lower jaw, *ICC* intraclass correlation coefficient

Table 4 Objectivity expressed as between-examiner differences identified for each method
Tab. 4 Objektivität ausgedrückt als Unterschiede zwischen den Prüfenden, die für jede Methode identifiziert wurden

	Stone casts ($n=30$)			Paired *t*-test	Digital models ($n=30$)			Paired *t*-test
Parameter	Examiner 1 Mean	Examiner 2 Mean	Δ	*p*-value	Examiner 1 Mean	Examiner 2 Mean	Δ	*p*-value
ALD UJ (mm)	0.82	0.68	+0.14	<0.01**	0.33	0.12	+0.21	0.08
ALD LJ (mm)	0.47	0.03	+0.44	<0.01**	0.91	0.48	+0.43	<0.01**
Overbite (mm)	3.43	3.34	+0.09	<0.05*	2.39	2.46	–0.07	<0.01**
Overjet (mm)	4.69	4.43	+0.26	<0.01**	5.24	5.31	–0.07	0.29
McNamara (mm)	33.52	33.59	–0.07	0.12	32.34	34.17	–1.83	<0.01**

ALD arch length discrepancy, *UJ* upper jaw, *LJ* lower jaw, Δ difference of means from deviations
*Difference statistically significant at $p<0.05$, **difference highly significant at $p<0.01$

($p<0.05$) or highly significant ($p<0.01$) differences for all parameters except the transversal measurement according to McNamara ($p=0.12$). None of these differences exceeded the 0.5 mm threshold of clinical relevance. The interexaminer differences on digital models were significant for three parameters (ALD LJ, overbite, McNamara; $p<0.01$), but only one (McNamara: –1.83 mm) was clinically relevant (Table 4).

In consideration of these findings, the null hypothesis concerning the objectivity, had to be declined.

Discussion

The authors are not aware of any previous studies evaluating the validity, reliability, reproducibility, and objectivity of measurements performed on digital models of mixed dentitions. Because there is no direct way of determining the true dimensions of teeth, measurements on stone casts were considered the gold standard against which validity was to be evaluated.

Since the accuracy of digital models is known to depend on scanner types and intraoral conditions (e.g., saliva or restricted mouth opening) [6, 11], our use of surface scans performed on stone casts ruled out any loss of precision that may have been incurred by clinical intraoral capturing. In other words, none of the differences between stone casts and digital models we observed can be attributed to intraoral scanning distortion.

In the hands of the experience examiner, all mean differences in repeated measurements between the stone casts and the digital models were significant ($p<0.05$) or highly significant ($p<0.01$), the only exception being ALD UJ ($p=0.28$). However, clinically relevant differences were noted for only two parameters (overbite: 1.04 mm; transversal measurement according to McNamara: 1.18 mm). Differences in overbite measurements were most likely due to the delicate contact points on the digital images compared to the bulky calliper on the stone casts. Differences noted between the analogue and digital measurements of intermolar distance might be a function of the longer distance to be measured.

For three out of five measured parameters, our results, although based on a different method of digitization and different software (OnyxCeph3™), support previous findings that the validity of measurements on digital models (OrthoCad) is clinically acceptable [4, 20, 21, 28]. Radeke et al. [29], who previously compared measurements on stone casts and digital models with OnyxCeph3™, did report differences between software- and calliper-based measurements of mesiodistal tooth width but found them to be operator-associated and not statistically significant. Zilberman et al. [45] and Santoro et al. [34] found that stone casts resulted in slightly better measurements than digital models. Both studies relied on impressions sent to OrthoCad, so that any flaws during impression-taking or subsequent alginate shrinkage could have been passed on to the digital models. Alginate shrinkage has been previously reported to account for underestimated measurements on digital models, compared to the situation in vivo [24, 37], which might affect the validity.

Consistent with our own result that manual overbite measurements were significantly increased (Table 1), the manual method has repeatedly been implicated in exaggerated overbite measurements [34, 38]. Thus, in absolute terms of millimetres, smaller teeth should be expected to result in smaller overbite values, although inconsistent overbite measurements could also have been introduced by other factors such as incorrect probe angulation during manipulation of the traditional casts or rounding of the digital measurements to the nearest 0.5 mm. Another explanation might lie in digital measurements of tooth size consistently falling short of the values obtained on stone casts [23, 36]. In our study, this observation held true for the parameters ALD UJ, overbite, and the transversal measurement according to McNamara.

Notwithstanding, it stands to reason that the measuring tools of the OnyxCeph3™ software may allow for better defined measuring protocols resulting in more valid measurements than callipers used on stone casts. Further, automated identification of many landmarks requires the examiner merely to visually check these results (Fig. 3). Thus, one could speculate that inaccuracies of analogue measurements may be introduced by human error in the form of misreading callipers or misplacing reference points. That said, only the reference points for ALD UJ, ALD LJ, and overbite measurements are automatically set by OnxyCeph3™, whereas analogue versus digital analysis yielded significantly different mean values for all the parameters investigated but one (ALD UJ: $p = 0.28$).

Intermethod reliability indicates the agreement of results obtained with any two comparable measuring methods. A Pearson correlation coefficient of >0.499 testifies to a high degree of agreement between the two methods in our study (Table 2). Zilberman et al. [45] and Bell et al. [2] reported mean intra-examiner errors of 0.18 or 0.17 mm, indicating excellent results on both stone casts and digital models in line with our own findings.

Intra-examiner ICC values as calculated in the present study yield estimates of reproducibility under the model of equal marginal distributions. When marginal distributions are inaccurate, the deviations are captured and rated as unreliable. Based on all parameters, consistent agreement of repeated measurements on the stone casts emerged for the experienced examiner 1, given an ICC >0.918 as compared to >0.666 for the inexperienced examiner 2 (Table 3). All parameters measured by examiner 1 achieved ICCs >0.9 (= excellent reproducibility). Most of those performed by examiner 2 also exceeded the threshold of excellent reproducibility; one digitally measured parameter was rated as "good" (ICC: 0.8775) and one analogue parameter as "moderate" (ICC: 0.666). Overall, both examiners achieved very good ICCs. The fact that those for the experienced examiner were slightly higher suggest that different levels of orthodontic experience may have an influence on reproducibility.

The reproducibility of analogue measurements performed on stone casts may not only depend on an operator's ability to identify landmarks and to accurately transfer quantitative data to a computer, but there is a need for careful handling to avoid breakage of the casts. The reproducibility of digital measurements will also depend on the hardware and software used for digitization [18]. In the present study, excellent reproducibility was noted for measurements performed on both stone casts and digital models. Similar results have been reported previously by Wiranto et al. [42], Stevens et al. [36], and Czarnota et al. [8]. Measurements of arch length discrepancy were associated with slightly higher ICCs for the digital technique. Measurements of overbite, overjet, and the transversal measurement according to McNamara involved moderately higher ICCs for the calliper-based analogue technique. Arguably, this might be because the reference points for overjet and McNamara need to be manually selected on screen, thus being more prone to misjudgement, while all landmarks for tooth-width and space analysis are identified by OnyxCeph3™ on its own.

Four out of five interexaminer differences for measurements on stone casts were statistically significant (Table 4). All these differences, however, remained below the threshold of clinical relevance (<0.5 mm). Based on digital measurements, significant interexaminer disagreement was seen for three parameters (ALD LJ, overbite, transversal measurement according to McNamara) and reached clinical relevance for McNamara (–1.83 mm). Given this nearly complete absence of clinically relevant differences between both examiners, both methods may be regarded as objective. The only exception just mentioned, a large interexaminer difference between the digital measurements of upper intermolar distance, may be attributed to operator experience, as the landmarks for this parameter need to be identified manually in OnyxCeph3™. This finding is consistent with Radeke et al. [29]. Intermolar distance remains a major problem of intraoral scanners, with large deviations incurred by the matching/stitching algorithm [17, 32].

Our data further support findings of Dalstra et al. [9] about the effect that interexaminer variation was generally low. Rheude et al. [30] investigated how orthodontic experience might influence treatment planning. In accordance with our own results, they found statistically significantly differences in diagnostic decisions, which did not, however, result in different treatment decisions. It should be noted that the present study featured an inexperienced examiner solely to identify different measurements as a function of orthodontic expertise.

Limitations to our study arise from the fact that the intraoral scanner was used on stone casts rather than in patients' mouths and examination was done retrospectively. Though, the sample size was checked conscientiously, it might be limited for a study of diagnostic accuracy.

Conclusions

Digital models of permanent dentitions are today an accepted alternative to stone casts in orthodontics. Driven by both an emerging aesthetic idealism and intraoral scanners eliminating the need for impression-taking, the demand for orthodontic treatment of paediatric patients has been growing in recent years. Intraoral scanners can minimize discomfort, which is essential in children. This preliminary study confirms that digital models (here generated based on surface scans of physical casts) can be a reliable alternative to stone casts in analysing mixed dentitions. Measurements performed on these virtual models yield reproducible intra- and interarch relationships, but the results are affected by operator experience. We cannot confirm that longer distances have a greater deviation between measurements, even though the transversal measurement according to McNamara is one long distance to measure, but the space analysis requires many small measurements. Well-designed investigations dealing with intraoral scanning followed by digital model generation are needed to assess the time requirements for these procedures in young patients to evaluate the feasibility of this digital work-flow.

Author Contribution L. Schieffer: concept of experiments and writing the manuscript. A.G. Crismani: concept of experiments and supervision of the study. L. Schieffer and M. Paulus: experimental work. L. Schieffer and H. Ulmer: data analysis. All authors contributed to the article and approved the submitted version.

Funding Open access funding provided by University of Innsbruck and Medical University of Innsbruck.

Declarations

Conflict of interest L. Schieffer, L. Latzko, H. Ulmer, N. Schenz-Spisic, U. Lepperdinger, M. Paulus and A.G. Crismani declare that they have no competing interests.

Ethical standards All procedures performed in studies involving human participants or on human tissue were in accordance with the ethical standards of the institutional and/or national research committee and with the 1975 Helsinki declaration and its later amendments or comparable ethical standards. Approval was obtained from the ethics committee of the Medical University of Innsbruck.

References

1. Akyalcin S, Cozad BE, English JD, Colville CD, Laman S (2013) Diagnostic accuracy of impression-free digital models. Am J Orthod Dentofacial Orthop 144:916–922. https://doi.org/10.1016/j.ajodo.2013.04.024
2. Bell A, Ayoub AF, Siebert P (2003) Assessment of the accuracy of a three-dimensional imaging system for archiving dental study models. J Orthod 30:219–223. https://doi.org/10.1093/ortho/30.3.219
3. Bolton WA (1962) The clinical application of a tooth-size analysis. Am J Orthod 48:504–529. https://doi.org/10.1016/0002-9416(62)90129-X
4. Bootvong K, Liu Z, McGrath C, Hägg U, Wong RWK, Bendeus M, Yeung S (2010) Virtual model analysis as an alternative approach to plaster model analysis: reliability and validity. Eur J Orthod 32:589–595. https://doi.org/10.1093/ejo/cjp159
5. Burhardt L, Livas C, Kerdijk W, van der Meer WJ, Ren Y (2016) Treatment comfort, time perception, and preference for conventional and digital impression techniques: a comparative study in young patients. Am J Orthod Dentofacial Orthop 150:261–267. https://doi.org/10.1016/j.ajodo.2015.12.027
6. Burzynski JA, Firestone AR, Beck FM, Fields HW, Deguchi T (2018) Comparison of digital intraoral scanners and alginate impressions: time and patient satisfaction. Am J Orthod Dentofacial Orthop 153:534–541. https://doi.org/10.1016/j.ajodo.2017.08.017
7. Camardella LT, Breuning H, de Vasconcellos Vilela O (2017) Accuracy and reproducibility of measurements on plaster models and digital models created using an intraoral scanner. J Orofac Orthop 78:211–220. https://doi.org/10.1007/s00056-016-0070-0
8. Czarnota J, Hey J, Fuhrmann R (2016) Measurements using orthodontic analysis software on digital models obtained by 3D scans of plaster casts: Intrarater reliability and validity. J Orofac Orthop 77:22–30. https://doi.org/10.1007/s00056-015-0004-2
9. Dalstra M, Melsen B (2009) From alginate impressions to digital virtual models: accuracy and reproducibility. J Orthod 36:36–41. https://doi.org/10.1179/14653120722905
10. Fleming P, Marinho V, Johal A (2011) Orthodontic measurements on digital study models compared with plaster models: a systematic review. Orthod Craniofac Res 14:1–16. https://doi.org/10.1111/j.1601-6343.2010.01503.x
11. Flügge TV, Schlager S, Nelson K, Nahles S, Metzger MC (2013) Precision of intraoral digital dental impressions with iTero and extraoral digitization with the iTero and a model scanner. Am J Orthod Dentofacial Orthop 144:471–478. https://doi.org/10.1016/j.ajodo.2013.04.017
12. Frambach JM, van der Vleuten CP, Durning SJ (2013) Quality criteria in qualitative and quantitative research. Acad Med 88:552. https://doi.org/10.1097/ACM.0b013e31828abf7f
13. Franco AR, Pritchard A, Calhoun VD, Mayer AR (2009) Interrater and intermethod reliability of default mode network selection. Hum Brain Mapp 30:2293–2303. https://doi.org/10.1002/hbm.20668
14. Gracco A, Buranello M, Cozzani M, Siciliani G (2007) Digital and plaster models: a comparison of measurements and times. Prog Orthod 8:252–259
15. Gül Amuk N, Karsli E, Kurt G (2019) Comparison of dental measurements between conventional plaster models, digital models obtained by impression scanning and plaster model scanning. Int Orthod 17:151–158. https://doi.org/10.1016/j.ortho.2019.01.014
16. Keating AP, Knox J, Bibb R, Zhurov AI (2008) A comparison of plaster, digital and reconstructed study model accuracy. J Orthod 35:191–201. https://doi.org/10.1179/146531207225022626
17. Keul C, Güth JF (2020) Accuracy of full-arch digital impressions: an in vitro and in vivo comparison. Clin Oral Investig 24:735–745. https://doi.org/10.1007/s00784-019-02965-2
18. Koretsi V, Tingelhoff L, Proff P, Kirschneck C (2018) Intra-observer reliability and agreement of manual and digital orthodontic model analysis. Eur J Orthod 40:52–57. https://doi.org/10.1093/ejo/cjx040

19. Lecocq G (2016) Digital impression-taking: fundamentals and benefits in orthodontics. Int Orthod 14:184–194. https://doi.org/10.1016/j.ortho.2016.03.003
20. Leifert MF, Leifert MM, Efstratiadis SS, Cangialosi TJ (2009) Comparison of space analysis evaluations with digital models and plaster dental casts. Am J Orthod Dentofacial Orthop 136:16.e1–16.e4. https://doi.org/10.1016/j.ajodo.2008.11.019
21. Mayers M, Firestone AR, Rashid R, Vig KWL (2005) Comparison of peer assessment rating (PAR) index scores of plaster and computer-based digital models. Am J Orthod Dentofacial Orthop 128:431–434. https://doi.org/10.1016/j.ajodo.2004.04.035
22. McNamara JA (2000) Maxillary transverse deficiency. Am J Orthod Dentofacial Orthop 117:567–570. https://doi.org/10.1016/s0889-5406(00)70202-2
23. Mullen SR, Martin CA, Ngan P, Gladwin M (2007) Accuracy of space analysis with emodels and plaster models. Am J Orthod Dentofacial Orthop 132:346–352. https://doi.org/10.1016/j.ajodo.2005.08.044
24. Naidu D, Scott J, Ong D, Ho CTC (2009) Validity, reliability and reproducibility of three methods used to measure tooth widths for bolton analyses. Aust Orthod J 25:97–103
25. Okunami TR, Kusnoto B, BeGole E, Evans CA, Sadowsky C, Fadavi S (2007) Assessing the American board of orthodontics objective grading system: digital vs plaster dental casts. Am J Orthod Dentofacial Orthop 131:51–56. https://doi.org/10.1016/j.ajodo.2005.04.042
26. Patzelt SBM, Lamprinos C, Stampf S, Att W (2014) The time efficiency of intraoral scanners. J Am Dent Assoc 145:542–551. https://doi.org/10.14219/jada.2014.23
27. Proffit WR, Ackerman JL (2000) Orthodontic diagnosis: the development of a problem list. Contemporary orthodontics, 3rd edn. Mosby, St. Louis, pp 165–170
28. Quimby ML, Vig KWL, Rashid RG, Firestone AR (2004) The accuracy and reliability of measurements made on computer-based digital models. Angle Orthod 74:298–303. https://doi.org/10.1043/0003-3219(2004)074%3C0298:taarom%3E2.0.co;2
29. Radeke J, von der Wense C, Lapatki BG (2014) Comparison of orthodontic measurements on dental plaster casts and 3D scans. J Orofac Orthop 75:264–274. https://doi.org/10.1007/s00056-014-0217-9
30. Rheude B, Sadowsky PL, Ferriera A, Jacobson A (2005) An evaluation of the use of digital study models in orthodontic diagnosis and treatment planning. Angle Orthod 75:300–304. https://doi.org/10.1043/0003-3219(2005)75[300:aeotuo]2.0.co;2
31. Rossini G, Parrini S, Castroflorio T, Deregibus A, Debernardi CL (2016) Diagnostic accuracy and measurement sensitivity of digital models for orthodontic purposes: a systematic review. Am J Orthod Dentofacial Orthop 149:161–170. https://doi.org/10.1016/j.ajodo.2015.06.029
32. Schmidt A, Klussmann L, Wöstmann B, Schlenz MA (2020) Accuracy of digital and conventional full-arch impressions in patients: an update. J Clin Med 9:688
33. Sandler PJ, Murray AM, Bearn D (2002) Digital records in orthodontics. Dent Update 29:18–24. https://doi.org/10.12968/denu.2002.29.1.18
34. Santoro M, Galkin S, Teredesai M, Nicolay OF, Cangialosi TJ (2003) Comparison of measurements made on digital and plaster models. Am J Orthod Dentofacial Orthop 124:101–105. https://doi.org/10.1016/s0889-5406(03)00152-5
35. Stausberg J, Kuklik N, Jöckel KH (2018) Opportunities and pitfalls in the definition of data validity. Stud Health Technol Inform 247:566–570
36. Stevens DR, Flores-Mir C, Nebbe B, Raboud DW, Heo G, Major PW (2006) Validity, reliability, and reproducibility of plaster vs digital study models: comparison of peer assessment rating and Bolton analysis and their constituent measurements. Am J Orthod Dentofacial Orthop 129:794–803. https://doi.org/10.1016/j.ajodo.2004.08.023
37. Torassian G, Kau CH, English JD, Powers J, Bussa HI, Salas-Lopez AM, Corbett JA (2010) Digital models vs plaster models using alginate and alginate substitute materials. Angle Orthod 80:474–481. https://doi.org/10.2319/072409-413.1
38. Verma R, Singh SP, Verma S, Kumar V, Bhupali NR, Arora S (2019) Comparison of reliability, validity, and accuracy of linear measurements made on pre- and posttreatment digital study models with conventional plaster study models. J Orthod Sci 8:18. https://doi.org/10.4103/jos.jos_14_19
39. Vogel AB, Kilic F, Schmidt F, Rübel S, Lapatki BG (2015) Dimensional accuracy of jaw scans performed on alginate impressions or stone models: a practice-oriented study. J Orofac Orthop 76:351–365. https://doi.org/10.1007/s00056-015-0296-2
40. Wesemann C, Muallah J, Mah J, Bumann A (2017) Accuracy and efficiency of full-arch digitalization and 3D printing: a comparison between desktop model scanners, an intraoral scanner, a CBCT model scan, and stereolithographic 3D printing. Quintessence Int 48:41–50. https://doi.org/10.3290/j.qi.a37130
41. Whetten JL, Williamson PC, Heo G, Varnhagen C, Major PW (2006) Variations in orthodontic treatment planning decisions of class II patients between virtual 3-dimensional models and traditional plaster study models. Am J Orthod Dentofacial Orthop 130:485–491. https://doi.org/10.1016/j.ajodo.2005.02.022
42. Wiranto MG, Engelbrecht WP, Tutein Nolthenius HET, van der Meer WJ, Ren Y (2013) Validity, reliability, and reproducibility of linear measurements on digital models obtained from intraoral and cone-beam computed tomography scans of alginate impressions. Am J Orthod Dentofacial Orthop 143:140–147. https://doi.org/10.1016/j.ajodo.2012.06.018
43. Yuzbasioglu E, Kurt H, Turunc R, Bilir H (2014) Comparison of digital and conventional impression techniques: evaluation of patients' perception, treatment comfort, effectiveness and clinical outcomes. BMC Oral Health 14:10. https://doi.org/10.1186/1472-6831-14-10
44. Zhao Y, Sampson MG, Wen X (2020) Quantify and control reproducibility in high-throughput experiments. Nat Methods 17:1207–1213. https://doi.org/10.1038/s41592-020-00978-4
45. Zilberman O, Huggare JAV, Parikakis KA (2003) Evaluation of the validity of tooth size and arch width measurements using conventional and three-dimensional virtual orthodontic models. Angle Orthod 73:301–306. https://doi.org/10.1043/0003-3219(2003)073%3C0301:eotvot%3E2.0.co;2