Reproduzierbarkeit bei der Bewertung von kieferorthopädischen Plattenapparaturen

Holm M[1], Reißmann DR[2], Conradt K[1], Nagel T[1],
Kassem W[1], Jost-Brinkmann P-G[1], Sierwald I[1]

[1] Abteilung für Kieferorthopädie, Orthodontie und Kinderzahnmedizin
Charité – Universitätsmedizin Berlin

[2] Poliklinik für Zahnärztliche Prothetik, Universitätsklinikum Hamburg-Eppendorf, Hamburg

Ziel

Ziel der Untersuchung war die Beurteilung der Reproduzierbarkeit der Bewertung von kieferorthopädischen Plattenapparaturen durch Zahnärzte im Rahmen des kieferorthopädischen Technikkurses.

Material und Methode

Insgesamt 38 Studierende des ersten klinischen Semesters stellten jeweils zwei gleiche kieferorthopädische Plattenapparaturen (N = 76) her (Abb. 1 und 2), die von vier Weiterbildungsassistenten für Kieferorthopädie beurteilt wurden.

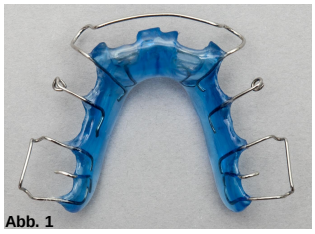

Abb. 1

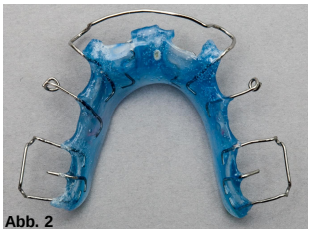

Abb. 2

Die Kriterien Ausarbeitung (Frage 1), Politur (Frage 2) und Fehlerfreiheit des Kunststoffs (Frage 3) sowie Gesamteindruck (Frage 4) wurden anhand eines Notensystems von 1 („optimal“) bis 6 („schlecht“) beurteilt (Abb. 3).

Bewerter:_____ Datum:_____

Platten-ID:_____

	Note					
	optimal					schlecht
	1	2	3	4	5	6
Ausarbeitung	○	○	○	○	○	○
Politur	○	○	○	○	○	○
Fehlerfreiheit des Kunststoffs	○	○	○	○	○	○
Gesamteindruck	○	○	○	○	○	○

Abb. 3

Die Gesamtnote wurde als gerundeter Mittelwert der vier gleichwertigen Einzelkriterien berechnet. Im Abstand von sieben Tagen erfolgte eine erneute Bewertung nach denselben Kriterien. Die Weiterbildungsassistenten wurden vor Durchführung der Bewertung nicht kalibriert. Insgesamt lagen 608 Bewertungen für die Analysen vor. Die Reproduzierbarkeit der Gesamt- und der Einzelnoten der einzelnen Bewerter (Intra-Rater-Reliabilität) wurde mittels gewichtetem Kappa bestimmt. Die Übereinstimmung zwischen den Bewertern (Inter-Rater-Reliabilität) erfolgte durch Berechnung eines Intraclass Correlation Coefficients (ICC) basierend auf einer zwei-faktoriellen Varianzanalyse (ANOVA).

Ergebnisse

Die Bewerter urteilten mit unterschiedlicher Strenge. Die genutzten Notenbereiche reichten von 1-3 bis 1-5 und die entsprechenden Durchschnittsnoten von 1,5 bis 3,0 (Tab. 1).

Tab. 1: Durchschnittliche Einzel- und Gesamtnoten der Untersucher

	Alle Untersucher	Untersucher				
		# 1	# 2	# 3	# 4	
	Mittelwert (SD) [Bereich]					P-Wert*
Frage 1	2,1 (1,0) [1 - 5]	1,5 (0,6) [1 - 3]	2,4 (0,7) [1 - 5]	2,9 (0,9) [1 - 5]	1,4 (0,6) [1 - 3]	<0,001
Frage 2	2,1 (0,9) [1 - 5]	1,4 (0,6) [1 - 3]	2,5 (0,9) [1 - 5]	2,8 (0,8) [1 - 5]	1,8 (0,6) [1 - 3]	<0,001
Frage 3	2,6 (1,1) [1 - 6]	1,8 (0,7) [1 - 4]	2,9 (1,2) [1 - 5]	2,9 (1,0) [1 - 6]	2,6 (1,0) [1 - 5]	<0,001
Frage 4	2,3 (0,9) [1 - 5]	1,5 (0,6) [1 - 3]	2,6 (0,7) [1 - 4]	3,0 (0,9) [1 - 5]	1,9 (0,6) [1 - 4]	<0,001
Gesamt	2,3 (0,9) [1 - 5]	1,5 (0,6) [1 - 3]	2,6 (0,7) [1 - 4]	3,0 (0,8) [1 - 5]	2,0 (0,6) [1 - 3]	<0,001
* Kruskal-Wallis-Test						

Etwa 2/3 (65,5 %) der Gesamtnoten unterschieden sich nicht zwischen beiden Bewertungsdurchgängen. Abweichungen bestanden nur um eine Note. Die Übereinstimmungen bei den Einzelnoten reichten von 57 % (Fehlerfreiheit) bis 68 % (Gesamteindruck) (Abb. 4).

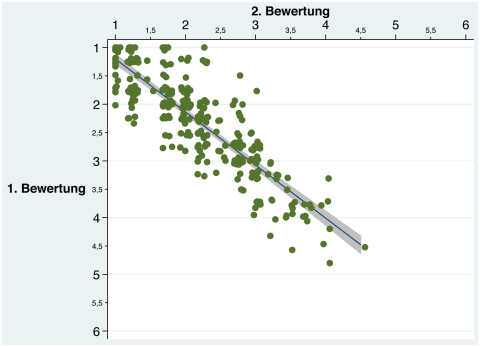

Abb. 4: Zusammenhang zwischen Durchschnittsnoten der ersten und zweiten Bewertung mit Regressionsgerade und 95%-Konfidenzintervall

Die Intra-Rater-Reliabilität war sowohl für die Gesamtnoten (gewichtetes Kappa = 0,79) als auch für die Einzelnoten (gewichtetes Kappa = 0,75 bis 0,81) ausgezeichnet. Demgegenüber war die Inter-Rater-Reliabilität für die Gesamtnoten (ICC = 0,47) nur mittelmäßig und bei den Einzelnoten (ICC = 0,27 bis 0,55) zum Teil gering.

Schlussfolgerung

Insgesamt zeigte eine erneute Bewertung der Apparaturen eine gute Reliabilität. Wenn die Beurteilung von Arbeiten auf mehrere Bewerter aufgeteilt wird, so sollten diese kalibriert werden, um eine Vergleichbarkeit der Noten zu gewährleisten.