Untersuchungen zur Optionsgewichtung als Methode für die Erfassung von Teilwissen in Multiple-Choice-Tests

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Birk Diedenhofen

aus Siegburg

Düsseldorf, Dezember 2015

Aus dem Institut für Experimentelle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Jochen Musch
Korreferent: Prof. Dr. Axel Buchner

Tag der mündlichen Prüfung: 26.01.2016

Inhaltsverzeichnis

Zusammenfassung **4**

Abstract **7**

1 Einleitung **9**

2 Optionsgewichtung **13**

3 Artikel 1: Eine experimentelle Untersuchung zur Validität der empirischen Optionsgewichtung **22**

4 Artikel 2: Eine Substichproben-Validierung zum Vergleich der empirischen Optionsgewichtung und der Experten-Optionsgewichtung **31**

5 Artikel 3: Ein R-Paket und Web-Interface zum Vergleich von abhängigen und unabhängigen Korrelationen **41**

6 Allgemeine Diskussion **50**

Literaturverzeichnis **56**

Anhang: Artikel **63**

Zusammenfassung

Multiple-Choice-Tests werden üblicherweise nach der Anzahl-Korrekt-Auswertung bepunktet, bei der 1 Punkt für die Wahl der richtigen Antwortoption und 0 Punkte für die Wahl eines Distraktors vergeben werden. Diese Auswertung des Multiple-Choice-Testformats wurde jedoch vielfach kritisiert, da keine Teilpunkte gewährt werden und nützliche Informationen über das Teilwissen der Testteilnehmer unberücksichtigt bleiben. Die Optionsgewichtung ist ein alternatives Auswertungsverfahren für Multiple-Choice-Tests, das für jede Antwortoption eines Items ein spezifisches Gewicht vergibt. Der Testscore eines Testteilnehmers berechnet sich aus der Summe der Optionsgewichte aller im Test ausgewählter Antwortoptionen. Diese individuelle Bepunktung der Antwortoptionen ermöglicht es, Teilpunkte auch für die Wahl eines Distraktors zu vergeben, wenn dieser Teilwissen widerspiegelt. Die Optionsgewichtung kann damit potenziell mehr Informationen aus einem einzelnen Item gewinnen als eine konventionelle Anzahl-Korrekt-Auswertung. Optionsgewichte lassen sich einerseits empirisch über die punkt-biseriale Korrelation zwischen den Optionswahlen und den Scores der Testteilnehmer ermitteln; andererseits können Optionsgewichte von Experten auf dem Gebiet des Tests bestimmt werden. Bisherige korrelative Studien fanden, dass die empirische Optionsgewichtung die Reliabilität eines Multiple-Choice-Tests gegenüber einer Anzahl-Korrekt-Auswertung verbesserte. Hinsichtlich der Validität blieb die Befundlage jedoch unklar. In Studie 1 der vorliegenden Arbeit wurde zum ersten Mal ein experimenteller Ansatz zur Validierung der empirischen Optionsgewichtung verfolgt. Als Außenkriterium wurden dazu unter den Studienteilnehmern unterschiedliche Wissensstände induziert, indem die Teilnehmer Sachtexte zu einem unvertrauten Thema lasen, deren Informationsgehalt zwischen drei Experimentalgruppen variierte. Die Teilnehmer aller drei Gruppen absolvierten anschließend den gleichen Wissenstest über die Inhalte der Sachtexte und beantworteten Multiple-Choice-Items, deren Antwortoptionen unterschiedlich viel Wissen zur Bewertung ihrer Korrektheit erforderten. Im Vergleich zur konventionellen Anzahl-Korrekt-Auswertung verbesserte die empirische

Optionsgewichtung sowohl die Reliabilität als auch die Validität des Wissenstests. Frühere Studien zu der Frage, ob Optionsgewichte empirisch oder auf der Grundlage von Expertenurteilen bestimmt werden sollten, lieferten kein schlüssiges Befundmuster. In den Untersuchungen wurde bei der Evaluation der empirischen Optionsgewichtung allerdings nur eine einzige Stichprobenteilung zur Kreuzvalidierung durchgeführt, ohne zu berücksichtigen, welchen Einfluss die Teilung der Stichprobe auf die Ergebnisse hatte. Die Wahl, welche der vielen möglichen Stichprobenteilungen ausgewertet und berichtet wird, muss jedoch als *Forscherfreiheitsgrad* (*researcher degree of freedom*; Simmons, Nelson & Simonsohn, 2011) betrachtet werden. Zum Vergleich der beiden Optionsgewichtungsverfahren untereinander und mit einer Anzahl-Korrekt-Auswertung, kam aus diesem Grund in Studie 2 erstmalig, neben einer konventionellen Analyse, eine wiederholte randomisierte Substichproben-Validierung zum Einsatz, bei der die Resultate von 10,000 Stichprobenteilungen aggregiert wurden. Die Ergebnisse der konventionellen Analyse hingen stark von der vorgenommenen Teilung der Stichprobe ab. Die über viele Stichprobenteilungen aggregierten Ergebnisse der Substichproben-Validierung ergaben dagegen eindeutig, dass die empirischen Optionsgewichte – im Gegensatz zu Experten-Optionsgewichten – die Testgüte eines Wissenstests im Vergleich zur Anzahl-Korrekt-Auswertung erhöhten. Die Verbesserung beschränkte sich jedoch auf die Reliabilität des Tests. Zusammengenommen legen die Ergebnisse der vorliegenden Arbeit nahe, dass Optionsgewichte neben der Reliabilität auch potenziell die Validität von Multiple-Choice-Tests verbessern können, wenn die Distraktoren der Items zwischen unterschiedlichen Fähigkeitsniveaus differenzieren. Anstatt Experten zu befragen, sollten Optionsgewichte empirisch bestimmt werden. Auf Wissensinduktion basierende experimentelle Validierungsmethoden werden als Verbesserung bisheriger korrelativer Ansätze empfohlen. Zukünftige Evaluationen der empirischen Optionsgewichtung sollten Kreuzvalidierungsmethoden einsetzen, die über eine Vielzahl von Stichprobenteilungen aggregieren. Da die meisten gängigen Statistikprogramme die Durchführung von Signifikanztests für abhängige und unabhängige Cronbach-α-

Koeffizienten sowie die Durchführung von Signifikanztests für abhängige und unabhängige Korrelationen nicht unterstützen, wurden im Rahmen der Arbeit R-Pakete und leicht benutzbare Web-Interfaces entwickelt, die frei zur Verfügung gestellt werden.

Abstract

Multiple-choice tests are generally scored using number-right scoring by awarding 1 point for choosing the correct answer and 0 points for choosing a distractor. This scoring procedure for the multiple-choice test format has often been criticized for not granting partial credit and for leaving out valuable information on the test takers' partial knowledge. Option weighting is an alternative scoring procedure for multiple-choice tests that assigns individual weights to each answer option. The score of a test taker is calculated by summing the option weights of all answer options that were selected in the test. This individual scoring of answer options allows granting partial credit for the choice of a distractor if it reflects partial knowledge. Option weighting captures potentially more information from a single item than conventional number-right scoring. Option weights may be determined empirically by calculating the point-biserial correlation between the option choices and scores of the test takers; alternatively, option weights can be assigned by experts in the domain of the test. Extant correlational studies found that empirical option weighting improved test reliability in comparison to number-right scoring. With regard to validity, results were however ambiguous. Study 1 of the present thesis pursued the first experimental approach to validate empirical option weighting. As an external criterion, different levels of knowledge were induced among three groups of participants by presenting essays about an unfamiliar topic with varying amounts of information. Subsequently, participants in all three groups completed the same knowledge test covering the content of the essays, and responded to multiple-choice items with answer options that required different amounts of knowledge to assess their correctness. Compared to conventional number-right scoring, empirical option weighting improved the reliability and validity of the knowledge test. Results of previous studies that investigated whether option weights should be determined empirically or on the basis of expert judgements remained inconclusive. When evaluating empirical option weights, however, the studies performed only a single sample

split for cross-validation, without considering the influence of the sample split on the results. Moreover, the decision which of the many sample splits is evaluated and reported must be considered a *researcher degree of freedom* (Simmons, Nelson, & Simonsohn, 2011). For this reason, to compare both option weighting procedures with each other and with number-right scoring, a repeated randomized subsampling validation was conducted in Study 2 that aggregated the results of 10,000 sample splits, in addition to a conventional analysis. The results of the conventional analysis strongly depended on the sample split that was performed. In contrast, the subsampling validation that aggregated the results of many sample splits clearly showed that empirical option weights – rather than expert option weights – improved the psychometric properties of a knowledge test in comparison with number-right scoring. This improvement was, however, limited to the reliability of the test. Taken together, the results of the present thesis suggest that option weighting may improve not only the reliability but also potentially the validity of a multiple-choice test if the distractors of the items differentiate between different ability levels. Instead of consulting experts, options weights should be determined empirically. Experimental validations based on the induction of knowledge are recommended as an improvement over correlational validation approaches. Future evaluations of empirical option weighting should use cross-validation procedures that aggregate across many sample splits. Most of the conventional statistical software packages support neither significance tests for dependent or independent correlations, nor significance tests for dependent or independent internal consistencies according to Cronbach. R software packages and easy-to-use web interfaces supporting such tests were therefore developed and made freely available.

1 Einleitung

Das Multiple-Choice-Testformat zählt zu den beliebtesten Antwortformaten in der Wissensdiagnostik. Das bereits seit Beginn des 20. Jahrhunderts eingesetzte Testformat besteht in seiner klassischen Form aus einer Frage oder einem Satzstamm und zwei oder mehr Antwortalternativen, die zur Auswahl stehen (Downing & Haladyna, 2006). Die Aufgabe der Testteilnehmer ist es, die gestellte Frage über die Wahl einer der verfügbaren Antwortoptionen zu beantworten. Dabei ist immer nur eine der Antwortalternativen richtig; bei den restlichen falschen Antwortoptionen handelt es sich um Distraktoren.

Die Popularität des Multiple-Choice-Formats lässt sich vermutlich auf seine zahlreichen vorteilhaften Eigenschaften zurückführen. Das Antwortformat ist vielseitig einsetzbar. Tests zur Messung verschiedenster kognitiver Fähigkeiten – von niedrigen bis zu hohen kognitiven Prozessen – lassen sich im Multiple-Choice-Format umsetzen (Downing & Haladyna, 2006). Das geschlossene Antwortformat erlaubt es dabei, eine hohe Auswertungs- und Interpretationsobjektivität zu erzielen. Eine sorgfältige Testkonstruktion vorausgesetzt, können Multiple-Choice-Tests dazu eine hohe Reliabilität und Validität erreichen (Downing & Haladyna, 2006; Haladyna, 2004). Neben Einzeltestungen erlauben es Multiple-Choice-Tests ebenfalls, effizient große Gruppen zu testen. Die Testantworten der Teilnehmer lassen sich anschließend leicht automatisiert auswerten. Typischerweise werden Tests im Multiple-Choice-Format mit Hilfe der *Anzahl-Korrekt-Auswertung* (*number-right scoring*; Frary, 1989) bepunktet. Bei dieser Auswertungsmethode entspricht der Testscore eines Testteilnehmers der Anzahl richtig beantworteter Items. Für ein Item wird dichotom 1 Punkt vergeben, wenn die korrekte Antwortoption gewählt wurde, und 0 Punkte, wenn die Wahl auf eine inkorrekte Antwortoption fällt.

Ein häufig geäußerter Nachteil von Multiple-Choice-Items ist, dass die Auswertung dem Alles-oder-nichts-Prinzip folgt und keine Teilpunkte gewährt werden, wenn die Bepunktung nach der konventionellen Anzahl-Korrekt-Auswertung erfolgt (Frary, 1989). Damit wird die Varianz in den Antworten der Testteilnehmer ignoriert, die sich aus der

Wahl unterschiedlicher Distraktoren ergibt. Dabei gibt es plausible Gründe dafür, warum diese Varianz aufschlussreich für das zu messende Konstrukt sein könnte. So ist nämlich anzunehmen, dass Testteilnehmer, denen es nicht gelingt die richtige Antwort eines Multiple-Choice-Items zu identifizieren, diejenige Antwortoption wählen, von der sie am ehesten glauben, dass sie richtig ist. Dies gilt insbesondere für Multiple-Choice-Tests, dessen Ergebnisse persönliche Konsequenzen für die Testteilnehmer haben, wie z. B. in einer Prüfungssituation. Auch wenn Testteilnehmer die Lösung eines Items nicht kennen, können sie möglicherweise Antwortoptionen durch Teilwissen ausschließen und dadurch zumindest die Wahrscheinlichkeit erhöhen, die richtige Antwort auszuwählen. Für Testteilnehmer, die sich bei einem Item für einen Distraktor entscheiden, ist folglich zu erwarten, dass sich leistungsstärkere Teilnehmer aufgrund ihres Teilwissens in der Wahl des Distraktors von leistungsschwächeren Teilnehmern unterscheiden. Die Wahl eines Distraktors kann damit – abgesehen davon, dass es sich um eine falsche Antwort handelt – zusätzliche Informationen über die Fähigkeit eines Testteilnehmers liefern (Davis, 1959) und damit möglicherweise die Messgenauigkeit eines Tests verbessern. Eine konventionelle Anzahl-Korrekt-Auswertung ist jedoch nicht in der Lage, diese Informationen über das Teilwissen der Testteilnehmer zu berücksichtigen, da die Vergabe von Teilpunkten für Teilwissen bei dieser Auswertungsmethode nicht vorgesehen ist.

Tabelle 1. Vergebene Punktzahl für die Antwortoptionen des Multiple-Choice-Items „Wer war Lehrer von Alexander dem Großen?" unter Anwendung der Anzahl-Korrekt-Auswertung und der Optionsgewichtung.

Antwortoption	**Anzahl-Korrekt**	**Optionsgewichtung**
Aristoteles (korrekt)	1	1.00
Sokrates	0	0.25
Beckenbauer	0	0.00

Tabelle 1 zeigt ein Multiple-Choice-Item zur Geschichte der Antike mit drei Antwortoptionen. Für das Item wird nach der klassischen Anzahl-Korrekt-Auswertung nur für die Wahl der Antwortoption „Aristoteles" 1 Punkt gewährt; für die Wahl einer der beiden

Distraktoren gibt es hingegen 0 Punkte. Dabei unterscheidet sich die Wahl des Distraktors „Sokrates" qualitativ von der Wahl des Distraktors „Beckenbauer". Entscheidet sich ein Testteilnehmer nicht für die korrekte Antwort „Aristoteles", sondern für die Antwortoption „Sokrates", und damit für einen anderen berühmten griechischen Philosophen der Antike, so kann man dem Testteilnehmer zumindest geschichtliches Teilwissen attestieren. Die Wahl der Antwortoption „Beckenbauer" ist hingegen abwegig und ein Hinweis auf ein geringes geschichtliches Wissen über die Antike.

In der Literatur wurden zahlreiche alternative Antwortformate vorgeschlagen, die viele Vorteile des klassischen Multiple-Choice-Formats teilen, jedoch darüber hinaus erlauben, Teilwissen zu erfassen. Das *Antworte-Bis-Es-Stimmt-Testformat* (*Answer-Until-Correct test format*; Gilman & Ferry, 1972) gibt Testteilnehmern bei Multiple-Choice-Items beispielsweise direkt nach jeder Antwort eine Rückmeldung, ob ihre Antwort richtig ist, und gewährt ihnen so viele Antwortversuche, bis sie die richtige Antwort ausgewählt haben. Ist ein Testteilnehmer nicht in der Lage, ein Item im ersten Versuch zu lösen, kann dennoch Teilwissen berücksichtigt werden, wenn die Person insgesamt nur wenige Versuche benötigt, um die richtige Antwort zu wählen.

Eine weitere Alternative zum konventionellen Multiple-Choice-Format, das auch als Einfachauswahlverfahren bezeichnet werden kann, ist das *Mehrfachauswahlverfahren* (*free-choice testing*; Dressel & Schmid, 1953). Bei diesem Verfahren haben die Testteilnehmer die Möglichkeit mehrere Antwortoptionen eines Multiple-Choice-Items auszuwählen, die sie potenziell für die richtige Antwort halten. Je geringer die Anzahl der ausgewählten Antwortoptionen, die benötigt wird, um die Lösung zu markieren, desto höher fällt die vergebene Punktzahl aus. Beim *Eliminations-Testformat* (*elimination testing*; Coombs, 1953) ist die Aufgabe genau umgekehrt: Testteilnehmer sollen alle Antwortoptionen eines Multiple-Choice-Items auswählen, die sie nicht für die richtige Lösung halten. Je mehr Antwortoptionen als falsch zurückgewiesen werden können, desto mehr Punkte werden gewährt. Ein Nachteil dieser alternativen Antwortformate ist, dass sie Testteilnehmern

meist unvertraut sind und zunächst erlernt werden müssen. Außerdem ist unklar, welchen Einfluss Persönlichkeitsmerkmale, wie Risikoneigung oder Testängstlichkeit, auf die psychometrischen Eigenschaften des Tests haben (Frary, 1989). Darüber hinaus nimmt die Bearbeitung pro Item meistens mehr Zeit in Anspruch als bei klassischen Multiple-Choice-Tests (Frary, 1989). Verfahren, wie das Antworte-Bis-Es-Stimmt-Format, benötigen darüber hinaus technische Hilfsmittel, um den Testteilnehmern eine Rückmeldung über die Richtigkeit ihrer Antworten zu geben.

Im folgenden Kapitel wird die Optionsgewichtung (Davis, 1959; Davis & Fifer, 1959) vorgestellt, eine Auswertungsmethode, die es ermöglicht auch unter Verwendung klassischer Multiple-Choice-Items, Teilwissen zu erfassen.

2 Optionsgewichtung

Die Optionsgewichtung ist ein alternatives Auswertungsverfahren für Multiple-Choice-Tests (Davis, 1959; Davis & Fifer, 1959). Anders als bei der häufig eingesetzten Anzahl-Korrekt-Auswertung, bei der dichotom 1 Punkt für die Wahl der richtigen Antwortoption und 0 Punkte für die Wahl eines Distraktors vergeben werden, kann bei der polytomen Optionsgewichtung die Bepunktung der Antwortoptionen frei variieren. Ein Optionsgewicht bestimmt die Punktzahl, die dem Testteilnehmer für die Wahl einer Antwortoption gutgeschrieben wird. Optionsgewichte gewichten damit über die gewährte Punktzahl die Antwortoptionen eines Tests relativ zueinander. Für die Vergabe der Optionsgewichte gilt generell: Je mehr Wissen die Wahl einer Antwortoption widerspiegelt, desto höher das Gewicht und die Punktzahl, die vergeben wird. Der Testscore eines Testteilnehmers berechnet sich schließlich aus der Summe der Optionsgewichte aller im Test gewählten Antwortoptionen. Für das Beispielitem in Tabelle 1 könnte das Optionsgewicht der falschen Antwortoption „Sokrates" beispielsweise 0.25 Punkte betragen. Das Auswertungsschema bildet auf diese Weise Teilwissen, das sich in der Wahl dieses Distraktors manifestiert, über die Vergabe von Teilpunkten ab und kann damit möglicherweise die Messgenauigkeit verbessern. Das Gewähren von Teilpunkten für Antwortoptionen, die Teilwissen widerspiegeln, hat im Gegenzug zur Folge, dass die Wahl abwegiger Antwortoptionen stärker bestraft wird als bei einer klassischen Anzahl-Korrekt-Auswertung.

Die polytome Auswertung der Optionsgewichtung erweitert den Wertebereich der Punkte, die für ein Multiple-Choice-Item vergeben werden können. Dadurch ermöglicht das Verfahren besser zwischen Testteilnehmern zu differenzieren. Ein Multiple-Choice-Test, der beispielsweise aus 20 Items mit je drei Antwortoptionen besteht, erlaubt es nach klassischer Anzahl-Korrekt-Auswertung nur zwischen 21 Wissenszuständen zu unterscheiden. Die Optionsgewichtung vermag hingegen bis zu $3^{20} = 3.486.784.401$ verschiedene

Wissenszustände voneinander zu trennen (Haladyna, 1990). Da sich die Optionsgewichtung insbesondere hinsichtlich der Bepunktung von unvollständigem Wissen von der klassischen Anzahl-Korrekt-Auswertung unterscheidet, verbessert sich die Differenzierung vor allem im niedrigen Fähigkeitsbereich (Haladyna, 1990).

Für die Bestimmung von Optionsgewichten wurden unterschiedliche Verfahren vorgeschlagen, die sich einerseits in die *empirische Optionsgewichtung* (*empirical option weighting*) und andererseits in die *Experten-Optionsgewichtung* (*a priori weighting* oder *expert option weighting*) unterteilen lassen (Stanley & Wang, 1970). Empirische Optionsgewichte werden nach folgendem Prinzip bestimmt: Testteilnehmer, die einen Distraktor wählen, der populär unter leistungsstarken (leistungsschwachen) Testteilnehmern ist, werden mit einer hohen (niedrigen) Punktzahl belohnt (bestraft). Ein geläufiger Ansatz, um ein empirisches Optionsgewicht für eine Antwortoption zu bestimmten, ist die Berechnung der *Option-Gesamt-Korrelation* (*option-total correlation*; Haladyna, 1990). Die Option-Gesamt-Korrelation ist die punkt-biseriale Korrelation zwischen den Optionswahlen – 1, wenn vom Testteilnehmer gewählt, und 0, wenn nicht vom Testteilnehmer gewählt – und dem erzielten Testscore, berechnet nach der Anzahl-Korrekt-Auswertung (Haladyna, 1990). In früheren Studien wurde häufig der hohe Aufwand für die Berechnung von empirischen Optionsgewichten kritisiert (z. B. Haladyna, 1990; Raffeld, 1975; Wang & Stanley, 1970). Heutzutage liegen die Daten für viele Tests jedoch elektronisch vor und die Auswertung erfolgt mit Hilfe eines Computers. Wurde ein Auswertungsverfahren, wie das der empirischen Optionsgewichtung, einmal als Algorithmus implementiert, ergibt sich hinsichtlich des Auswertungsaufwandes kein Unterschied im Vergleich zur konventionellen Anzahl-Korrekt-Auswertung. Bei der empirischen Optionsgewichtung hängt der Testscore eines Testteilnehmers allerdings vom Abschneiden anderer Testteilnehmer ab. Dies ist ein Charakteristikum des Verfahrens, das zu Kritik und im Extremfall sogar zu Akzeptanzproblemen auf Seiten der Testteilnehmer führen könnte.

Die Experten-Optionsgewichtung versucht dieses Problem zu vermeiden, indem die Einschätzungen von einem oder mehreren Experten im Bereich der Testdomäne eingeholt

werden (Downey, 1979; Patnaik & Traub, 1973). Es wird angenommen, dass Experten dazu qualifiziert sind, die Menge an Teilwissen einzuschätzen, die von der Wahl eines Distraktors widergespiegelt wird. Daher sollten Experten in der Lage sein, Optionsgewichte zu vergeben, die Teilwissen angemessen belohnen und Unwissen entsprechend bestrafen. Wenn mehrere Experten ihre Einschätzungen abgeben, werden üblicherweise die Optionsgewichte über alle Experten aggregiert. Testergebnisse, die mit Experten-Optionsgewichten bestimmt wurden, genießen potenziell eine größere Akzeptanz unter Testteilnehmern, da sie im Vergleich zu empirischen Optionsgewichten über eine höhere Augenscheinvalidität und Transparenz verfügen.

In fast allen Studien, die die empirische Optionsgewichtung untersuchten, konnten durch die Gewichtung reliablere Testscores erzielt werden als unter der Verwendung konventioneller Auswertungsmethoden (Claudy, 1978; Cross & Frary, 1978; Cross, Ross & Geller, 1980; Downey, 1979; Echternacht, 1976; Haladyna, 1990; Hendrickson, 1971; Raffeld, 1975; Reilly & Jackson, 1973; Waters, 1976). Einzig Sabers und White (1969) berichteten inkonsistente Ergebnisse. Das Befundmuster von Studien, die die Reliabilität der Experten-Optionsgewichtung analysierten, ist dagegen heterogen. Einige Untersuchungen fanden, dass Experten-Optionsgewichte, im Vergleich zu konventionellen Auswertungsverfahren, die Reliabilität von Multiple-Choice-Tests erhöhten (Downey, 1979; Patnaik & Traub, 1973); andere Studien berichteten jedoch keine Verbesserung (Cross et al., 1980; Echternacht, 1976; Hambleton, Roberts & Traub, 1970) oder kein eindeutiges Ergebnismuster (Kansup & Hakstian, 1975).

Die Befundlage hinsichtlich möglicher Validitätsverbesserungen durch die empirische Optionsgewichtung ist ebenfalls nicht eindeutig. Obwohl die Validität das wichtigere Testgütekriterium ist, wurde sie nicht in allen Studien zusätzlich zur Reliabilität untersucht (Claudy, 1978; Davis & Fifer, 1959; Hendrickson, 1971). Die Studien von Reilly und Jackson (1973) und Downey (1979) beobachteten eine Verschlechterung der Validität durch die empirische Optionsgewichtung im Vergleich zu konventionellen Auswertungsmethoden. Cross und Frary (1978), Echternacht (1976), Raffeld (1975) und Haladyna

(1990) fanden hingegen eine Verbesserung der Validität durch die Verwendung von empirischen Optionsgewichten. Einige Untersuchungen fanden keinen Unterschied hinsichtlich der Validität zwischen empirischer Optionsgewichtung und der Anzahl-Korrekt-Auswertung (Cross et al., 1980; Davis & Fifer, 1959; Sabers & White, 1969; Waters, 1976). Studien zur Experten-Optionsgewichtung fanden meist keine Verbesserung der Validität gegenüber einer klassischen Auswertung (Cross et al., 1980; Echternacht, 1976; Hambleton et al., 1970; Kansup & Hakstian, 1975; Patnaik & Traub, 1973). Allein Downey (1979) berichtete eine Verbesserung der Validität durch Experten-Optionsgewichte.

Nur wenige der oben genannten Studien, die die Reliabilität oder Validität von Optionsgewichtungsverfahren untersuchten, führten Signifikanztests zur zufallskritischen Absicherung der Ergebnisse durch (Cross & Frary, 1978; Davis & Fifer, 1959; Patnaik & Traub, 1973; Waters, 1976). Haladyna (1990) sowie Kansup und Hakstian (1975) testeten zumindest die gefundenen Validitätsunterschiede auf Signifikanz. Alle anderen Studien interpretierten ihre Ergebnisse auf der Grundlage deskriptiver Statistiken. Eine nicht unwahrscheinliche Erklärung für das Versäumnis früherer Studien, die beobachteten Ergebnisse zufallskritisch abzusichern, ist das Fehlen von Signifikanztests zur Prüfung von Unterschieden in abhängigen und unabhängigen Reliabilitätskoeffizienten sowie Korrelationen in den gängigen Statistikpaketen, wie z. B. SPSS oder SAS.

Das unklare Befundmuster der Optionsgewichtungsverfahren im Hinblick auf die Validität könnte möglicherweise auch auf die zur Validierung verwendete Methodik zurückgehen. Ausnahmslos alle Studien haben zur Bestimmung konvergenter oder prädiktiver Validitäten ein korrelatives Studiendesign verwendet. Der wahre Wissensstand der Testteilnehmer war in diesen Studien unbekannt und konnte nicht als Außenkriterium zur Validierung herangezogen werden. Stattdessen wurde die Validität über die Korrelation der Testscores mit den Ergebnissen anderer Testverfahren oder Expertenurteilen ermittelt (z. B. Davis & Fifer, 1959; Downey, 1979). Diese Außenkriterien stellen allerdings fehlerbehaftete Schätzungen des wahren Wissensstandes eines Testteilnehmers dar. Ob sie eine

hinreichend reliable Messung zur Bestimmung der Validität der Optionsgewichtung zulassen ist fraglich. Eine hohe Korrelation zwischen zwei Tests muss nicht notwendigerweise allein auf der Messung desselben Konstrukts beruhen, sondern kann beispielsweise auch auf die gemeinsame Methodenvarianz der Tests zurückgehen (Podsakoff, MacKenzie, Lee & Podsakoff, 2003). Die Existenz dieser Alternativerklärung für eine hohe externe Validität ist problematisch, weil viele Studien zur Optionsgewichtung das gleiche Testformat sowohl für die Prädiktoren als auch für das Kriterium verwendeten und deshalb eine alternative Interpretation zulassen (z. B. Cross & Frary, 1978; Echternacht, 1976; Haladyna, 1990).

Zur Untersuchung, ob empirische Optionsgewichte oder Optionsgewichte, die auf der Grundlage von Expertenurteilen bestimmt wurden, besser geeignet sind, um gegenüber der Anzahl-Korrekt-Auswertung die Testgüte eines Multiple-Choice-Tests zu verbessern, ist es erforderlich, alle drei Auswertungsverfahren unter identischen Bedingungen zu vergleichen. Zu diesem Zweck müssen die konkurrierenden Verfahren auf den gleichen Satz an Antworten angewendet werden, die eine Gruppe von Personen zur Beantwortung eines Multiple-Choice-Tests gegeben hat. In der bestehenden Literatur erfüllen dieses Kriterium nur die Studien von Echternacht (1976), Downey (1979) und Cross et al. (1980). Bedauerlicherweise wurden in keiner dieser drei Studien die beobachteten Unterschiede zwischen den Auswertungsverfahren hinsichtlich der Testgüte auf statistische Signifikanz geprüft.

In der ersten der drei Studien verwendete Echternacht (1976) einen gezeiteten Test, der aus 30 Items bestand. Da nur 17% der Teilnehmer es schafften, alle Items zu beantworten, wurden die letzten 12 Items des Tests entfernt, um eine Beendigungsrate von über 90% zu erreichen. Die empirischen Optionsgewichte wurden nicht auf der Basis der Testscores einer Anzahl-Korrekt-Auswertung bestimmt, sondern auf der Grundlage von Testscores, für die eine Ratekorrektur (*formula scoring*; Lord, 1975) vorgenommen wurde. Um für Rateverhalten zu korrigieren, wurden von den Testscores einer Anzahl-

Korrekt-Auswertung jeweils die aufgrund einer zufälligen Beantwortung erwartete Punktzahl abgezogen. Die auf diese Weise bestimmten empirischen Optionsgewichte erreichten eine höhere Reliabilität und eine höhere konvergente Validität als eine Anzahl-Korrekt-Auswertung mit oder ohne Ratekorrektur. Darüber hinaus variierten in der Studie von Echternacht (1976) die Experten-Optionsgewichte nicht zwischen den Items. Für jedes Item wurden 6 Punkte für die korrekte Antwort vergeben, 1 Punkt für den Distraktor, der zumindest auf Teilwissen schließen ließ, und -4 Punkte für den Distraktor, der keinen Aufschluss über Teilwissen erlaubte. Die nach diesem Schema festgelegten Experten-Optionsgewichte verbesserten weder die Reliabilität noch die Validität des untersuchten Multiple-Choice-Tests. Eine mögliche Erklärung für diese Nullbefunde könnte die Simplizität der eingesetzten Experten-Optionsgewichte sein. Da diese für alle Items identisch waren, konnte das Teilwissen der Testteilnehmer möglicherweise nicht adäquat abgebildet werden.

In der zweiten Studie von Downey (1979) verbesserte die empirische Optionsgewichtung zwar die Reliabilität, jedoch weder die prädiktive noch die konkurrente Validität im Vergleich zur Anzahl-Korrekt-Auswertung. Das verwendete Testmaterial liefert allerdings eine mögliche Erklärung für das Ausbleiben einer Validitätsverbesserung. Für 9 der insgesamt 30 Items war das Optionsgewicht einer der Distraktoren höher als das Optionsgewicht der korrekten Antwort. Für fast ein Drittel der Items wählten leistungsstarke Testteilnehmer demnach eine Antwortoption, die nicht die Lösung war. Die Experten-Optionsgewichte wurden ermittelt, indem 7 Experten gebeten wurden, für jede Antwortoption die darin reflektierte Menge an Teilwissen auf einer 7-Punkte-Skala einzuschätzen. Optionsgewichte, die über alle Experten gemittelt wurden, waren in der Lage, die Testreliabilität gegenüber einer Anzahl-Korrekt-Auswertung zu verbessern. Die Verbesserung fiel jedoch deskriptiv geringer aus als bei der empirischen Optionsgewichtung. Die Studie von Downey (1979) ist die einzige Untersuchung, die auch eine moderate Verbesserung der prädiktiven Validität durch die Experten-Optionsgewichtung feststellte. Die Validitätsverbesserung fand sich jedoch nur in einer Hälfte der Stichprobe. In keiner der beiden

Stichprobenhälften konnte die Experten-Optionsgewichtung die konkurrente Validität erhöhen.

In der dritten Studie von Cross et al. (1980) verbesserte die empirische Optionsgewichtung die Reliabilität, aber nicht die Validität verschiedener Multiple-Choice-Tests. Hinsichtlich der Experten-Optionsgewichtung vermochten Gewichte, die durch den jeweiligen Testentwickler festgelegt wurden, weder die Reliabilität noch die Validität der Tests zu erhöhen. Bei vielen Items wiesen die Testentwickler, die hier die Rolle der Experten einnahmen, jedoch allen Distraktoren ein Optionsgewicht von null zu. Ein solches Gewichtungsschema kann unmöglich zu einer Verbesserung gegenüber einer Anzahl-Korrekt-Auswertung führen, weil es äquivalent zu einer dichotomen Auswertung ist. Möglicherweise waren die Einschätzungen der Experten zu ungenau oder das Testmaterial war für eine Optionsgewichtung ungeeignet, da die Distraktoren kein Teilwissen widerspiegelten.

Zusammenfassend lässt sich sagen, dass kein eindeutiges Ergebnismuster aus früheren Vergleichen zwischen der empirischen Optionsgewichtung und der Experten-Optionsgewichtung hervorging. Daher war es nicht möglich, eine klare Empfehlung zu formulieren, ob Optionsgewichte empirisch oder auf der Grundlage von Expertenurteilen bestimmt werden sollten.

Generell sollte sich eine Verbesserung der Testgüte durch die Optionsgewichtung gegenüber einer konventionellen Auswertung am deutlichsten bei schwierigen Items zeigen (Haladyna, 1990). Da Testteilnehmer bei leichten Items meist die korrekte Antwort wählen, können gewichtete Distraktoren hier keine zusätzlichen Informationen liefern. Dies könnte auch eine Erklärung dafür sein, warum einige der früheren Studien bei leichten Tests keine Verbesserung durch Optionsgewichtungsverfahren im Hinblick auf die Validität fanden. In den Studien von Cross et al. (1980) und Downey (1979) beantworteten die Testteilnehmer beispielsweise 68% bzw. 67% der Items korrekt.

Gemischte Ergebnisse hinsichtlich der Validität der empirischen Optionsgewichtung könnten auch das Resultat einer verzerrten Testkonstruktion sein. Beim Vergleich dichotomer und polytomer Auswertungsverfahren für Multiple-Choice-Tests ist es wichtig

zu berücksichtigen, ob der für den Vergleich verwendete Test für ein bestimmtes Auswertungsverfahren konstruiert wurde. Die meisten Tests, die in früheren Studien eingesetzt wurden, waren ursprünglich für die Anzahl-Korrekt-Auswertung entwickelt worden (z. B. Kansup & Hakstian, 1975; Patnaik & Traub, 1973; Waters, 1976). Die empirische Optionsgewichtung ist jedoch am effektivsten, wenn die Antwortoptionen der Multiple-Choice-Items jeweils für Testteilnehmer mit unterschiedlichen Fähigkeitsniveaus attraktiv sind (Haladyna, 1990; Raffeld, 1975). Derartige Itemeigenschaften werden bei der Konstruktion eines für die Anzahl-Korrekt-Auswertung entwickelten Multiple-Choice-Tests üblicherweise nicht berücksichtigt (Haladyna, 2004). Nur zwei Studien haben bisher die empirische Optionsgewichtung auf einen Test angewendet, der im Hinblick auf eine polytome Auswertung konstruiert wurde (Davis & Fifer, 1959; Echternacht, 1976). Davis und Fifer (1959) fanden, dass auf der empirischen Optionsgewichtung basierende Testscores höher mit einer parallelen Testversion korrelierten als konventionell berechnete Testscores, ohne dabei die konkurrente Validität zu verringern. Die Autoren setzten in ihrer Untersuchung jedoch eine ungewöhnliche Methode zur Berechnung der empirischen Optionsgewichte ein. Sie verwendeten auf Experten-Optionsgewichten basierende Testscores als Grundlage für die Berechnung der empirischen Optionsgewichte, anstatt – wie üblich – empirische Optionsgewichte auf der Basis einer Anzahl-Korrekt-Auswertung zu berechnen. Davis und Fifer (1959) nutzten außerdem eine vereinfachte Berechnungsmethode, die nur die Daten der jeweils leistungsstärksten und -schwächsten 27% der Teilnehmer berücksichtigte. Die Optionsgewichte wurden darüber hinaus nach der Berechnung so angepasst, dass die korrekte Antwort stets ein höheres Gewicht als die Distraktoren hatte. In der Studie von Echternacht (1976) verwendeten die Autoren ebenfalls einen Test, der für eine polytome Auswertung entwickelt wurde. Deskriptiv fand Echternacht (1976) sogar eine Verbesserung der Reliabilität und Validität durch die empirische Optionsgewichtung im Vergleich zu einer konventionellen Auswertung. Da jedoch viele der früheren Studien überwiegend Tests verwendeten, die für eine Anzahl-Korrekt-Auswertung konstruiert wurden, können

die Ergebnisse dieser Studien hinsichtlich des Potenzials der Optionsgewichtungsverfahren nur eingeschränkt interpretiert werden.

Im folgenden Kapitel wird Artikel 1 zusammengefasst, der eine experimentelle Studie zur Validierung der empirischen Optionsgewichtung schildert (Studie 1). Darauf folgt in Kapitel 4 eine Vorstellung des zweiten Artikels, in dem eine Studie zum Vergleich der empirischen Optionsgewichtung und der Experten-Optionsgewichtung berichtet wird (Studie 2). In Kapitel 5 wird der dritte Artikel zusammengefasst, in dem eine Statistiksoftware zum Vergleich von Korrelationen vorgestellt wird, die im Rahmen der vorliegenden Arbeit entwickelt und angewendet wurde. Eine detaillierte Darstellung aller Arbeiten ist im Anhang zu finden.

3 Artikel 1: Eine experimentelle Untersuchung zur Validität der empirischen Optionsgewichtung

Zur Beantwortung der Frage, ob die empirische Optionsgewichtung die Validität von Multiple-Choice-Tests verbessert, kamen in früheren Studien ausschließlich korrelative Studiendesigns zum Einsatz. In Studie 1 der vorliegenden Arbeit wurde zum ersten Mal ein experimentelles Verfahren zur Bestimmung der Validität eingesetzt (vgl. Poizner, Nicewander & Gettys, 1978). Dazu wurden die Wissensstände der Studienteilnehmer experimentell manipuliert, indem ihnen im Hinblick auf einen späteren Wissenstest unterschiedlich informative Texte zum Lesen vorgegeben wurden. Anders als in korrelativen Studien waren somit die Wissensstände, über die die Teilnehmer höchstens verfügen konnten, bekannt und konnten als Außenkriterium zur Bestimmung der Validität herangezogen werden. Der Test war außerdem so konstruiert, dass die Antwortoptionen der Multiple-Choice-Items für Teilnehmer mit verschiedenen Wissensständen unterschiedlich attraktiv waren. Im Gegensatz zur Mehrzahl früherer Studien hatte das Testmaterial damit günstige Eigenschaften für die Anwendung der empirischen Optionsgewichtung. Die Testantworten wurden zweifach ausgewertet: zum einen durch die Anzahl-Korrekt-Auswertung und zum anderen unter Verwendung der empirischen Optionsgewichtung. Es wurde erwartet, dass die Testscores der empirischen Optionsgewichtung die Wissensstände der Teilnehmer reliabler und valider erfassen würden als die Testscores der Anzahl-Korrekt-Auswertung. Als Maß für die Validität wurde die Präzision bestimmt, mit der die beiden Auswertungsverfahren in der Lage waren, die experimentell manipulierten – und daher bekannten – Wissensstände der Testteilnehmer abzubilden.

Das Experiment hatte ein 3×2-faktorielles Design mit dem Zwischensubjektfaktor *Wissen* (kein Wissen, Teilwissen und vollständiges Wissen) und dem Innersubjektfaktor *Auswertungsmethode* (Anzahl-Korrekt-Auswertung vs. empirische Optionsgewichtung). Die Untersuchung wurde in Form einer Onlinestudie durchgeführt, zu der insgesamt 567 Teilnehmer (55% weiblich) per E-Mail rekrutiert wurden. Die Teilnehmer wurden

zunächst randomisiert einer der drei Wissensbedingungen zugewiesen. Um verschiedene Wissensstände zu induzieren, lasen die Teilnehmer kurze Informationstexte, die entweder keine Informationen, einen Teil der Informationen oder alle Informationen enthielten, die benötigt wurden, um die Fragen in einem anschließenden Wissenstest richtig beantworten zu können. Die Informationstexte beschrieben seltene und obskure Details über die Geschichte und Kultur eines Stammes südamerikanischer Ureinwohner – der Arawaken. In einer Vorstudie im Rahmen einer Psychologievorlesung waren die Arawaken einer Stichprobe von 120 Studenten unbekannt. Durch die Wahl dieses Themas wurde sichergestellt, dass die Teilnehmer über kein Vorwissen verfügten, das mit der experimentellen Wissensmanipulation interferieren könnte.

Als abhängige Variablen wurden die Reliabilität und Validität des Wissenstests ermittelt. Als Maß für die Reliabilität wurde die interne Konsistenz über Cronbachs α (Cronbach, 1951) bestimmt. Zur Bestimmung der Validität wurden zwei unterschiedliche Maße berechnet: Auf der einen Seite wurde eine Varianzanalyse mit den Testscores der Teilnehmer als abhängige Variable und dem Wissensstand der Teilnehmer als unabhängige Variable durchgeführt, um zu prüfen, unter welchem Auswertungsverfahren der experimentell manipulierte Wissensstand der Teilnehmer mehr Varianz in den Testscores aufklärte. Auf der anderen Seite wurde für beide Auswertungsverfahren die Trefferrate bestimmt, mit der eine eindimensionale lineare Diskriminanzanalyse die Teilnehmer anhand der Testscores korrekt ihren bekannten Wissensständen zuordnen konnte.

Der Wissenstest, den die Teilnehmer nach dem Lesen des Informationstextes absolvierten, bestand aus 10 Multiple-Choice-Items mit je drei Antwortoptionen. Bei sieben Items handelte es sich um Experimentalitems, die in die anschließende Datenauswertung eingingen. Die Items waren so konstruiert, dass die Teilnehmer in der Gruppe ohne Wissen nicht in der Lage waren, aufgrund der Informationen, die sie im Text erhielten, Distraktoren auszuschließen. Folglich wurde erwartet, dass Teilnehmer in dieser Bedingung gezwungen sein würden, Antwortoptionen zufällig auszuwählen. Die Lösungswahrscheinlichkeit eines Items sollte für diese Teilnehmergruppe auf dem Rateniveau von 33% liegen. Die

Gruppe von Teilnehmern, die nur über Teilwissen verfügte, las einen Informationstext, der ihnen erlaubte einen der zwei Distraktoren bei jedem Item auszuschließen. Es wurde erwartet, dass die Lösungswahrscheinlichkeit für diese Gruppe, unter Annahme eines perfekten Gedächtnisses, 50% betragen würde. Teilnehmer, die sich in der Gruppe mit vollständigem Wissen befanden, erhielten einen detaillierten Informationstext, der es ihnen erlaubte, alle Items korrekt zu beantworten und so den maximalen Testscore zu erreichen – ein perfektes Gedächtnis vorausgesetzt. Tabelle 2 veranschaulicht anhand eines Beispielitems, wie die Informationstexte, die sich zwischen den Wissensgruppen hinsichtlich ihres Informationsgehaltes unterschieden, zu verschiedenen Wissensständen führten. Dadurch schnitten die Teilnehmer, abhängig von ihrem Wissensstand, entweder erfolgreich, mäßig erfolgreich oder nicht erfolgreich im Wissenstest ab. Das Beispielitem wurde aufgrund seiner Kürze ausgewählt. Die Antwortoptionen anderer Items bestanden nicht aus additiven Wortlisten, sondern aus längeren Sätzen. Alle Items hatten jedoch gemeinsam, dass sie eine Hierarchie etablierten, die die Distraktoren für Teilnehmer mit unterschiedlichen Wissensstufen attraktiv machte. Dies wurde durch die Erstellung von Distraktoren mit einer unterschiedlichen Anzahl von Propositionen erreicht, welche entweder in den Informationstexten, die den Teilnehmern zur Induktion der drei Wissensstufen vorgegeben wurden, enthalten waren oder nicht. Neben den sieben Experimentalitems bestand der Wissenstest außerdem aus drei Füllitems, für die die Teilnehmer aller Wissensbedingungen ausreichend Informationen erhielten, um die korrekte Antwort zu identifizieren. Diese Items gingen nicht in die Auswertung ein, sondern sollten den Teilnehmern in der Gruppe, die über kein Wissen verfügte, die Frustration ersparen, keines der Items beantworten zu können.

Tabelle 2. Beispielitem („Welche Nahrungsmittel waren den Arawaken bekannt?") des Arawaken-Wissenstests mit den Punkten, die nach der Anzahl-Korrekt-Auswertung und der empirischen Optionsgewichtung für jede der Antwortoptionen in den Stichproben 1 und 2 gewährt wurden, sowie das Wissen, das benötigt wurde, um bestimmte Antwortoptionen auszuschließen.

		Empirische Optionsgewichte		
		Stichproben		
Antwortoptionen	**Anzahl-Korrekt**	**1**	**2**	**Für Ausschluss benötigtes Wissen**
Kartoffeln, Süßkartoffeln, Bohnen	0	−.38	−.35	Teilwissen oder vollständiges Wissen
Süßkartoffeln, Tomaten, Kakao	0	−.22	−.18	Vollständiges Wissen
Süßkartoffeln, Bohnen, Erdnüsse	1	.50	.42	(Richtige Antwort, kein Ausschluss möglich)

Anmerkungen. Die Teilnehmer in der Gruppe ohne Wissen erhielten keine Informationen über die Nahrungsmittel, die den Arawaken bekannt waren. Teilnehmer in der Gruppe mit Teilwissen bekamen die Information: „Sie kannten zwar die sogenannten Süßkartoffeln und aßen sie häufig; die heute verbreiteten Kartoffeln waren ihnen jedoch unbekannt." Zusätzlich zu dieser Information erfuhr die Gruppe mit vollständigem Wissen: „Außerdem ernährten sie sich von Bohnen und Erdnüssen. Die bei anderen Stämmen bekannten Tomaten und die in einigen Regionen Südamerikas verbreitete Kakaopflanze kannten sie nicht."

Nachdem alle Teilnehmer den Test absolviert hatten, wurden die Antworten zweifach ausgewertet: einerseits durch die Anzahl-Korrekt-Auswertung und anderseits unter Verwendung der empirischen Optionsgewichtung. Zur Bestimmung der empirischen Optionsgewichte wurde die part-whole-korrigierte Option-Gesamt-Korrelation für alle Antwortoptionen aller Items im Test berechnet (Haladyna, 1990). Um *capitalization on chance* zu vermeiden und eine doppelte Kreuzvalidierung der Ergebnisse vorzunehmen (Stanley & Wang, 1970), wurde die Stichprobe zufällig in zwei Substichproben mit 284 und 283 Teilnehmern geteilt. Die an der ersten Hälfte der Teilnehmer (Stichprobe 1) bestimmten empirischen Optionsgewichte wurden zur Auswertung der Antworten der zweiten Hälfte (Stichprobe 2) verwendet und umgekehrt. Um sicherzustellen, dass beide Stichproben über die gleiche Zahl an Teilnehmern aus jeder der drei Wissensgruppen verfügten, wurde die Stichprobenteilung separat für jede Wissensgruppe durchgeführt. Tabelle 2 zeigt die empirischen Optionsgewichte, die für die Antwortoptionen des Beispielitems ermittelt wurden.

Die sieben Experimentalitems hatten eine mittlere Schwierigkeit von .52 bzw. .53 in den Stichproben 1 und 2. Wie erwartet, war bei allen Items das Optionsgewicht für die korrekte Antwort am höchsten (im Mittel, $M = .48$). Der Distraktor, der mit vollständigem Wissen, aber nicht mit Teilwissen, ausgeschlossen werden konnte, erhielt das zweithöchste Optionsgewicht ($M = -.18$). Der Distraktor, der von Teilnehmern mit Teilwissen oder vollständigem Wissen ausgeschlossen werden konnte, hatte das niedrigste Optionsgewicht ($M = -.38$), mit der Ausnahme eines Items, für das in Stichprobe 2 die Optionsgewichte der beiden Distraktoren in umgekehrter Reihenfolge waren ($M = -.25$ vs. $M = -.21$).

Die von den Teilnehmern erreichten Testscores sind in Tabelle 3 für die Anzahl-Korrekt-Auswertung und die empirische Optionsgewichtung aufgeführt. Wie erwartet, beantworteten Teilnehmer mit Teilwissen mehr Experimentalitems korrekt als Teilnehmer ohne Wissen, sowohl in Stichprobe 1 ($M = 3.06$ vs. $M = 1.84$; $t(192) = 7.65$, $p < .001$, $d = 1.10$) als auch in Stichprobe 2 ($M = 3.17$ vs. $M = 1.84$; $t(191) = 8.04$, $p < .001$,

$d = 1.16$). Teilnehmer mit vollständigem Wissen beantworteten wiederum mehr Items richtig als Teilnehmer mit Teilwissen, sowohl in Stichprobe 1 ($M = 6.30$ vs. $M = 3.06$; $t(184) = 21.92$, $p < .001$, $d = 3.22$) als auch in Stichprobe 2 ($M = 6.26$ vs. $M = 3.17$; $t(183) = 18.95$, $p < .001$, $d = 2.79$). Diese Unterschiede in der durchschnittlichen Anzahl korrekt beantworteter Items belegen, dass die Wissensmanipulation erfolgreich war. Die Teilnehmer der drei Wissensbedingungen schnitten bei der Beantwortung der drei Füllitems in Stichprobe 1 ($F(2, 281) = 0.41$, $p = .663$, $\eta_g^2 < .01$) und Stichprobe 2 ($F(2, 280) = 0.84$, $p = .435$, $\eta_g^2 < .01$) gleich gut ab, wie aus einfaktoriellen Varianzanalysen hervorging.

In beiden Stichproben wurde Cronbachs α für die Testscores der Anzahl-Korrekt-Auswertung und die Testscores der empirischen Optionsgewichtung berechnet und verglichen. Dazu wurde der Signifikanztest von Feldt, Woodruff und Salih (1987) für abhängige Cronbach-α-Koeffizienten benutzt, der im R-Paket *cocron* (Version 1.0-0; das Statistikpaket wurde im Rahmen dieser Arbeit entwickelt; Diedenhofen, 2013) implementiert ist. Die empirische Optionsgewichtung erreichte eine signifikant höhere Reliabilität als die Anzahl-Korrekt-Auswertung, sowohl in Stichprobe 1 ($\alpha = .78$ vs. $\alpha = .76$; $\chi^2(1) = 28.92$, $p < .001$) als auch in Stichprobe 2 ($\alpha = .79$ vs. $\alpha = .77$; $\chi^2(1) = 38.25$, $p < .001$). Außerdem wurde die h-Statistik (Haladyna, 1990, p. 236) ermittelt, die angibt, um welchen Faktor ein Test bei einer Anzahl-Korrekt-Auswertung verlängert werden müsste, um dieselbe interne Konsistenz zu erreichen, wie eine Auswertung mit der empirischen Optionsgewichtung (Tabelle 3).

Tabelle 3. Mittelwert (M), Standardabweichung (SD), Minimum (min), Maximum (max), Reliabilität und Validität der Testscores, die von den Teilnehmern in beiden Stichproben unter der Anzahl-Korrekt-Auswertung und der empirischen Optionsgewichtung erzielt wurden.

Stichprobe	**Auswertungsverfahren**	***M***	***SD***	**min**	**max**	$\boldsymbol{\alpha}$	***h***	$\boldsymbol{\eta_g^2}$	**Trefferrate**
1	Anzahl-Korrekt	3.67	2.13	0.00	7.00	.76	–	.77	78.52%
	Empirische Optionsgewichtung	0.75	1.78	−2.70	3.40	.78	1.15	.82	87.32%
2	Anzahl-Korrekt	3.69	2.15	0.00	7.00	.77	–	.73	74.56%
	Empirische Optionsgewichtung	0.75	1.76	−2.56	3.34	.79	1.15	.77	84.10%

Anmerkungen. Zur Bestimmung der Reliabilität wurde Cronbachs α ermittelt. Die h-Statistik gibt den Faktor an, um den ein Test bei einer Anzahl-Korrekt-Auswertung verlängert werden müsste, um dieselbe interne Konsistenz zu erzielen, wie die Auswertung des Tests mit der empirischen Optionsgewichtung. Als Maß für die Validität gibt das generalisierte Eta-Quadrat (η_g^2) den Anteil der Varianz an, der in einer Varianzanalyse mit den Testscores als abhängige Variable und den Wissensbedingungen als unabhängige Variable aufgeklärt werden konnte. Als ein zweites Maß für die Validität wurde die Trefferrate der Teilnehmer berechnet, die in einer Rückklassifikation mit Hilfe einer Diskriminanzanalyse auf der Basis der Testscores ihrer Wissensbedingung korrekt zugeordnet werden konnten.

Um zu bestimmten, welches der zwei Auswertungsmethoden die Wissensstände der Testteilnehmer besser abbildet, wurden für beide Verfahren Varianzanalysen berechnet, mit dem Testscore als abhängige Variable und der Wissensbedingung als unabhängige Variable. Für Stichprobe 1 klärten die mit empirischer Optionsgewichtung berechneten Testscores deskriptiv mehr Varianz im Kriterium auf ($F(2, 281) = 646.15$, $p < .001$, $\eta_g^2 = .82$) als Testscores der Anzahl-Korrekt-Auswertung ($F(2, 281) = 475.20$, $p < .001$, $\eta_g^2 = .77$). Das gleiche Ergebnismuster fand sich auch in Stichprobe 2. Der Anteil der aufgeklärten Varianz war für die empirische Optionsgewichtung deskriptiv höher ($F(2, 280) = 474.46$, $p < .001$, $\eta_g^2 = .77$) als für die konventionelle Anzahl-Korrekt-Auswertung ($F(2, 280) = 384.33$, $p < .001$, $\eta_g^2 = .73$). Darüber hinaus wurde mit Hilfe einer eindimensionalen linearen Diskriminanzanalyse auf der Grundlage der Testscores beider Auswertungsverfahren eine Rückklassifikation der Teilnehmer zu ihren Wissensbedingungen vorgenommen. Tabelle 4 zeigt die Anzahl richtiger und falscher Rückklassifikationen separat für die beiden Auswertungsverfahren in den zwei Stichproben. Liddell-Tests für abhängige Proportionen (Liddell, 1983) ergaben, dass die Trefferrate bei der Rückklassifikation für die empirische Optionsgewichtung höher war als für die Anzahl-Korrekt-Auswertung, sowohl in Stichprobe 1 (87.32% vs. 78.52%; $F(4, 52) = 13.00$, $p < .001$) als auch in Stichprobe 2 (84.10% vs. 74.56%; $F(6, 58) = 9.67$, $p < .001$).

Die Ergebnisse zeigen, dass die empirische Optionsgewichtung die Reliabilität des Wissenstests gegenüber der Anzahl-Korrekt-Auswertung verbesserte. Die h-Statistik veranschaulicht, dass die Verbesserung der internen Konsistenz durch die empirische Optionsgewichtung einer Testverlängerung eines konventionell ausgewerteten Tests von 15% entsprach. Darüber hinaus konnten zwei unabhängige Maße belegen, dass die empirische Optionsgewichtung die Validität des Multiple-Choice-Wissenstests verbesserte. Zum einen ergaben Varianzanalysen mit dem Testscore als abhängige Variable und den Wissensbedingungen als unabhängige Variable, dass die Wissensbedingung mehr Varianz in den Testscores aufklärte, wenn diese durch die empirische Optionsgewichtung berechnet wurden. Zum anderen konnten durch eine Diskriminanzanalyse mehr Teilnehmer anhand

der Testscores korrekt ihrer Wissensbedingung zugeordnet werden, wenn die empirische Optionsgewichtung verwendet wurde. Damit belegen beide Validitätsmaße, dass die Testscores das Wissen der Teilnehmer genauer widerspiegelten, wenn diese über die empirische Optionsgewichtung berechnet wurden.

Tabelle 4. Anzahl der richtigen und falschen Rückklassifikationen der Teilnehmer zu ihren bekannten Wissensbedingungen auf der Grundlage von Diskriminanzanalysen über die Testscores der Anzahl-Korrekt-Auswertung und der empirischen Optionsgewichtung.

Stichprobe			**Empirische Optionsgewichtung**		
			Richtig	**Falsch**	**∑**
1	**Anzahl-Korrekt**	**Richtig**	222 (78.17%)	1 (0.35%)	223 (78.52%)
		Falsch	26 (9.15%)	35 (12.32%)	61 (21.48%)
		∑	248 (87.32%)	36 (12.68%)	284 (100.00%)
2	**Anzahl-Korrekt**	**Richtig**	209 (73.85%)	2 (0.71%)	211 (74.56%)
		Falsch	29 (10.25%)	43 (15.19%)	72 (25.44%)
		∑	238 (84.10%)	45 (15.90%)	283 (100.00%)

4 Artikel 2: Eine Substichproben-Validierung zum Vergleich der empirischen Optionsgewichtung und der Experten-Optionsgewichtung

In der zweiten Studie wurde untersucht, welche Methode zur Bestimmung von Optionsgewichten geeigneter ist, um gegenüber einer Anzahl-Korrekt-Auswertung die Testgüte eines Multiple-Choice-Tests zu verbessern: eine empirische Ermittlung der Optionsgewichte oder eine Methode, die auf den Einschätzungen von Experten beruht. Da sich die experimentelle Induktion von Wissen, aufgrund der artifiziell erzeugten Wissensstände, nicht mit der Experten-Optionsgewichtung kombinieren ließ, wurde für die Untersuchung ein korrelatives Studiendesign gewählt.

In früheren Studien konnte die Anwendung der Experten-Optionsgewichtung häufig nicht überzeugen. Möglicherweise könnte das schlechte Abschneiden mit der gewählten Methode zur Bestimmung der Experten-Optionsgewichte im Zusammenhang stehen. Wenn Experten-Optionsgewichte über die eingeschätzte Rangordnung der Antwortoptionen im Hinblick auf das widergespiegelte Wissen bestimmt wurden, konnte keine Verbesserung der Validität festgestellt werden (Hambleton et al., 1970; Patnaik & Traub, 1973). Hambleton et al. (1970) fanden bei Anwendung dieser Methode nicht einmal eine Verbesserung der Reliabilität. Um mehr Informationen pro Experte mit einer weniger problematischen Prozedur zu gewinnen, wurden in Studie 2 Verhältnisskalen zur Bestimmung der Experten-Optionsgewichte genutzt. Die verwendete Methode glich damit der Methode in der einzigen Studie, die eine Verbesserung sowohl der Reliabilität als auch der Validität durch die Experten-Optionsgewichtung fand (Downey, 1979). Mit Ausnahme von Echternacht (1976), wurden in allen früheren Studien, in denen mehr als ein Experte involviert war, die Gewichte über alle Experten hinweg aggregiert. Es ist jedoch denkbar, dass sich Experten in der Güte der von ihnen vergebenen Optionsgewichte unterscheiden

und dass einige Experten geeignetere Optionsgewichte wählen als andere. Individuelle Unterschiede zwischen den Experten werden jedoch möglicherweise durch die vorgenommene Aggregierung nivelliert. Um eine umfassende Evaluation der Experten-Optionsgewichtung vorzunehmen, wurden im Gegensatz zu vorherigen Studien daher auch Analysen separat für jeden einzelnen Experten durchgeführt, neben einer Analyse in der die Gewichte über alle Experten hinweg aggregierten wurden.

Bei der Untersuchung der Reliabilität und Validität von Testscores, die über die empirische Optionsgewichtung ermittelt wurden, ist eine Kreuzvalidierung zwingend erforderlich (Stanley & Wang, 1970). Wenn die Berechnung der empirischen Optionsgewichte und die Anwendung der Gewichte zur Auswertung von Testantworten auf derselben Stichprobe beruhen, kann auf der einen Seite *overfitting* und *capitalization on chance* zu einer Überschätzung der Effektivität der empirischen Optionsgewichtung führen. Auf der anderen Seite kann die Verwendung von unabhängigen Stichproben bei der Berechnung der empirischen Optionsgewichte an einer Stichprobe und der Anwendung der Gewichte an einer anderen Stichprobe zu einer Unterschätzung der Nützlichkeit der empirischen Optionsgewichtung führen, weil die Gewichte zwangsläufig auch zu einem Teil Stichprobenfehler abbilden. Mit Ausnahme von Waters (1976) und Cross et al. (1980) haben alle vorangegangenen Studien die erhobene Stichprobe für eine Kreuzvalidierung randomisiert in zwei Hälften geteilt, um die Antworten der einen Hälfte mit empirischen Optionsgewichten auszuwerten, die aus den Antworten der anderen Hälfte berechnet wurden. Einige der Studien nahmen dabei auch eine doppelte Kreuzvalidierung vor (z. B. Cross & Frary, 1978). Alle Studien, die eine Kreuzvalidierung durchführten, berichteten jedoch nur die Ergebnisse einer der vielen möglichen Stichprobenteilungen, die vorgenommen werden konnten. Dabei hat keine der früheren Studien untersucht, ob die Wahl der Stichprobenteilung einen Einfluss auf das Ergebnis hatte. Wenn zufällige Stichprobenteilungen wiederholt durchgeführt werden, können sich die Ergebnisse jedoch von einer Teilung zur anderen stark unterscheiden. Weil die Wahl der zufälligen Stichprobenteilung, die berichtet wird, im Ermessen des Forschers liegt, muss sie als *Forscherfreiheitsgrad* (*researcher*

degree of freedom; Simmons, Nelson & Simonsohn, 2011) angesehen werden. Jegliche Flexibilität in der Datenauswertung und der Berichtlegung erhöht potenziell die Rate an falsch-positiven Ergebnissen. Um diesem Problem entgegenzutreten, mahnten Simmons et al. (2011) an, dass Forscher stets darlegen müssen, dass ihre Ergebnisse nicht von willkürlichen Entscheidungen bei der Auswertung abhängen, da die Erzielung jedes beliebigen Ergebnisses sonst in ihrem eigenen Ermessen liegt. Aus diesem Grund sollte die Stichprobenteilung für eine aussagekräftige und zuverlässige Einschätzung der psychometrischen Eigenschaften unterschiedlicher Auswertungsverfahren als potenzieller Moderator der Ergebnisse berücksichtigt werden.

In Studie 2 absolvierten insgesamt 675 Teilnehmer (57% weiblich) ein Online-Experiment, zum dem per E-Mail eingeladen wurde. Zu Beginn des Experimentes beantworteten die Teilnehmer zunächst eine Skala mit 11 Items zur Erfassung ihrer Fußballexpertise. Anschließend absolvierten alle Teilnehmer einen Fußballwissenstest, der aus 27 Multiple-Choice-Items mit jeweils drei Antwortoptionen bestand, und der Fakten und Ereignisse der deutschen und internationalen Fußballgeschichte abfragte. Der Test war mit einer mittleren Itemschwierigkeit von unter .5 schwierig genug, um Teilwissen der Teilnehmer über die Wahl eines Distraktors erfassen zu können.

Um den Wissenstest auszuwerten, wurde die Stichprobe in zwei Hälften geteilt, und nur die Antworten einer Stichprobenhälfte ausgewertet. Zur Auswertung verwendet wurde entweder die Anzahl-Korrekt-Auswertung, Experten-Optionsgewichte oder empirische Optionsgewichte, die für jede Antwortoption aus den Antworten der anderen Stichprobenhälfte bestimmt wurden. Zur Ermittlung der empirischen Optionsgewichte wurde die part-whole-korrigierte Option-Gesamt-Korrelation für alle Antwortoptionen der Items berechnet (Haladyna, 1990). Zehn bekannte deutsche Fußballjournalisten nahmen als Experten teil und vergaben Experten-Optionsgewichte für alle Items des Fußballwissenstests. Alle Experten hatten in der Vergangenheit entweder im Fernsehen oder Radio über Fußball in der deutschen Bundesliga berichtet. Die Experten wurden aufgefordert, für jeden

Tabelle 5. Beispielitem („1992 wurde überraschend Dänemark Europameister. Wen schlug das Team damals im Finale 2:0?") des Fußballwissenstests und das dazugehörige Auswertungsschema für die Anzahl-Korrekt-Auswertung, die empirische Optionsgewichtung und die Experten-Optionsgewichtung.

Antwortoption	**Anzahl-Korrekt**	**Empirische Optionsgewichtung**	**Experten-Optionsgewichtung**
Deutschland (korrekt)	1	.41	100.00
Niederlande	0	−.20	33.50
Schweden	0	−.28	23.50

Anmerkungen. Die empirischen Optionsgewichte wurden als mittlere punkt-biseriale Korrelation zwischen der Wahl einer Antwortoption und dem Testscore berechnet. Die Experten-Optionsgewichte wurden auf einer Skala von 0 bis 100 Punkten angegeben.

Distraktor die Punktzahl zu bestimmen, die ihrer Meinung nach für die Wahl dieser Antwortoption noch vergeben werden sollte. Für die Experten-Optionsgewichtung war das Gewicht für die korrekte Antwortoption immer auf 100 Punkte festgelegt. Dies entsprach dem Vorgehen von Cross et al. (1980), Hambleton et al. (1970) und Kansup und Hakstian (1975), die ebenfalls die Gewichte der korrekten Antwortoptionen konstant hielten. Diese Prozedur erleichterte den Experten die Aufgabe, da sie auf diese Weise nur Gewichte für die beiden Distraktoren angeben brauchten – jeweils auf einer Skala von 0 bis 100 Punkten. Neben den individuellen Optionsgewichten der Experten wurde außerdem für jede Antwortoption das mittlere Experten-Optionsgewicht über alle Experten hinweg bestimmt. Tabelle 5 zeigt ein Item des Fußballwissenstests und das Auswertungsschema für die drei Auswertungsverfahren.

Als abhängige Variablen wurden die Reliabilität und Validität der Testscores der drei konkurrierenden Verfahren ermittelt. Für jedes Auswertungsverfahren wurde die interne Konsistenz über Cronbachs α (Cronbach, 1951) als Maß für die Reliabilität bestimmt. Als Maß für die konvergente Validität wurde die Korrelation zwischen den Testscores und der selbsteingeschätzten Fußballexpertise der Teilnehmer als externes Kriterium ermittelt. Um zu prüfen, inwieweit die Ergebnisse der Vergleiche zwischen der empirischen Optionsgewichtung, der Experten-Optionsgewichtung und der Anzahl-Korrekt-

Auswertung von der vorgenommenen Stichprobenteilung abhing, wurde erstmalig eine *wiederholte randomisierte Substichproben-Validierung* (*repeated random subsampling validation*) durchgeführt. Dazu wurden alle Analysen 10,000-mal wiederholt, und für jeden Durchgang wurde eine neue Stichprobenteilung sowie die Berechnung der Reliabilitäts- und Validitätsmaße vorgenommen. Über alle 10,000 Durchgänge hinweg ließen sich anschließend Mittelwerte und Konfidenzintervalle berechnen. Neben der Möglichkeit den Einfluss der Stichprobenteilung auf das Ergebnis zu untersuchen, sollten Ergebnisse, die auf einer großen Zahl an Kreuzvalidierungen beruhen, außerdem robuster und besser generalisierbar sein als Ergebnisse, die nur auf einer einzigen Stichprobenteilung beruhen. Zum Vergleich wurde zusätzlich eine Auswertung mit Hilfe konventioneller Methoden aus früheren Studien durchgeführt.

Um Unterschiede zwischen den drei Auswertungsverfahren in der Reliabilität und Validität auf statistische Signifikanz zu prüfen, wurden zwei verschiedene Analysen durchgeführt. Auf der einen Seite wurde eine konventionelle Analyse vorgenommen, bei der die Reliabilität und Validität der Auswertungsverfahren separat für jeden der 10,000 Durchgänge verglichen wurden. Zum Vergleich der Reliabilität wurde der Signifikanztest von Feldt et al. (1987) für den Vergleich zweier abhängiger Cronbach-α-Koeffizienten angewendet, der im R-Paket *cocron* (Version 1.0-0; Diedenhofen, 2013) implementiert ist. Zum Vergleich der Validität, die über die Korrelation zwischen Testscore und Außenkriterium ermittelt wurde, kam der Test von Steiger (1980) für den Vergleich von zwei abhängigen Korrelationen zum Einsatz, der im R-Paket *cocor* (Version 1.1-0; das Statistikpaket wurde im Rahmen dieser Arbeit entwickelt und ist ausführlich im nächsten Kapitel beschrieben; Diedenhofen & Musch, 2015) verfügbar ist. Auf der anderen Seite wurde die Verteilung der Cronbach-α-Koeffizienten sowie die Verteilung der Validitätskoeffizienten über alle 10,000 Durchgänge der wiederholten randomisierten Substichproben-Validierung hinweg ermittelt, und das 95%-Konfidenzintervall über die Berechnung der 2.5%- und 97.5%-Perzentile der Verteilungen bestimmt. Unterschiede in der Reliabilität und Validität zwischen zwei

Auswertungsverfahren wurden als statistisch signifikant erachtet, wenn die so konstruierten Konfidenzintervalle sich nicht überlappten.

Alle Analysen wurden im Rahmen der wiederholten randomisierten Substichproben-Validierung 10,000-mal durchgeführt. Falls nicht anders angegeben, handelt es sich bei den folgenden Statistiken um Mittelwerte über 10,000 Durchgänge. Wie beabsichtigt, war der Wissenstest herausfordernd für die Teilnehmer, mit einer mittleren Itemschwierigkeit von .48 (SD = .14). Die Teilnehmer konnten einige der Items nicht lösen und entschieden sich deshalb häufig für einen Distraktor. Hinsichtlich der Itemschwierigkeit waren damit günstige Voraussetzungen für die Erfassung von Teilwissen durch die Optionsgewichtung gegeben.

Um die Reliabilität der Testscores für die Anzahl-Korrekt-Auswertung, die empirische Optionsgewichtung und die Experten-Optionsgewichtung zu vergleichen, wurde Cronbachs α berechnet (Tabelle 6). Deskriptiv erreichten die empirischen Optionsgewichte die höchste Reliabilität (α = .82), wohingegen es keinen Unterschied zwischen der Anzahl-Korrekt-Auswertung (α = .77) und den über alle Experten hinweg gemittelten Experten-Optionsgewichten gab (α = .77). Wurde jeder Experte separat ausgewertet, erzielte keiner der Experten deskriptiv eine höhere Reliabilität als die Anzahl-Korrekt-Auswertung. Die Verbesserung der Reliabilität, die durch die empirische Optionsgewichtung erreicht wurde, war laut h-Statistik (Haladyna, 1990, p. 236) äquivalent zu einer Verbesserung, die durch eine Testverlängerung um 36% bei einer konventionellen Auswertung erwartet werden konnte.

Um die Unterschiede in der Reliabilität auf statistische Signifikanz zu prüfen, wurden zwei verschiedene Ansätze verfolgt. Zum einen wurde in einer konventionellen Analyse, separat für jeden Durchgang der wiederholten randomisierten Substichproben-Validierung, die Cronbach-α-Koeffizienten auf statistische Signifikanz überprüft. Zum anderen wurde die Verteilung der α-Koeffizienten über alle 10,000 Durchgänge der Substichproben-Validierung analysiert. Die Ergebnisse zeigten, dass in jeder der 10,000 Durchgänge die empirische Optionsgewichtung zu einer signifikant höheren internen Konsistenz führte als

die Anzahl-Korrekt-Auswertung (Tabelle 6). Die Experten-Optionsgewichtung verbesserte die Reliabilität nur in 1% der Durchgänge. Außerdem wurden die Mittelwerte und 95%-Konfidenzintervalle (95%-CI) für die Cronbach-α-Koeffizienten über alle 10,000 Durchgänge der Substichproben-Validierung hinweg berechnet (Tabelle 6). Ein Unterschied in der Reliabilität zwischen zwei Auswertungsverfahren wurde als statistisch signifikant erachtet, wenn sich die Konfidenzintervalle für die α-Koeffizienten der beiden Verfahren nicht überlappten. Die empirische Optionsgewichtung, 95%-CI für α: [.80, .83], erzielte eine höhere Reliabilität als die Experten-Optionsgewichtung, [.74, .79], wenn die Gewichte über alle Experten gemittelt wurden. Darüber hinaus übertraf die empirische Optionsgewichtung auch die Anzahl-Korrekt-Auswertung, [.74, .79]. Nur die Optionsgewichte eines Experten (Experte 10) führten zu einer signifikant niedrigeren Reliabilität, [.64, .71], im Vergleich zur Anzahl-Korrekt-Auswertung. Da sich die Konfidenzintervalle der über alle Experten gemittelten Experten-Optionsgewichte, [.74, .79], und der Anzahl-Korrekt-Auswertung, [.74, .79], überlappten, unterschieden sich die beiden Auswertungsmethoden folglich nicht in der Reliabilität.

Tabelle 6. Cronbachs α-Reliabilität und konvergente Validität (r_c) ermittelt aus den Antworten der Teilnehmer zu den 27 Items des Fußballwissenstests, ausgewertet mit der Anzahl-Korrekt-Auswertung, der empirischen Optionsgewichtung und der Experten-Optionsgewichtung.

	Reliabilität				**Validität**		
			Anteil $p < .05$			**Anteil $p < .05$**	
Auswertungsverfahren	**α [95%-CI]**	**h**	**$<$ AK**	**$>$ AK**	**r_c [95%-CI]**	**$<$ AK**	**$>$ AK**
Anzahl-Korrekt-Auswertung (AK)	.77 [.74 .79]	–	–	–	.71 [.67 .75]	–	–
Empirische Optionsgewichtung	.82 [.80 .83]	1.36	.00	1.00	.73 [.69 .76]	.00	.46
Experten-Optionsgewichtung							
Mittel über alle Experten	.77 [.74 .79]	1.00	.04	.01	.71 [.67 .75]	.00	.05
Experte 1	.76 [.73 .79]	0.97	.40	.00	.71 [.67 .75]	.00	.02
Experte 2	.75 [.72 .78]	0.90	.99	.00	.70 [.66 .74]	.01	.00
Experte 3	.75 [.71 .77]	0.88	1.00	.00	.70 [.66 .74]	.39	.00
Experte 4	.75 [.72 .78]	0.92	1.00	.00	.71 [.67 .75]	.00	.00
Experte 5	.75 [.72 .78]	0.90	1.00	.00	.70 [.67 .74]	.01	.00
Experte 6	.77 [.74 .79]	1.00	.03	.01	.71 [.67 .75]	.00	.07
Experte 7	.77 [.74 .79]	1.00	.05	.00	.71 [.67 .74]	.00	.00
Experte 8	.75 [.72 .78]	0.90	.98	.00	.70 [.65 .74]	.19	.00
Experte 9	.76 [.73 .79]	0.95	.81	.00	.71 [.67 .75]	.00	.02
Experte 10	.68 [.64 .71]	0.63	1.00	.00	.68 [.64 .72]	.42	.00

Anmerkungen. Gemittelte Werte über die 10,000 Durchgänge der wiederholten randomisierten Substichproben-Validierung; in Klammern sind 95%-Konfidenzintervalle (95%-CI) angegeben. Es wird außerdem der Anteil an Durchgängen aufgeführt, in denen die empirische Optionsgewichtung bzw. die Experten-Optionsgewichtung eine signifikant niedrigere ($<$ AK) oder höhere ($>$ AK) Reliabilität oder Validität als die Anzahl-Korrekt-Auswertung (AK) erzielte. Die h-Statistik gibt den Faktor an, um den ein mit der Anzahl-Korrekt-Auswertung ausgewerteter Test verlängert werden müsste, um die gleiche interne Konsistenz zu erreichen, wie die empirische Optionsgewichtung bzw. die Experten-Optionsgewichtung.

Zur Bestimmung der Validität der drei Auswertungsverfahren wurde jeweils die Korrelation zwischen dem Testscore und der selbsteingeschätzten Fußballexpertise der Teilnehmer als Außenkriterium berechnet (Tabelle 6). Die interne Konsistenz der Skala zur Erfassung der Fußballexpertise war zufriedenstellend ($\alpha = .73$). Die Ermittlung einer Skalensumme, über die Items der Skala hinweg, schien damit gerechtfertigt. Deskriptiv erzielte nur die empirische Optionsgewichtung, und nicht die Experten-Optionsgewichtung, eine höhere Validität als die Anzahl-Korrekt-Auswertung. Wurden die Experten separat ausgewertet, erreichten ihre Optionsgewichte Validitäten, die genauso hoch oder niedriger waren als die Validität, die durch die Anzahl-Korrekt-Auswertung erreicht wurde.

Um zu überprüfen, ob sich die Unterschiede in der Validität zwischen den Auswertungsverfahren statistisch signifikant voneinander unterschieden, wurde wieder sowohl der konventionelle Analyseansatz als auch der Ansatz der wiederholten randomisierten Substichproben-Validierung verfolgt. In 46% der 10,000 Durchgänge war die Korrelation zwischen Testscore und Außenkriterium für die empirische Optionsgewichtung signifikant höher als für die Anzahl-Korrekt-Auswertung. Im Gegensatz dazu erreichten die auf Experten-Optionsgewichten basierenden Testscores nur in 5% der 10,000 Durchgänge eine signifikant höhere Korrelation mit dem Außenkriterium als die Anzahl-Korrekt-Auswertung. Für die wiederholte randomisierte Substichproben-Validierung wurden die Mittelwerte und 95%-Konfidenzintervalle jeweils für die Validitätskoeffizienten der drei Auswertungsverfahren über die 10,000 Durchgänge hinweg berechnet (Tabelle 6). Da sich die Konfidenzintervalle aller Auswertungsverfahren überlappten, konnte keiner der beobachteten Unterschiede als statistisch signifikant eingestuft werden.

Studie 2 untersuchte, ob Optionsgewichte, die entweder auf empirischen Daten oder auf Experten-Urteilen beruhen, in der Lage sind, die Reliabilität und Validität eines Multiple-Choice-Wissenstests im Vergleich zur konventionellen Anzahl-Korrekt-Auswertung zu verbessern. Um *overfitting* und *capitalization on chance* zu vermeiden, teilten frühere Studien die Stichprobe zufällig und führten die Berechnung und Anwendung der empirischen Optionsgewichtung an zwei unabhängigen Substichproben durch. Die Auswahl

einer zufälligen Teilung ist ein Forscherfreiheitsgrad, der einen bedeutsamen Einfluss auf das Ergebnis haben kann. Daher wurde eine wiederholte randomisierte Substichproben-Validierung durchgeführt, indem alle Analysen 10,000-mal wiederholt und die Kennwerte für die Reliabilität und Validität über alle Durchgänge hinweg gemittelt wurden. Die Ergebnisse der Auswertung zeigten, dass die empirische Optionsgewichtung die Reliabilität eines Wissenstests im Vergleich zur konventionellen Anzahl-Korrekt-Auswertung verbesserte. Diese Verbesserung entsprach einer Testverlängerung von 36%. Die Berechnung empirischer Optionsgewichte birgt folglich das Potenzial, Testentwicklern die Erstellung zusätzlicher Items zu ersparen. Die Verwendung von Optionsgewichten, die durch Experten festgelegt wurden, verbesserte hingegen nicht die Reliabilität und verminderte sie sogar in einem Fall. Hinsichtlich der Validität konnten weder empirische Optionsgewichte noch Experten-Optionsgewichte Verbesserungen gegenüber der Anzahl-Korrekt-Auswertung erzielen. Dies galt unabhängig davon, ob die Experten-Optionsgewichte über alle Experten gemittelt wurden, oder ob für jeden einzelnen der 10 Experten separat eine Auswertung durchgeführt wurde. Die wiederholte Durchführung einer konventionellen Analyse, die dem Vorgehen in früheren Studien entsprach, offenbarte eine starke Abhängigkeit der Ergebnisse von der vorgenommenen Stichprobenteilung. In etwa der Hälfte der 10,000 Durchgänge fand die konventionelle Analyse eine Verbesserung der Validität durch die empirische Optionsgewichtung, für die übrigen Durchgänge wurde, in Übereinstimmung mit dem Ergebnis der Substichproben-Validierung, keine Validitätsverbesserung festgestellt. Hinsichtlich der Reliabilität entsprachen die Ergebnisse aller wiederholten konventionellen Analysen dem Ergebnis der Substichproben-Validierung.

5 Artikel 3: Ein R-Paket und Web-Interface zum Vergleich von abhängigen und unabhängigen Korrelationen

Die Bestimmung des Zusammenhanges zweier Variablen ist das Ziel vieler Forschungsvorhaben. Eine der besonders häufig verwendeten statistischen Methoden, die zur Ermittlung der Beziehung zweier Variablen in den empirischen Wissenschaften zum Einsatz kommt, ist die Produkt-Moment-Korrelation. Sie quantifiziert die Stärke eines linearen Zusammenhanges zwischen zwei Variablen auf einem Wertebereich von -1 bis 1, und kann folglich positiv, negativ oder null sein. In vielen Forschungskontexten ist der Vergleich zweier Korrelationen von Interesse. Dies kann beispielsweise der Fall sein, wenn ermittelt werden soll, ob sich der Zusammenhang zwischen zwei Variablen nach einer Intervention verändert hat, oder ob sich zwei Gruppen hinsichtlich eines Zusammenhanges zwischen zwei Variablen unterscheiden.

In der psychologischen Diagnostik werden Korrelationen häufig für die Bestimmung der psychometrischen Eigenschaften von Testverfahren verwendet: Die Trennschärfe eines Items, die Test-Retest-Reliabilität, die Split-Half-Reliabilität und die Paralleltest-Reliabilität werden beispielsweise über die Berechnung von Korrelationen ermittelt (Amelang, Schmidt-Atzert, Fydrich & Zielinski, 2006). Auch zur Bestimmung der konkurrenten, divergenten oder prädiktiven Validität eines Tests kommen häufig Korrelationen zum Einsatz. Um festzustellen inwieweit die Scores von Testteilnehmern mit einem Validitätskriterium assoziiert sind, bietet sich die Berechnung der Korrelation zwischen Testscore und Kriterium an. Bei der Entwicklung und Anwendung von Testverfahren werden oft unterschiedliche Tests hinsichtlich ihrer psychometrischen Eigenschaften verglichen. Dies ist beispielsweise der Fall, wenn eine neue überarbeitete Version eines Testes mit einer alten Version verglichen werden soll. Für einen aussagekräftigen Vergleich der psychometrischen Eigenschaften zweier Testverfahren ist es notwendig, Unterschiede zwischen den

Verfahren zufallskritisch abzusichern. Da viele psychometrische Kennwerte auf Korrelationen beruhen, ist es oft erforderlich, Korrelationskoeffizienten auf ihre Unterschiedlichkeit hin zu prüfen.

Bisherige Studien, die den Einfluss der Optionsgewichtung auf die Testvalidität untersuchten, setzten dabei fast ausschließlich korrelative Methoden ein (z. B. Hambleton et al., 1970; Cross et al., 1980). Die Studien analysierten, ob auf Optionsgewichten basierende Testscores höher mit einem Außenkriterium korrelierten, als Testscores, die über eine klassische Anzahl-Korrekt-Auswertung bestimmt wurden. Häufig kamen jedoch keine Signifikanztests zum Einsatz, um den Unterschied zwischen den Korrelationen zufallskritisch abzusichern (z. B. Hambleton et al., 1970; Cross et al., 1980). Schlussfolgerungen, die allein auf der Grundlage deskriptiver Statistiken gezogen werden, sind jedoch nur eingeschränkt interpretierbar. Fehlende Signifikanztestungen sind wahrscheinlich nicht auf einen Mangel an geeigneten statistischen Testverfahren zurückzuführen, sondern gehen vermutlich auf die unzureichende Verfügbarkeit geeigneter Software und das mangelnde Wissen um die Notwendigkeit der Durchführung eines Signifikanztests zurück. Populäre Statistikpakete, wie SPSS oder SAS, bieten beispielsweise keine Tests zum Vergleich von Korrelationen an. Auch in Forschungsfeldern abseits der psychologischen Diagnostik wird die Anwendung von Korrelationsvergleichen häufig vernachlässigt. In neurowissenschaftlichen Untersuchungen werden beispielsweise Verhaltensmaße mit Aktivitäten in bestimmten Hirnarealen korreliert, um Areale zu identifizieren, die am stärksten bei der Durchführung einer bestimmten Aufgabe beteiligt sind. Rousselet und Pernet (2012) kritisierten, dass derartige Untersuchungen selten statistische Tests durchführten, um Unterschiede zwischen Korrelationen zufallskritisch abzusichern. Stattdessen fielen Autoren häufig auf einen statistischen Trugschluss herein, indem sie einen Unterschied zwischen zwei Korrelationen für statistisch bedeutsam hielten, wenn die eine Korrelation signifikant und die andere Korrelation nicht signifikant von null verschiedenen war (Nieuwenhuis, Forstmann & Wagenmakers, 2011). Für einen gültigen und aussagekräftigen Vergleich zwischen zwei

Korrelationen, ist es jedoch notwendig, beide Korrelationen mit einem geeigneten statistischen Test direkt zu vergleichen. Um dem Mangel an Statistiksoftware zum Vergleich von Korrelationen zu begegnen und ein stärkeres Bewusstsein für Signifikanztests zum Vergleich von Korrelationen zu schaffen, war es ein Ziel dieser Arbeit, eine Software zum Vergleich von abhängigen und unabhängigen Korrelationen zu entwickeln. Die Software sollte eine umfangreiche Auswahl an Testverfahren über eine einfache Benutzeroberfläche einer Vielzahl von Anwendern zur Verfügung stellen, und gleichzeitig zur Automatisierung in R-Auswertungsskripte integriert werden können.

Selbst wenn die Notwendigkeit eines direkten statistischen Vergleiches zweier Korrelationen erkannt worden ist, stellt sich die entscheidende Frage, welcher der vielen zur Verfügung stehenden Tests für den vorliegenden Anwendungsfall angemessen ist. Um einen geeigneten Test zu wählen, muss zunächst unterschieden werden, welcher der folgenden drei Fälle zutrifft (siehe auch Abbildung 1): (1) Die beiden zu vergleichenden Korrelationen wurden in zwei unabhängigen Gruppen A und B ermittelt. Dies ist der Fall, wenn beispielsweise die Korrelation zwischen Verträglichkeit und Extraversion in zwei verschiedenen Gruppen A und B ($\rho_A = \rho_B$) verglichen werden soll. Sind die beiden Gruppen jedoch abhängig, muss zunächst die Art der Abhängigkeit festgestellt werden: (2) Die beiden Korrelationen können sich überlappen ($\rho_{A12} = \rho_{A23}$). Dieser Fall liegt vor, wenn beide Korrelationen eine gemeinsame Variable haben. Hier beziehen sich ρ_{A12} und ρ_{A23} auf die in Gruppe A geschätzte Populationskorrelation zwischen den Variablen 1 und 2 bzw. den Variablen 2 und 3. Dies ist zum Beispiel der Fall, wenn innerhalb derselben Gruppe A getestet werden soll, ob die Korrelation zwischen Verträglichkeit und Extraversion höher ist als die Korrelation zwischen Verträglichkeit und Ängstlichkeit. (3) Zwei abhängige Korrelationen können sich auch nicht überlappen ($\rho_{A12} = \rho_{A34}$), dann haben beide Korrelationen keine Variable gemeinsam. Dieser Fall trifft beispielsweise zu, wenn innerhalb derselben Gruppe A überprüft werden soll, ob die Korrelation zwischen Verträglichkeit und Extraversion kleiner ist als die Korrelation zwischen Ängstlichkeit und

Gewissenhaftigkeit. Es handelt sich ebenfalls um einen Vergleich zweier nicht-überlappender Korrelationen, wenn getestet werden soll, ob die Korrelation zwischen zwei Variablen nach einer Intervention gestiegen bzw. gefallen ist.

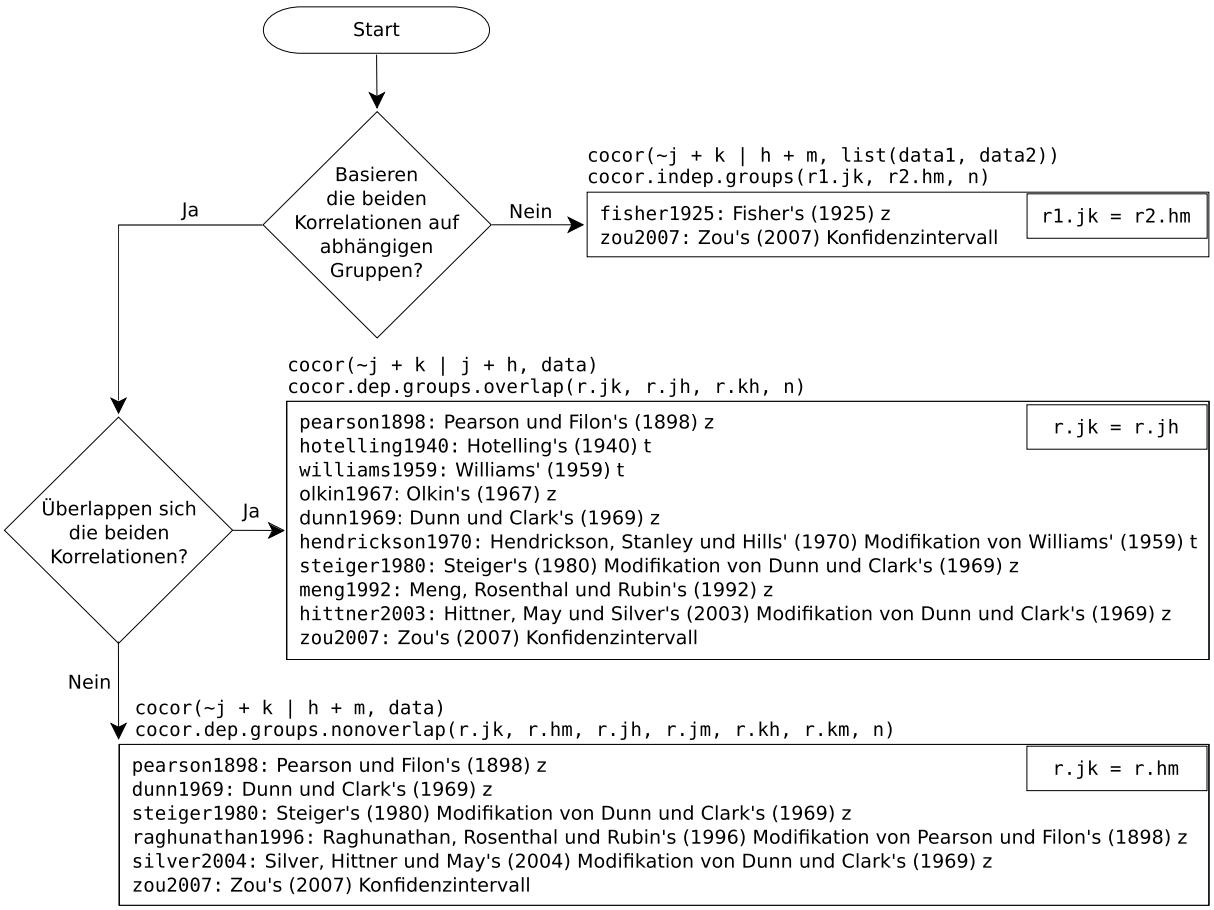

Abbildung 1. Ein Flussdiagramm, das alle verfügbaren Tests, die vier Hauptfunktionen von *cocor* und deren Verwendung darstellt. Für jeden Fall ist ein Beispiel für die Formel aufgeführt, die als Argument an die `cocor()`-Funktion übergeben wird, sowie die Korrelationskoeffizienten, die für die Funktionen `cocor.indep.groups()`, `cocor.dep.groups.overlap()` und `cocor.dep.groups.nonoverlap()` benötigt werden. Die Test-Bezeichnung vor dem Doppelpunkt kann den Funktionen als Argument übergeben werden, um nur bestimmte Tests zu berechnen.

Für jeden der drei Fälle wurden jeweils verschiedene Signifikanztests vorgeschlagen. Ein Überblick über alle Tests ist im Flussdiagramm in Abbildung 1 gegeben. May

und Hittner (1997) verglichen die Teststärke und die α-Fehler-Wahrscheinlichkeit verschiedener Tests für abhängige Korrelationen mit überlappenden Variablen und fanden, dass keiner der Tests bezüglich beider Maße den anderen Tests eindeutig überlegen war. Stattdessen berichteten sie, dass die optimale Wahl eines Tests von der Stichprobengröße, der Prädiktor-Interkorrelation, der Effektgröße und der Prädiktor-Kriterium-Interkorrelation abhängt. Da keine klare Empfehlung für einen dieser Tests formuliert werden kann, die auf alle denkbaren Konstellationen zutrifft, und da unterschiedliche Tests für eine Fragestellung optimal sein können, ist es wichtig, dass Anwendern alle Tests zur Verfügung stehen. Ausführliche Diskussionen der konkurrierenden Tests für abhängige Korrelationen mit überlappenden Variablen finden sich in Dunn und Clark (1971), Hittner, May und Silver (2003), May und Hittner (1997), Neill und Dunn (1975) und Steiger (1980). Für abhängige Korrelationen mit nicht-überlappenden Variablen werden die Vor- und Nachteile der unterschiedlichen Tests in Raghunathan, Rosenthal und Rubin (1996), Silver, Hittner und May (2004) und Steiger (1980) diskutiert. Zou (2007) schlug einen Test vor, der im Gegensatz zu den meisten anderen Tests, auf der Berechnung von Konfidenzintervallen beruht. Zou's (2007) Konfidenzintervalle lassen sich für unabhängige und abhängige Korrelationen mit entweder überlappenden oder nicht-überlappenden Variablen berechnen. Konfidenzintervalle werden häufig gegenüber Signifikanztests als überlegen angesehen, weil sie die Größe und Präzision eines geschätzten Effekts unabhängig voneinander quantifizieren (Cohen, 1994; Olkin & Finn, 1995). Konfidenzintervalle ermöglichen es ebenfalls zu testen, ob sich eine Korrelation signifikant von null oder einer beliebigen Konstante unterscheidet, oder ob die Differenz zwischen zwei Korrelationen einen bestimmten Grenzwert überschreitet.

Viele populäre Statistikprogramme bieten keine oder nur eine geringe Teilmenge der verfügbaren Signifikanztests zum Vergleich von Korrelationen an. Bei Programmen, die speziell für den statistischen Vergleich von Korrelationen entwickelt wurden, handelt es sich meist um isolierte eigenständige Programme, die über keine grafische Benutzeroberfläche verfügen. DEPCOR (Silver, Hittner & May, 2006) ist beispielsweise eine Software,

die auf Vergleiche zwischen zwei abhängigen Korrelationen mit überlappenden und nicht-überlappenden Variablen beschränk ist. Das Programm ist in Fortran geschrieben und läuft unter Windows in der MS-DOS-Eingabeaufforderung. DEPCORR (Hittner & May, 1998) ist ein SAS-Makro für den Vergleich von zwei abhängigen überlappenden Korrelationen. Die neuste Version von SAS/STAT-Software (Version 9.4) läuft unter Windows- und Linux-Systemen. DEPCORR verfügt jedoch über keine grafische Benutzeroberfläche und deckt nur einen der drei beschriebenen Fälle beim Vergleich von Korrelationen ab. Die beiden Pakete *psych* (Revelle, 2014) und *multilevel* (Bliese, 2013) für die Programmiersprache R (R Core Team, 2015) bieten ebenfalls Funktionen zum Vergleich von zwei abhängigen und unabhängigen Korrelationen an. Die Funktionen implementieren jedoch jeweils nur einen oder zwei der vielen verfügbaren Tests zum Vergleich von Korrelationen. Darüber hinaus existiert keine grafische Benutzeroberfläche, über die sich die Funktionen ausführen lassen. Weaver und Wuensch (2013) veröffentlichten Skripte, zum Vergleich von abhängigen und unabhängigen Korrelationen, die jedoch nur unter SPSS und SAS lauffähig sind.

Mit *cocor* (Version 1.1-0) wurde im Rahmen der vorliegenden Arbeit eine umfassende Lösung zum Vergleich von zwei Korrelationen entwickelt, die entweder auf abhängigen oder unabhängigen Gruppen beruhen können. Das Statistikpaket ergänzt die R-Statistikumgebung (R Core Team, 2015), die für Windows-, Mac- und Linux-Systeme frei verfügbar ist. Das Programm kann als R-Paket aus dem *Comprehensive R Archive Network* (CRAN) heruntergeladen werden (http://cran.r-project.org/package=cocor). Für eine Installation muss lediglich `install.packages("cocor")` in die R-Konsole eingegeben werden; die Funktionalität des Pakets wird über die Eingabe `library("cocor")` verfügbar. Die Funktion `cocor()` berechnet und vergleicht Korrelationen auf der Basis von Rohdaten. Die zugrundeliegenden Variablen werden über ein Formel-Interface spezifiziert (Abbildung 1). Wenn Rohdaten nicht zur Verfügung stehen, bietet cocor drei Funktionen zum Vergleich von Korrelationen an, die bereits berechnet wurden: Die Funktion `cocor.indep.groups()` vergleicht zwei unabhängige Korrelationen, wohingegen die

Funktionen `cocor.dep.groups.overlap()` und `cocor.dep.groups.nonoverlap()` zwei abhängige Korrelationen mit überlappenden bzw. nicht-überlappenden Variablen vergleichen. Alle Funktionen erlauben das Argument `null.value` zu spezifizieren, um mit Hilfe der Konfidenzintervalle von Zou (2007) zu überprüfen, ob der Unterschied zwischen zwei Korrelationen einen bestimmten Grenzwert übersteigt. Standardmäßig werden alle verfügbaren Tests berechnet. Um nur bestimmte Tests auszuführen, können entsprechende Test-Bezeichnungen über das `test`-Argument an die Funktionen übergeben werden. Das Flussdiagramm in Abbildung 1 listet alle verfügbaren Tests auf, und gibt an, über welche Bezeichnung sie aufgerufen werden können (z. B. `zou2007` für Zou's (2007) Konfidenzintervall). Im Vergleich zu bestehender Software bietet cocor die umfassendste Auswahl an Testverfahren an, um Korrelationen auf Unterschiedlichkeit zu testen. Insbesondere ist cocor das erste R-Paket, das die Tests von Zou (2007) implementiert. Außerdem ist hervorzuheben, dass über eine einzige Funktion (`cocor()`) mit Hilfe eines Formel-Interfaces Vergleiche von unabhängigen und abhängigen Korrelationen mit entweder überlappenden oder nicht-überlappenden Variablen durchgeführt werden können.

Als Einschränkung ist zu nennen, dass cocor auf den Vergleich von zwei Korrelationen begrenzt ist. Der simultane Vergleich von mehr als zwei Korrelationen benötigt Tests, die über den Umfang der Software hinausgehen (Levy, 1976; Paul, 1989; Silver, Zaikina, Hittner & May, 2008). Zweitens erlaubt cocor nicht die Verwendung von Strukturgleichungsmodellen, die benötigt werden, um komplexere Ansätze zum statistischen Vergleich von Korrelationen durchzuführen (Cheung & Chan, 2004; Cheung, 2009).

Es stehen zwei komfortable Möglichkeiten zur Verfügung cocor über eine grafische Benutzeroberfläche zu verwenden. Erstens beinhaltet das R-Paket ein Plug-in (Abbildung 2) für die plattformunabhängige R-Entwicklungsumgebung RKWard (Rödiger, Friedrichsmeier, Kapat & Michalke, 2012). Zweitens gibt es für Nutzer, die keine Erfahrung mit R haben, ein Web-Interface unter http://comparingcorrelations.org (Abbildung 3). Damit bietet cocor das Beste aus zwei Welten: Auf der einen Seite verfügt cocor über die Mächtigkeit der R-Skriptsprache, die es erlaubt Auswertungen zu automatisieren.

Auf der anderen Seite ermöglichen zwei grafische Benutzeroberflächen auch unerfahrenen Anwendern cocor komfortabel einzusetzen. Da cocor Teil der R-Statistikumgebung ist, kann es nahtlos in R-Analysen integriert werden. R-Code kann über die grafischen Benutzeroberflächen generiert werden und anschließend in Batch-Analysen verwendet werden. Weil cocor unter der *GNU General Public License* (GPL; Version ≥ 3) veröffentlicht wurde, können alle Nutzer das Paket begutachten, verwenden, kopieren, verändern und unter der gleichen Lizenz Weiterentwicklungen veröffentlichen.

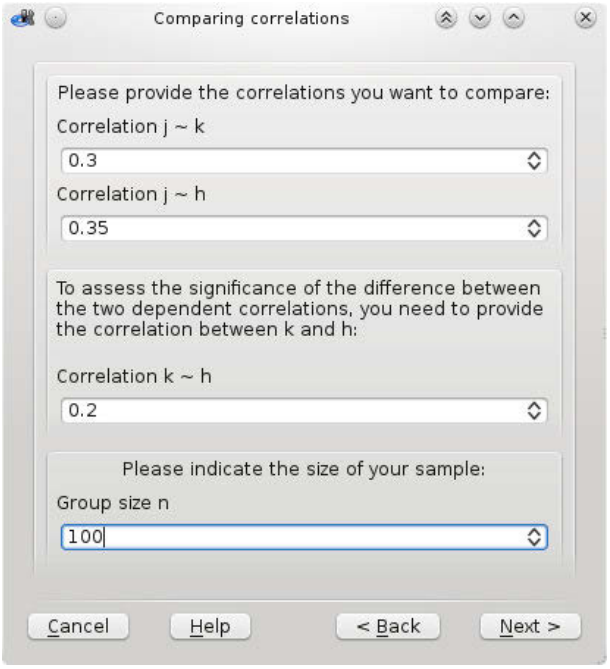

Abbildung 2. Screenshot des cocor-Plug-ins für RKWard.

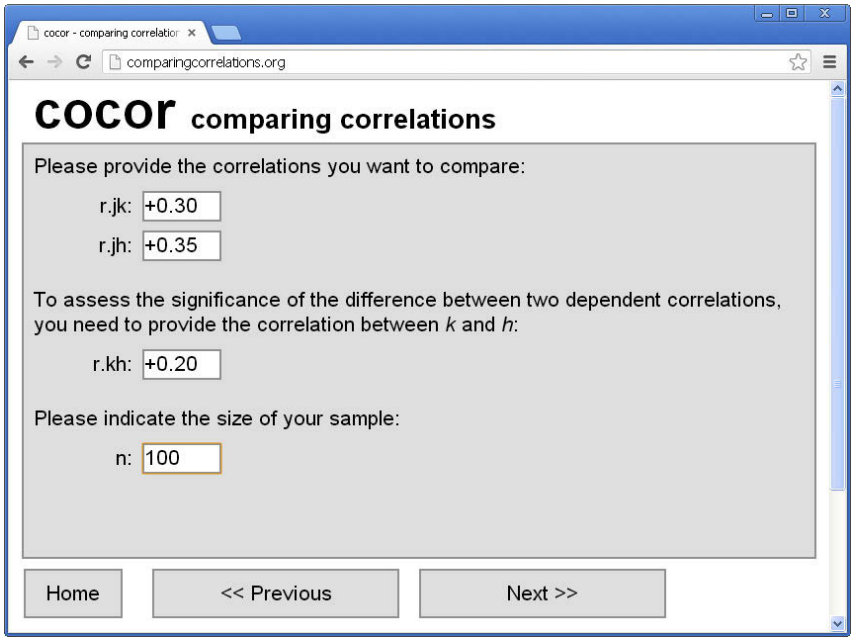

Abbildung 3. Screenshot des cocor-Web-Interfaces auf http://comparingcorrelations.org.

6 Allgemeine Diskussion

Das vielfach eingesetzte Multiple-Choice-Format wurde häufig dafür kritisiert, dass es bei einer klassischen dichotomen Anzahl-Korrekt-Auswertung keine Teilpunkte gewährt und Teilwissen, das durch die Wahl eines Distraktors widergespiegelt wird, unberücksichtigt bleibt (Frary, 1989). Um diesen Kritikpunkten zu begegnen, sieht die Optionsgewichtung vor, jede Antwortoption eines Multiple-Choice-Items mit einem individuellen Gewicht zu versehen, das bei der Wahl einer Option dem Testteilnehmer gutgeschrieben wird. Für empirische Optionsgewichte, die aus den Antworten von Testteilnehmern berechnet wurden, fielen die Ergebnisse vorheriger Studien zur Validität nicht eindeutig aus. Viele der untersuchten Tests wurden ursprünglich für eine dichotome Anzahl-Korrekt-Auswertung konstruiert und waren daher möglicherweise ungeeignet, um das Potenzial der polytomen Optionsgewichtung umfassend zu evaluieren. Darüber hinaus kamen in den Untersuchungen ausschließlich korrelative Studiendesigns zum Einsatz. In Studie 1 wurde zum ersten Mal ein experimenteller Ansatz zur Validierung der empirischen Optionsgewichtung verfolgt. Dazu wurden die Wissensstände der Teilnehmer experimentell manipuliert und als Außenkriterium zur Bestimmung der Validität verwendet. Die Items des untersuchten Tests verfügten über Antwortoptionen, die für Teilnehmer mit verschiedenen Wissensständen unterschiedlich attraktiv waren. Die Ergebnisse zeigten, dass die empirische Optionsgewichtung sowohl die Reliabilität als auch die Validität gegenüber einer konventionellen Anzahl-Korrekt-Auswertung verbessern konnte. Frühere Studien zur Beantwortung der Frage, ob Optionsgewichte empirisch oder durch Expertenurteile bestimmt werden sollten, lieferten keine eindeutig zu interpretierenden Ergebnisse. Vorherige Studien führten zur Evaluation der empirischen Optionsgewichtung jedoch Kreuzvalidierungen durch, die nur auf einer einzigen Stichprobenteilung basierten, ohne zu untersuchen, ob die Wahl der Stichprobenteilung die Ergebnisse beeinflusste. Um diesen Forscherfreiheitsgrad (Simmons et al., 2011) zu vermeiden, wurde in Studie 2 sowohl eine konventionelle Analyse als auch eine wiederholte randomisierte Substichproben-Validierung mit insgesamt

10,000 Stichprobenteilungen durchgeführt, um Vergleiche zwischen der empirischen Optionsgewichtung, der Experten-Optionsgewichtung und der Anzahl-Korrekt-Auswertung vorzunehmen. Das Ergebnis der konventionellen Analyse zeigte, dass die durchgeführte Stichprobenteilung einen maßgeblichen Einfluss auf das Ergebnis hatte. Die wiederholte randomisierte Substichproben-Validierung ergab hingegen, dass nur die empirische Optionsgewichtung, und nicht die Experten-Optionsgewichtung, die Testgüte des untersuchten Wissenstest verbesserte; die Verbesserung beschränkte sich jedoch auf die Reliabilität.

Generell hängt das Ausmaß, in dem Multiple-Choice-Tests von der Optionsgewichtung profitieren können, von den Eigenschaften des Tests ab. Die Optionsgewichtung kann einen Test durch die Vergabe von Teilpunkten für Teilwissen nur verbessern, wenn eine ausreichende Zahl an Testteilnehmern über unvollständiges Wissen verfügt und daher bei einigen der Items einen Distraktor wählt. Schwierige Tests sollten daher eher von einer Optionsgewichtung profitieren als leichte Tests (Haladyna, 1990). Sowohl in Studie 1 als auch in Studie 2 wurde diese Voraussetzung berücksichtigt, indem den Teilnehmern Tests mit einer mittleren Schwierigkeit von .48 bzw. .52 vorgegeben wurden. In Studie 1 konnte die erwartete Verbesserung der Testgüte durch die Optionsgewichtung beobachtet werden; in Studie 2 stellte sich nur eine Verbesserung hinsichtlich der Reliabilität ein.

Studie 2 und einige frühere Studien konnten keine Verbesserung der Validität durch die Optionsgewichtung feststellen. Eine mögliche Erklärung dafür ist, dass die Distraktoren der verwendeten Testitems nicht in der Lage waren, zwischen verschiedenen Abstufungen von Teilwissen zu differenzieren. In Studie 1 gingen die Validitätsverbesserungen durch die empirische Optionsgewichtung aus einer experimentellen Wissensinduktion hervor. Die drei experimentell induzierten Wissensstände wurden dabei innerhalb eines Items von den Antwortoptionen abgebildet. Die Wahrscheinlichkeit, eine bestimmte Antwortoption zu wählen, variierte folglich mit dem Wissensstand der Testteilnehmer. Diese Testeigenschaften boten günstige Voraussetzungen dafür, dass die empirische Optionsgewichtung nützliche Zusatzinformationen über das Teilwissen der Testteilnehmer gewinnen kann, die von der Anzahl-Korrekt-Auswertung nicht erfasst werden. Studie 1 demonstriert

damit, dass die empirische Optionsgewichtung die Validität eines Multiple-Choice-Tests verbessern kann, wenn dieser über Distraktoren verfügt, die im Hinblick auf die Optionsgewichtung entwickelt wurden. Die Übertragung auf andere Multiple-Choice-Tests ist am erfolgversprechendsten – und möglicherweise darauf beschränkt –, wenn die Tests ebenfalls über Distraktoren verfügen, die unterschiedlich attraktiv für Testteilnehmer mit verschiedenen Wissensständen sind. Die Antwortoptionen der Items in Studie 2 bildeten scheinbar nur bedingt die Wissensstände der Testteilnehmer ab. Für einige Items waren die empirischen Optionsgewichte und Experten-Optionsgewichte nahezu äquivalent zur konventionellen Anzahl-Korrekt-Auswertung, da beiden Distraktoren ungefähr gleich hohe Gewichte zugewiesen wurden. Die Optionsgewichtung kann die Testgüte gegenüber der Anzahl-Korrekt-Auswertung jedoch nur verbessern, wenn die Optionsgewichte sich von einem dichotomen Auswertungsschema hinreichend unterscheiden. Dies könnte erklären, warum in Studie 2 die beiden Optionsgewichtungsverfahren nicht in der Lage waren, die Testvalidität gegenüber der Anzahl-Korrekt-Auswertung zu verbessern.

Für die Konstruktion von Multiple-Choice-Tests, die mit einem Optionsgewichtungsverfahren ausgewertet werden, ist der Ansatz des *cognitive design modeling* von Embretson (1998) zu empfehlen. Um zwischen Testteilnehmern zu differenzieren, denen es nicht gelingt die korrekte Antwort zu identifizieren, sollten Distraktoren von vornherein mit dem Ziel entwickelt werden, Misskonzepte und Teilwissen aufzudecken. Distraktoren, die beispielsweise eklatante Misskonzepte repräsentieren, stellen einen effizienten Weg dar, um Testteilnehmer mit sehr wenig Wissen in einer bestimmten Domäne zu identifizieren (Lukas, 1997). Auf diese Weise können Distraktoren kreiert werden, die verschiedene potenzielle Zustände einer Wissensstruktur widerspiegeln. Eine individuelle Bepunktung der Distraktoren durch die Optionsgewichtung ermöglicht es dann, diese Wissensstruktur auch über die Wahl von Distraktoren abzubilden.

Ein weiterer denkbarer Grund, warum die Experten-Optionsgewichte in Studie 2 schlecht abschnitten, könnte die mangelnde Expertise der ausgewählten Experten sein. Da jedoch alle Experten, die an der Studie teilnahmen, professionelle Fußballjournalisten

waren, sollten sie in der Lage gewesen sein, die Antwortoptionen des Fußballwissenstests zu beurteilen. Das schlechte Abschneiden der Fußballexperten könnte allerdings darauf zurückgehen, dass es ihnen nicht gelungen ist, die Testantworten aus der Perspektive eines Novizen zu betrachten. Wenn das Urteil eines Experten darüber ungenau ist, welche Antwortoptionen Testteilnehmer mit geringem Wissen am ehesten wählen, werden die Optionsgewichte dieses Experten höchstwahrscheinlich nicht zu einer präziseren Erfassung der Wissensstände beitragen. Die Verbesserung durch die Experten-Optionsgewichtung, die von Downey (1979) berichtet wurde, könnte darauf zurückgehen, dass in der Studie Lehrer als Experten befragt wurden. Die Lehrer in der Studie verfügten möglicherweise über Erfahrung mit der Einschätzung, welche Antworten für Schüler mit unterschiedlichen Wissensständen attraktiv sind. Schlussendlich ist vorstellbar, dass zusätzlich zu einem fundierten Wissen über eine Domäne ein gewisses Testkonstruktionswissen notwendig ist, um angemessene Optionsgewichte vergeben zu können. Für die Fußballexperten, die an Studie 2 teilnahmen, kann nicht angenommen werden, dass sie über ein solches Wissen verfügten. Die Testentwickler, die von Cross et al. (1980) als Experten herangezogen wurden, erzielten ebenfalls keine Verbesserung der Testgüte durch die Vergabe von Experten-Optionsgewichten, obwohl sie über ausreichendes Testkonstruktionswissen verfügten. Zukünftige Untersuchungen sollten sich genauer mit der Frage beschäftigen, welche Anforderungen ein Experte erfüllen muss, um geeignete Optionsgewichte bestimmen zu können.

Die Ergebnisse der konventionellen Analyse, die separat für jeden Experten in Studie 2 durchgeführt wurde, legen nahe, dass Optionsgewichte über mehrere Experten gemittelt werden sollten, wenn die Experten-Optionsgewichtung verwendet wird. Die Mehrzahl der Experten bestimmte Optionsgewichte, die zu einer signifikanten Reduktion der Reliabilität in nahezu allen 10,000 Durchgängen der wiederholten randomisierten Substichproben-Validierung führten. Die Optionsgewichte der Experten 3, 8 und 10 reduzierten sogar signifikant die Validität des Wissenstests in einer beträchtlichen Anzahl von Durchgängen. Wurden die Optionsgewichte jedoch über alle Experten gemittelt, gab

es in den Durchgängen nur selten eine Reduktion der Reliabilität und in keinem Fall eine Minderung der Validität.

Frühere Evaluationen der empirischen Optionsgewichtung führten lediglich eine einzige Stichprobenteilung zur Kreuzvalidierung durch. Das Ergebnis kann jedoch davon abhängen, welche der vielen möglichen Teilungen einer Stichprobe gewählt wurde. In der Tat wurde in Studie 2 eine Instabilität der Ergebnisse über zahlreiche Stichprobenteilungen hinweg gefunden. Die Ergebnisse vorheriger Studien, in denen der Einfluss der Stichprobenteilung nicht untersucht wurde, müssen daher mit Vorsicht interpretiert werden. In Studie 2 war die Wahrscheinlichkeit, auf der Basis einer einzigen Stichprobenteilung eine signifikante Verbesserung der Testvalidität zu beobachten, 46%. Der Ausgang der Analyse würde daher einem Münzwurf gleichen, wenn sie auf den Methoden früherer Studien beruhte. Im Gegensatz dazu lieferten die Konfidenzintervalle, die über eine große Zahl von Stichprobenteilungen berechnet wurden, ein eindeutiges Ergebnis. Wann immer ein Zufallsprozess Teil eines Forschungsdesigns ist (z. B. wenn zur Kreuzvalidierung eine randomisierte Auswahl getroffen wird), ist es wichtig, dass eine Sensitivitätsanalyse durchgeführt wird, um zu untersuchen, in welchem Maße die Ergebnisse von diesem Zufallsprozess abhängen. Die Ergebnisse derartiger Studien sollten über eine große Anzahl an Zufallsziehungen aggregiert werden. Dabei sollten Schwankungen im Ergebnis zwischen den wiederholten Zufallsziehungen ebenfalls berichtet werden. Angesichts der sinkenden Kosten für die Rechenleistung von Computern ist der zusätzliche Programmieraufwand meist das einzige Hindernis eine Analyse mehrfach durchzuführen. In zukünftigen Studien, die eine Evaluation der empirischen Optionsgewichtung zum Ziel haben, ist daher eine wiederholte randomisierte Substichproben-Validierung zu empfehlen.

Zusammenfassend legen die Ergebnisse der vorliegenden Arbeit nahe, dass Testentwickler zur Bestimmung von Optionsgewichten den empirischen Ansatz einer Befragung von Experten vorziehen sollten, um in Multiple-Choice-Tests Teilpunkte für Teilwissen zu vergeben. Empirische Optionsgewichte sind potenziell in der Lage, sowohl die Reliabilität als auch die Validität zu verbessern, wenn die Items über Distraktoren verfügen, die

zwischen unterschiedlichen Wissensständen differenzieren. Zukünftige Studien sollten bei der Evaluation der empirischen Optionsgewichtung eine Vielzahl von Stichprobenteilungen bei der Kreuzvalidierung berücksichtigen, um zuverlässige Ergebnisse zu erhalten und Forscherfreiheitsgrade zu vermeiden. Darüber hinaus sollten Signifikanztests eingesetzt werden, um Verbesserungen in der Testgüte zufallskritisch abzusichern. Die im Rahmen dieser Arbeit entwickelten Statistikpakete, cocor und cocron, ergänzen dabei die zur Verfügung stehenden Testverfahren um Tests zum Vergleich von Korrelationen bzw. Cronbach-α-Koeffizienten. Zukünftige Validierungen von Auswertungsverfahren sollten außerdem die Verwendung experimenteller Methoden, wie z. B. eine experimentelle Induktion von Wissensständen, als Verbesserung gegenüber korrelativer Ansätze in Betracht ziehen.

Literaturverzeichnis

Amelang, M., Schmidt-Atzert, L., Fydrich, T. & Zielinski, W. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Heidelberg, Germany: Springer-Verlag.

Bliese, P. (2013). multilevel: Multilevel Functions (Version 2.5). Zugriff 18. Dezember 2015, unter http://cran.R-project.org/package=multilevel

Cheung, M. W.-L. (2009). Constructing approximate confidence intervals for parameters with structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 267–294. doi:10.1080/10705510902751291

Cheung, M. W.-L. & Chan, W. (2004). Testing dependent correlations via structural equation modeling. *Organizational Research Methods*, *7*, 206–223. doi:10.1177/1094428104264024

Claudy, J. G. (1978). Biserial weights: A new approach to test item option weighting. *Applied Psychological Measurement*, *2*, 25–30. doi:10.1177/014662167800200102

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. doi:10.1037/0003-066X.49.12.997

Coombs, C. (1953). On the use of objective examinations. *Educational and Psychological Measurement*, *13*, 308–310. doi:10.1177/001316445301300214

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi:10.1007/BF02310555

Cross, L. H. & Frary, R. B. (1978). Empirical choice weighting under "guess" and "do not guess" directions. *Educational and Psychological Measurement*, *38*, 613–620. doi:10.1177/001316447803800302

Cross, L. H., Ross, F. K. & Geller, E. S. (1980). Using choice-weighted scoring of multiple-choice tests for determination of grades in college courses. *The Journal of Experimental Educational*, *48*, 296–301. doi:10.1080/00220973.1980.11011747

Davis, F. B. (1959). Estimation and use of scoring weights for each choice in multiple-choice test items. *Educational and Psychological Measurement*, *19*, 291–298. doi:10.1177/001316445901900301

Davis, F. B. & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, *19*, 159–170. doi:10.1177/001316445901900202

Diedenhofen, B. (2013). cocron: Statistical comparisons of two or more alpha coefficients (Version 1.0-0). Zugriff 18. Dezember 2015, unter http://comparingcronbachalphas.org/

Diedenhofen, B. & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*, e0121945. doi:10.1371/journal.pone.0121945

Downey, R. G. (1979). Item-option weighting of achievement tests: Comparative study of methods. *Applied Psychological Measurement*, *3*, 453–461. doi:10.1177/014662167900300403

Downing, S. M. & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Dressel, P. L. & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, *13*, 574–595. doi:10.1177/001316445301300404

Dunn, O. J. & Clark, V. A. (1971). Comparison of tests of the equality of dependent correlation coefficients. *Journal of the American Statistical Association*, *66*, 904–908. doi:10.2307/2284252

Echternacht, G. (1976). Reliability and validity of item option weighting schemes. *Educational and Psychological Measurement*, *36*, 301–309. doi:10.1177/001316447603600208

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 380–396. doi:10.1037//1082-989x.3.3.380

Feldt, L. S., Woodruff, D. J. & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*, 93–103. doi:10.1177/014662168701100107

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, *2*, 79–96. doi:10.1207/s15324818ame0201_5

Gilman, D. A. & Ferry, P. (1972). Increasing test reliability through self-scoring procedures. *Journal of Educational Measurement*, *9*, 205–207. doi:10.1111/j.1745-3984.1972.tb00953.x

Haladyna, T. M. (1990). Effects of empirical option weighting on estimating domain scores and making Pass/Fail decisions. *Applied Measurement in Education*, *3*, 231–244. doi:10.1207/s15324818ame0303_2

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3. Aufl.). Mahwah, NJ: Taylor & Francis.

Hambleton, R. K., Roberts, D. M. & Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement*, *7*, 75–82. doi:10.1111/j.1745-3984.1970.tb00698.x

Hendrickson, G. F. (1971). The effect of differential option weighting on multiple-choice objective tests. *Journal of Educational Measurement*, *8*, 291–296. doi:10.1111/j.1745-3984.1971.tb00941.x

Hittner, J. B. & May, K. (1998). DEPCORR: A SAS program for comparing dependent correlations. *Applied Psychological Measurement*, *22*, 93–94. doi:10.1177/014662169802210 10

Hittner, J. B., May, K. & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *The Journal of General Psychology*, *130*, 149–168. doi:10.1080/00221300309601282

Kansup, W. & Hakstian, A. R. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. *Journal of Educational Measurement*, *12*, 219–230. doi:10.1111/j.1745-3984.1975.tb01023.x

Levy, K. J. (1976). A multiple range procedure for independent correlations. *Educational and Psychological Measurement*, *36*, 27–31. doi:10.1177/001316447603600103

Liddell, F. D. K. (1983). Simplified exact analysis of case-referent studies: Matched pairs; dichotomous exposure. *Journal of Epidemiology and Community Health*, *37*, 82–84. doi:10.1136/jech.37.1.82

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, *12*, 7–11. doi:10.1111/j.1745-3984.1975.tb01003.x

Lukas, J. (1997). Modellierung von Fehlkonzepten in einer algebraischen Wissensstruktur. *Kognitionswissenschaft*, *6*, 196–204. doi:10.1007/s001970050042

May, K. & Hittner, J. B. (1997). Tests for comparing dependent correlations revisited: A Monte Carlo study. *The Journal of Experimental Education*, *65*, 257–269. doi:10.1080/00220973.1997.9943458

Neill, J. J. & Dunn, O. J. (1975). Equality of dependent correlation coefficients. *Biometrics*, *31*, 531–543. doi:10.2307/2529435

Nieuwenhuis, S., Forstmann, B. U. & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105–1107. doi:10.1038/nn.2886

Olkin, I. & Finn, J. D. (1995). Correlation redux. *Psychological Bulletin*, *118*, 155–164. doi:10.1037/0033-2909.118.1.155

Patnaik, D. & Traub, R. E. (1973). Differential weighting by judged degree of correctness. *Journal of Educational Measurement*, *10*, 281–286. doi:10.1111/j.1745-3984.1973.tb00805.x

Paul, S. R. (1989). A multiple range procedure for independent correlations. *Canadian Journal of Statistics*, *17*, 217–227. doi:10.2307/3314850

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879–903. doi:10.1037/0021-9010.88.5.879

Poizner, S. B., Nicewander, W. A. & Gettys, C. F. (1978). Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. *Applied Psychological Measurement*, *2*, 83–96. doi:10.1177/014662167800200109

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Zugriff 18. Dezember 2015, unter http://www.R-project.org/

Raffeld, P. (1975). The effects of Guttman weights on the reliability and predictive validity of objective tests when omissions are not differentially weighted. *Journal of Educational Measurement*, *12*, 179–185. doi:10.1111/j.1745-3984.1975.tb01020.x

Raghunathan, T. E., Rosenthal, R. & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, *1*, 178–183. doi:10.1037//1082-989X.1.2.178

Reilly, R. R. & Jackson, R. (1973). Effects of empirical option weighting on reliability and validity of an academic aptitude test. *Journal of Educational Measurement*, *10*, 185–193. doi:10.1111/j.1745-3984.1973.tb00796.x

Revelle, W. (2014). psych: Procedures for psychological, psychometric, and personality research (Version 1.4.8). Zugriff 21. Februar 2015, unter http://cran.R-project.org/package=psych

Rödiger, S., Friedrichsmeier, T., Kapat, P. & Michalke, M. (2012). RKWard: A comprehensive graphical user interface and integrated development environment for statistical analysis with R. *Journal of Statistical Software*, *49*, 1–34. Zugriff 18. Dezember 2015, unter http://www.jstatsoft.org/v49/i09

Rousselet, G. A. & Pernet, C. R. (2012). Improving standards in brain-behavior correlation analyses. *Frontiers in Human Neuroscience*, *6*, 1–11. doi:10.3389/fnhum.2012.00119

Sabers, D. L. & White, G. W. (1969). The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement*, *6*, 93–96. doi:10.1111/j.1745-3984.1969.tb00664.x

Silver, N. C., Hittner, J. B. & May, K. (2004). Testing dependent correlations with nonoverlapping variables: A Monte Carlo simulation. *The Journal of Experimental Education*, *73*, 53–69. doi:10.3200/JEXE.71.1.53-70

Silver, N. C., Hittner, J. B. & May, K. (2006). A FORTRAN 77 program for comparing dependent correlations. *Applied Psychological Measurement*, *30*, 152–153. doi:10.1177/0146621605277132

Silver, N. C., Zaikina, H., Hittner, J. B. & May, K. (2008). INCOR: A computer program for testing differences among independent correlations. *Molecular Ecology Resources*, *8*, 763–764. doi:10.1111/j.1755-0998.2008.02107.x

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Stanley, J. C. & Wang, M. W. (1970). Weighting test items and test-item options, an overview of the analytical and empirical literature. *Educational and Psychological Measurement*, *30*, 21–35. doi:10.1177/001316447003000102

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251. doi:10.1037/0033-2909.87.2.245

Wang, M. W. & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, *40*, 663–705. doi:10.3102/00346543040005663

Waters, B. K. (1976). The measurement of partial knowledge: A comparison between two empirical option-weighting methods and rights-only scoring. *The Journal of Educational Research*, *69*, 256–260. doi:10.1080/00220671.1976.10884892

Weaver, B. & Wuensch, K. L. (2013). SPSS and SAS programs for comparing pearson correlations and OLS regression coefficients. *Behavior Research Methods*, *45*, 880–895. doi:10.3758/s13428-012-0289-7

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*, 399–413. doi:10.1037/1082-989X.12.4.399

Anhang: Artikel

Artikel 1

Diedenhofen, B. & Musch, J. (2015). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment*. Advance online publication. doi: 10.1027/1015-5759/a000295

Artikel 2

Diedenhofen, B. & Musch, J. (2015). Option weights should be determined empirically and not by experts when assessing knowledge using multiple-choice items. Manuscript submitted for publication.

Artikel 3

Diedenhofen, B. & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*(4), e0121945. doi: 10.1371/journal.pone.0121945

Empirical option weights improve the validity of a multiple-choice knowledge test

Birk Diedenhofen & Jochen Musch

University of Duesseldorf

Author Note

Article type: Original article, Word count: 5402

Birk Diedenhofen & Jochen Musch, Department of Experimental Psychology, University of Duesseldorf.

Correspondence concerning this article should be addressed to Birk Diedenhofen, Department of Experimental Psychology, University of Duesseldorf, Universitaetsstrasse 1, Building 23.03, 40225 Duesseldorf, Germany. Tel.: +49 211 81-12065, Fax: +49 211 81-11753, E-mail: birk.diedenhofen@uni-duesseldorf.de

Table 1: An example item of the Arawaken knowledge test with the points awarded by number-right scoring (NR) and by empirical option weighting (EOW) for each answer option in samples 1 and 2, and the knowledge that is necessary to exclude the respective option from the set of potential solutions.

Table 2: Item difficulty, item discriminatory power, and empirical option weights for each answer option of all 7 experimental items (sorted by difficulty) separately for samples 1 and 2.

Table 3: Mean (M) and standard deviation (SD) of the difficulty and discriminatory power of the seven knowledge test items for samples 1 and 2, and the mean scores achieved in the three knowledge conditions under number-right scoring.

Table 4: Mean (M), standard deviation (SD), minimum (min), maximum (max), and Cronbach's coefficient α of the scores achieved by participants in the two samples under number-right scoring and empirical option weighting.

Table 5: Number of correct and incorrect classifications of participants to their known knowledge conditions based on a one-dimensional linear discriminant analysis under number-right scoring (NR) and empirical option weighting (EOW).

Abstract

Standard dichotomous scoring of multiple-choice test items grants no partial credit for partial knowledge. Empirical option weighting is an alternative, polychotomous scoring method that uses the point-biserial correlation between option choices and total score as a weight for each answer alternative. Extant studies demonstrate that the method increases reliability of multiple-choice tests in comparison to conventional scoring. Most previous studies employed a correlational validation approach, however, and provided mixed findings with regard to the validity of empirical option weighting. The present study is the first investigation using an experimental approach to determine the reliability and validity of empirical option weighting. To obtain an external validation criterion, we experimentally induced various degrees of knowledge in a domain of which participants had no knowledge. We found that in comparison to dichotomous scoring, empirical option weighting increased both reliability and validity of a multiple-choice knowledge test employing distractors that were appealing to test takers with different levels of knowledge. A potential application of the present results is the computation and publication of empirical option weights for existing multiple-choice knowledge tests that have previously been scored dichotomously.

Keywords: Empirical option weighting, multiple-choice test, reliability, validity, partial knowledge

Empirical option weights improve the validity of a multiple-choice knowledge test

Multiple-choice testing is applied broadly to assessment of knowledge in education and personnel selection contexts (Haladyna, 2004). Tests from various domains that employ the multiple-choice answer format have been shown repeatedly to possess satisfactory psychometric properties. *Number-right scoring* (Frary, 1989) has usually been used as the standard to score multiple-choice items. Under the method, test takers receive one point for a correct answer and zero points for choosing a distractor. One important aspect of this dichotomous scoring is that all items receive the same weight when calculating the total score. Number-right scoring rests on the implicit assumption that test takers failing to identify the correct solution should be rewarded with the same score regardless of what distractors they chose. However, test takers who did not identify the correct answer frequently did not choose randomly among available options; their choice of distractor often reflects partial knowledge (Lau, Lau, Hong, & Usop, 2011). Therefore, test takers possessing partial knowledge might demonstrate preference for a distractor not shared by test takers without knowledge. Dichotomous number-right scoring does not distinguish distractor choices, and therefore cannot grant partial credit for partial knowledge.

One way to reward partial knowledge is to score tests polychotomously by assigning weights to each answer option (Frary, 1989). The total score can then be calculated by summing the weights of all options selected. For polychotomous scoring procedures, weights of answer options are proportionate to the amount of partial knowledge reflected by choosing an option.

Various procedures have been proposed to determine weights. Generally, *a priori weighting* and *empirical option weighting* are the two primary methods (Stanley & Wang, 1970). Experts usually assign a priori weights, and quantify partial knowledge reflected by choices. Number-right scoring is a special case of a priori weighting that awards 1 point for choosing the correct answer, and no points for choosing a distractor. Some studies obtained better results for empirical approaches than for a priori weighting (Frary, 1989). The present study focuses on the empirical option weighting approach in

which empirical data, rather than the opinions of experts, decide weights, and thus partial credit awarded for answer alternatives.

Total test scores of participants opting for a particular response alternative are usually used as an empirical criterion to calculate empirical option weights (Frary, 1989). Empirical option weighting is then employed by awarding more (fewer) points for the choice of an option weight that is popular among high (low) scorers. Two ways of computing empirical option weights are available: One called the *method of reciprocal averages* awards each test taker with a score achieved by other test takers having decided to choose the same answer option (Echternacht, 1976; Haladyna, 1990). The alternative is to use option-total correlations as empirical option weights (Claudy, 1978; Haladyna, 1990). To employ this approach, a point-biserial correlation is computed for each answer option (1 if chosen, zero if not chosen) and the test score for the entire test under number-right scoring. Even if a test taker does not choose the correct answer, he or she might receive partial credit to the extent his or her answer is popular among better-scoring participants. For both approaches, a part-whole correction is typically performed. Claudy (1978) compared empirical option weighting based on the method of reciprocal averages and option-total correlations. His results suggest the two procedures offer comparable results.

Some authors of early studies of empirical option weighting discuss critically whether potential gains from using empirical option weighting outweigh the higher effort associated with computing empirical option weights (Raffeld, 1975; Wang & Stanley, 1970). The additional effort appears increasingly less troublesome in view of current and widespread availability of electronic testing environments that offer automated scoring of both simple number-right scores and empirically determined weights. If a test is scored electronically, there is little disparity in effort between number-right scoring and empirical option weighting. Consequently, whether empirically determined option weights improve test validity deserves closer examination.

Most studies that scrutinize empirical option weights found them to improve test reliability (Claudy, 1978; Cross & Frary, 1978; Cross, Ross, & Geller, 1980; Downey, 1979; Echternacht, 1976; Haladyna,

1990; Hendrickson, 1971; Raffeld, 1975; Reilly & Jackson, 1973; Waters, 1976). Sabers and White (1969) reported inconsistent results when comparing the reliability of test scores based on empirical option weighting and number-right scoring. They found an increase in reliability for three of the tests and a decrease for one of the tests they investigated. However, none of the observed differences in reliability were tested for significance. Regarding potential improvement of validity, however, findings are less conclusive and contradictory. Despite validity being the more important objective of testing, few studies examine validity in addition to reliability. Two studies by Reilly and Jackson (1973) and Downey (1979) observed a detrimental effect of empirical option weighting, finding it to lower validity in comparison to conventional scoring. This descriptive decrease, however, was not tested for significance. Cross and Frary (1978), Echternacht (1976), Raffeld (1975), and Haladyna (1990) found empirical option weighting to increase validity in comparison to number-right scoring, but again, the increase was not tested for significance. Several studies found no difference in validity between empirical option weighting and number-right scoring (Cross et al., 1980; Davis & Fifer, 1959; Sabers & White, 1969; Waters, 1976). Without exception, however, all studies investigate convergent or predictive validities of empirical option weighting using a correlational design. Therefore, a test taker's true knowledge was unknown and could never be used as an external criterion. Instead, validity was assessed using criteria based on expert judgments or scores from other tests. Both of these validity criteria are imperfect proxies of a test taker's true knowledge, however. Whether they provide reliable assessment of the validity of empirical option weighting is questionable. A high correlation between two tests need not be due to measurement of the same construct; two tests might correlate only because they share proportions of method variance. This alternative explanation for high external validity is troublesome because most validation studies use the same test format for both predictors and criteria.

Mixed results regarding the validity of empirical option weighting might indicate biased test construction. When comparing polychotomous and dichotomous scorings, it is important to consider the scoring procedure for which the test was constructed. Many tests used in extant investigations were

devised for traditional number-right scoring. Empirical option weighting works best, however, when multiple-choice answer options are designed such that each option appeals to test takers of varying abilities (Raffeld, 1975; Haladyna, 1990). Only two studies apply empirical option weighting to a test constructed with this peculiarity of polychotomous scoring in mind (Davis & Fifer, 1959; Echternacht, 1976). Davis and Fifer (1959) found that without reducing concurrent test validity, empirically weighted scores correlate more closely with a parallel test version than dichotomous scores did. However, Davis and Fifer employ an unusual option weighting approach that used test scores based on expert judgments as the criterion when computing empirical option weights. Weights were modified so points awarded for a correct answer always exceeded those awarded for distractors. The authors use a simplified method to calculate option weights that only takes data of the upper and lower 27% of participants into account. In a second study designed so each option appealed to test takers of a different ability, Echternacht (1976) descriptively found improvement in reliability and validity through empirical option weighting in comparison to conventional scoring. Unfortunately, Echternacht did not report a test of statistical significance. The available body of research, combined with the fact that extant studies apply empirical option weighting to tests constructed for number-right scoring, precludes strong conclusions regarding whether empirical option weighting enhances test validity. To overcome the shortcomings of previous investigations, we used an experiment to determine the validity of empirical option weighting.

The present study investigates the possibilities of empirical weighting for a specific type of multiple-choice items characterized by distractors differentially appealing to test takers with different levels of expertise. To answer whether empirical option weighting improves the validity of multiple-choice tests, we compared empirical option weighting and number-right scoring as alternative scoring procedures for a new-knowledge test we constructed. Unlike extant studies and to overcome the drawbacks of correlational validation procedures, we conducted an experiment based on the experimental induction of various degrees of knowledge of varying groups of participants (cf. Poizner,

Nicewander, & Gettys, 1978). This approach has never been used for assessment of the validity of empirical option weighting, but offers the advantage of establishing a benchmark to assess external validity, namely, the true level of the test takers' knowledge. To obtain complete control over the test takers' knowledge, we constructed short essays and a knowledge test for a domain with which all test takers were unfamiliar. We were thus able to provide varying amounts of knowledge to groups of participants at will. Participants first read an essay which contained either no, partial, or full information and then completed the knowledge test. We scored the test responses twice using number-right scoring and empirical option weighting, and compared the scores of both procedures with regard to reliability and validity. We expected empirical option weighting to provide an improved diagnostic precision by producing test scores that allow us to determine the knowledge of participants more accurately compared to number-right scoring.

Methods

Design

The present experiment had a 3×2 factorial design with the between-subjects factor knowledge level (no, partial, and complete knowledge), the within-subjects factor scoring procedure (number-right scoring vs. empirical option weighting), and reliability and validity of the knowledge test as dependent variables. We manipulated the amount of knowledge (no, partial, or complete knowledge) gained form a short essay we asked test takers to read. One-third of participants received no information that was useful for the knowledge test presented later. Another third received sufficient information to exclude one of three answer alternatives during the subsequent test; their chances of guessing the correct solution was increased from 33% to 50%. The remaining third received enough information to exclude two of three answer alternatives, allowing them to identify the correct solution (in theory) to each item, under the assumption of perfect memory. After all test takers completed the tests, we scored the test takers' responses twice using either number-right scoring or empirical option weighting. We then compared the reliability and validity of both scoring procedures. As a measure of reliability, we calculated Cronbach's coefficient α. To assess validity, we computed two indices. First, we conducted an analysis of variance with test scores as dependent variables and participants' knowledge as independent variables, allowing us to determine whether experimentally controlled knowledge accounted for more of a test score's variance when scores were calculated under empirical option weighting than when calculated under number-right scoring. As an additional measure of validity, we determined the hit rate by which one-dimensional linear discriminant analysis classified participants correctly regarding their knowledge based on scores achieved on the test. We used the R statistical computing environment (R Core Team, 2014) for all of our analyses.

Material

Participants were assigned randomly to one of three groups, differing by knowledge. To induce varying knowledge, participants were provided with short essays containing no, partial, or all information required to answer questions on the subsequent knowledge test. The essays described rare and obscure details of the history, life, and culture of the Arawaken, a Native American tribe we found unknown to each participant during a pretest we conducted with a German student sample consisting of 120 participants who attended a psychology lecture. By choosing this topic, we ensured participants did not possess knowledge that would interfere with knowledge manipulation.

The knowledge test consisted of 10 three-option multiple-choice items. Seven items constituted experimental items used for subsequent analysis. The items were constructed so that participants with no knowledge were unable to exclude distractors based on information they received from the essay. Consequently, we expected participants in this condition to be forced to select answer options randomly, resulting in a probability of 33% of choosing the correct answer. Participants in the group provided with partial knowledge read an essay that enabled them to exclude one of two distractors in each question. We expected the probability of correctly solving an item—assuming perfect memory—to be 50%. Participants in the group provided with complete knowledge were presented with a detailed essay that allowed them to answer all items correctly, and thus achieve a perfect score under the assumption of perfect memory. Table 1 shows an example item (Item 2) from the knowledge test, demonstrating how essays that differed by the information they contained resulted in varying knowledge that allowed participants to be successful, moderately successful, or unsuccessful. Item 2 was chosen as an illustration because of its briefness. Other items of the test were not based on such additive word lists; what all items of the test did have in common, however, is that they established a hierarchy that made different distractors appealing to respondents differing in their level of knowledge. This was achieved by creating distractors based on varying numbers of propositions that either were or were not included in the texts that were presented to induce different levels of knowledge among test

takers. In addition to the seven experimental items, we added three filler items for which essays in all conditions contained sufficient information to identify correct answers. These items were not used during analyses, but were useful during testing to spare participants in the no-knowledge group from frustration of not being able to answer any questions.

Sample

Participants were recruited by e-mail through invitations sent to members of an online panel consisting of participants of previous studies conducted in the Department of Experimental Psychology at the University of Düsseldorf. No member of the panel had participated previously in a test containing any of the materials used during the present investigation. After exclusion of participants who indicated a native language other than German (2), did not complete the test (9), or took part repeatedly using the same IP address (15) (Aust, Diedenhofen, Ullrich, & Musch, 2013), 567 cases were available for analysis. The average age of participants (55% female) was 36.24 years (SD = 12.59). The sample was highly educated, with 26% of participants reporting a high school degree and 62% indicating a college degree as their highest educational attainment; only 12% reported holding no high school degree. Random assignment of participants to conditions resulted in 196, 191, and 180 participants in the no, partial, and complete knowledge groups, respectively.

Procedure

The study was conducted online using Unipark EFS Survey 9.1 software (QuestBack, 2013). The first page of the survey welcomed participants and obtained their informed consent to participate. On the second page, participants received instructions to read the following essay carefully because they would later be questioned regarding its content. The following page presented one of three essays, differing concerning the information they contained depending on the knowledge group to which a participant had been assigned. After reading the essay, participants were instructed to choose for all items in the subsequent multiple-choice test the answer they found most plausible. On the following pages, the knowledge test was administered, with each of the 10 items presented on a separate page.

The order of the items and the order of all answer options within each item was randomized. After completing the knowledge test, participants indicated their ages, genders, first languages, and their highest education degrees. On the last two pages, we thanked and debriefed the participants, and provided them with feedback regarding performance.

Scoring

Responses were scored twice, once using number-right scoring and once using empirical option weighting. When scored dichotomously, a test taker received 1 point for selecting the correct answer and zero points for choosing a distractor. The total test score under number-right scoring was computed as the number of items answered correctly. To determine the empirical option weights, we computed the option-total correlation for all items and all answer alternatives (Echternacht, 1976; Haladyna, 1990). Calculation of option-total correlations was simple, and their use as empirical option weights provided results comparable to the method of reciprocal averages (Claudy, 1978). Thus, weights for all options were calculated as the point-biserial correlation between each answer option (1 if chosen, zero if not chosen) and the total test score under number-right scoring. A part-whole correction was applied for the computation of all option weights. The total test score under empirical option weighting was determined by summing the weights of all options a participant selected. To avoid capitalization on chance and to cross-validate results (Stanley & Wang, 1970), the sample was split randomly into two subsamples of 284 and 283 participants. Empirical option weights determined using the first half of participants (sample 1) were used to score responses of the second half (sample 2), and vice versa. To ensure that both samples contained equal numbers of participants concerning each knowledge condition, participants were split in halves for each knowledge group. Table 1 displays the empirical options weights that were awarded for each answer option of item 2.

Results

For both samples, all empirical option weights are detailed in Table 2. As expected and in all items, the correct answer received the highest option weight (on average, $M = .48$, which is the average option-total correlation for the correct answer). The answer option that could be excluded based on complete information but not on partial information received the second highest option weight ($M = -.18$). The distractor that could be excluded by participants in the partial and full-knowledge conditions obtained the lowest option weight ($M = -.38$), with the exception of item 7 in sample 2, for which the option weights for the two distractors were in reverse order ($M = -.25$ vs. $M = -.21$). However, since item 7 is the easiest item of the test, relatively few participants chose one of the two distractors for this item ($n = 55$ and $n = 22$, respectively). The reverse order of the empirical option weights for the two distractors was therefore probably the result of sampling error. Thus, the order of option weights for the answer alternatives followed the expectation nearly without exception.

Table 3 shows scores achieved by participants in the three knowledge groups and expected scores—assuming perfect memory—under number right-scoring. The differences in the mean number of correctly answered experimental items in the three groups suggest that the knowledge manipulation was successful. As expected, participants in the partial-knowledge group answered more items correctly than participants in the no-knowledge group, both for sample 1 ($t(192) = 7.65$, $p < .001$, $d = 1.10$, two-tailed) and sample 2 ($t(191) = 8.04$, $p < .001$, $d = 1.16$, two-tailed). Similarly, participants in the complete-knowledge group answered more items correctly than participants in the partial-knowledge group, both for sample 1 ($t(184) = 21.92$, $p < .001$, $d = 3.22$, two-tailed) and sample 2 ($t(183) = 18.95$, $p < .001$, $d = 2.79$, two-tailed). As indicated by generalized eta-squared (η_g^2) as a measure of effect size, scores achieved on the filler items under number-right scoring did not differ among the three knowledge conditions for sample 1 ($F(2, 281) = 0.41$, $p = .663$, $\eta_g^2 < .01$) and sample 2 ($F(2, 280) = 0.84$, $p = .435$, $\eta_g^2 < .01$). Averaged across the three conditions, the mean item difficulty

was .52, and the average discriminatory power of the items was .49 (Table 3). Scores achieved under number-right scoring and empirical option weighting are shown in Table 4.

For both samples, we compared Cronbach's coefficient α based on scores that were calculated under empirical option weighting or under number-right scoring. We used the significance test for dependent α coefficients by Feldt, Woodruff, and Salih (1987) as implemented in the R package *cocron* (version 1.0-0; Diedenhofen, 2013). Empirical option weighting demonstrated higher reliability than number-right scoring for sample 1 ($\alpha_{NR} = .758$, $\alpha_{EOW} = .782$, $\chi^2(1) = 28.92$, $p < .001$) and sample 2 ($\alpha_{NR} = .766$, $\alpha_{EOW} = .790$, $\chi^2(1) = 38.25$, $p < .001$). We computed the *h* statistic derived from Spearman-Brown's formula (Haladyna, 1990, p. 236) to determine by which factor the test would have to be lengthened under number-right scoring to achieve the same α as under empirical option weighting (Table 4).

We calculated analyses of variance for each scoring procedure and sample with test scores as dependent variables and knowledge conditions as independent variables. For sample 1, results suggest scores obtained from empirical option weighting explained more variance of the criterion (generalized η^2; $F(2, 281) = 646.15$, $p < .001$, $\eta_g^2 = .82$) than scores obtained from number-right scoring ($F(2, 281) = 475.20$, $p < .001$, $\eta_g^2 = .77$). The same pattern was observed for sample 2. Again, the proportion of explained variance was numerically higher for empirical option weighting ($F(2, 280) = 474.46$, $p < .001$, $\eta_g^2 = .77$) than for traditional number-right scoring ($F(2, 280) = 384.33$, $p < .001$, $\eta_g^2 = .73$). For both calculation methods, we reassigned participants to their original knowledge conditions using one-dimensional linear discriminant analysis based on respective test scores. Table 5 shows the resulting number of correct and incorrect classifications separately for both samples and both calculation methods. Liddell tests for dependent proportions (Liddell, 1983) revealed that hit rates were higher for empirical option weighting for both samples 1 ($Hit\ rate_{NR} = 78.52\%$, $Hit\ rate_{EOW} = 87.32\%$, $F(4, 52) = 13.00$, $p < .001$) and 2 ($Hit\ rate_{NR} = 74.56\%$, $Hit\ rate_{EOW} = 84.10\%$, $F(6, 58) = 9.67$, $p < .001$).

Discussion

The results suggest that empirical option weighting improves reliability and validity of a knowledge test in comparison to conventional number-right scoring. Using the *h* statistic as an indicator, improvement in internal consistency reliability obtained using empirical option weighting was equivalent to increasing the length of the dichotomously scored test by 15%. More importantly, two independent indices showed that empirical option weighting increases validity of a multiple-choice knowledge test. First, analyses of variance with test scores as dependent variables and knowledge conditions as independent variables indicated that the knowledge condition explains more test-score variance when scores are calculated using empirical option weighting. In a personnel selection context, the amount of additionally explained variance can be illustrated using the dichotomous categories of the binomial effect size display proposed by Rosenthal and Rubin (1982). According to this display, an increase of explained variance of 4% is equivalent to a raise in the success rate of selecting suitable candidates from 40% to 60%. Second, the rate of correct classifications obtained using one-dimensional linear discriminant analysis to classify participants according to knowledge condition was higher for empirical option weighting. Thus, both indices of validity suggest degrees of knowledge reflect test scores better when computed using empirical option weights.

Extant studies found no improved validity for empirical option weighting (e.g., Downey, 1979; Cross et al., 1980). One explanation is that empirical option weighting differs most prominently from number-right scoring regarding difficult items in which test takers frequently choose distractors. If test takers identify the answer, empirical option weighting and number-right scores converge. Indeed, Haladyna (1990) found empirical option weighting to be most effective in the lower half of the ability distribution because less able participants more frequently chose distractors for which empirical option weighting provides a finer scoring scheme (cf., Davis & Fifer, 1959; Hambleton, Roberts, & Traub, 1970; Haladyna, 1990). The high difficulty of the present test items, which on average were answered correctly by only 52% of participants, might have contributed to increased reliability and validity of

empirical option weighting. In studies by Downey (1979) and Cross et al. (1980), lower difficulty of items under investigation might have prevented empirical option weighting from providing additional information beyond that contained by number-right scores. In a study by Downey (1979), items were solved by 67% of participants on average, and in the study by Cross et al. (1980), items were solved by 68% of participants, averaged across all items and tests. In both of these studies, items were considerably easier than in the present investigation.

Our present finding that empirically derived option weights yield improvement over conventional number-right scoring was obtained using experimental validation. Another explanation for conflicting results obtained in extant studies is that they might have used items that were less capable of differentiating degrees of partial knowledge. The items used in the present study were constructed so that each answer option directly reflected one of the three different levels of knowledge. Under such circumstances, empirical option weighting is most likely to add helpful additional information that cannot be obtained using dichotomous scoring. The present study demonstrates that empirical option weighting can improve validity when distractors are gradated in a way that reflects different levels of knowledge. Generalizability to other multiple-choice tests is likely to be highest, and may be limited to, tests also employing distractors that are differentially attractive to test takers with different levels of expertise. Finally, we referred to the present test as a knowledge test. However, given that participants had to learn and to recall the material immediately after the learning phase, our test could justifiably also be referred to as a memory test.

As an index of how well the distractors of an item helped to distinguish between different levels of knowledge, we computed the difference between the weight for the correct solution and the best distractor (weight for answer option 1 minus weight for answer option 2, see Table 2), and the difference between the weight for the best distractor and the second best distractor (weight for answer option 2 minus weight for answer option 3). The sum of these two differences indicates how well the three available answer options represent different levels of knowledge. As expected, easy items that

were solved by most participants were less well suited to distinguish between different levels of knowledge than difficult items; across both samples, there was a strong negative correlation between the above gradation index and item difficulty ($r = -.57$, $p < .04$). Importantly, the more difficult items were also the ones that led to the observed increase in validity for empirical option weighting. Averaged across both samples, the hit rates obtained on the basis of the four difficult items improved from 70.73% to 80.43% when using empirical option weighting instead of number-right scoring. On the basis of the three easy items, there was no such increase in validity, and hit rates were identical (63.14%) for number-right scoring and empirical option weighting, respectively.

For multiple-choice items, it has been argued that many participants who cannot directly identify the correct solution to an item are trying to exclude distractors to maximize their chance of selecting the correct answer (Carpenter, Just, & Shell, 1990; Frary, 1982; Wilcox, 1981). They usually do so by excluding distractors in the order of their subjective implausibility. If distractor options differ with regard to how easily they can be dismissed based on partial or incomplete knowledge, the specific distractor that is chosen by a test taker offers valuable information in addition to the fact that the test taker was unable to identify the correct solution. One means to take advantage of this additional information is to count the number of attempts a test taker needs to arrive at the solution. This count is used as the participants' score in what is called the answer-until-correct procedure (Gilman & Ferry, 1972). As the present study shows, empirical option weighting can be used in a similar way as the answer-until-correct procedure to gain additional information from multiple-choice test items, but without having to change the answer format. To profit from the information contained in distractor choice, it is, however, important that items have the same properties that are also necessary for a successful implementation of the answer-until-correct format (Kane & Moloney, 1978). In particular, answer options should follow a knowledge hierarchy to make sure that the more partial knowledge a participant has available, the more distractors he or she can reject. To this end, the creator of a test first has to make sure that test items are difficult enough so that a sufficient number of participants fail to

identify the solution. If an answer option is chosen by either everybody or no one, there is no opportunity to gather information that goes beyond whether a test taker was able to solve the item or not. As in the cognitive design modeling approach proposed by Embretson (1998), distractor options should then be created in a way that allows to identify partial knowledge or misconceptions that discriminate between test takers who were not able to arrive at the correct solution for some reason. If distractor options are modeled to reflect various potential states within a knowledge structure, the choice of a particular distractor offers information that goes beyond the simple dichotomous scoring of answers as being either correct or wrong. Distractors offering blatant misconceptions, for example, offer an efficient way to identify test takers who have very little knowledge in a particular domain (Lukas, 1997).

Based on our results, we recommend that future studies using empirical option weighting should use test items that contain discriminating distractors that offer reliable assessment beyond one degree of partial knowledge. Future studies should also use experimental validation like the one used in the present study. In view of ubiquitous computerization of testing environments, the additional effort associated with employing advanced scoring is no longer a serious obstacle to using what we suggest is a superior scoring method regarding both reliability and validity. A potentially useful application of present results is computation and publication of empirical option weights for multiple-choice knowledge tests that have been published but were scored dichotomously. The results suggest that multiple-choice knowledge tests benefit from sophisticated scoring based on empirical option weighting.

References

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavioral Research Methods, 45*, 527–535. doi:10.3758/s13428-012-0265-2

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. *Psychological Review, 97*, 404–431. doi:10.1037/0033-295x.97.3.404

Claudy, J. G. (1978). Biserial weights: A new approach to test item option weighting. *Applied Psychological Measurement, 2*, 25–30. doi:10.1177/014662167800200102

Cross, L. H. & Frary, R. B. (1978). Empirical choice weighting under "guess" and "do not guess" directions. *Educational and Psychological Measurement, 38*, 613–620. doi:10.1177/001316447803800302

Cross, L. H., Ross, F. K., & Geller, E. S. (1980). Using choice-weighted scoring of multiple-choice tests for determination of grades in college courses. *The Journal of Experimental Educational, 48*, 296–301. doi:10.1080/00220973.1980.11011747

Davis, F. B. & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement, 19*, 159–170. doi:10.1177/001316445901900202

Diedenhofen, B. (2013). *cocron: Statistical comparisons of two or more alpha coefficients*. (Version 1.0-0). Retrieved from http://comparingcronbachalphas.org/

Downey, R. G. (1979). Item-option weighting of achievement tests: Comparative study of methods. *Applied Psychological Measurement, 3*, 453–461. doi:10.1177/014662167900300403

Echternacht, G. (1976). Reliability and validity of item option weighting schemes. *Educational and Psychological Measurement, 36*, 301–309. doi:10.1177/001316447603600208

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380–396. doi:10.1037//1082-989x.3.3.380

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*, 93–103. doi:10.1177/014662168701100107

Frary, R. B. (1982). A simulation study of reliability and validity of multiple-choice test scores under six response-scoring modes. *Journal of Educational and Behavioral Statistics, 7*, 333–351. doi:10.3102/10769986007004333

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2*, 79–96. doi:10.1207/s15324818ame0201_5

Gilman, D. A. & Ferry, P. (1972). Increasing test reliability through self-scoring procedures. *Journal of Educational Measurement, 9*, 205–207. doi:10.1111/j.1745-3984.1972.tb00953.x

Haladyna, T. M. (1990). Effects of empirical option weighting on estimating domain scores and making Pass/Fail decisions. *Applied Measurement in Education, 3*, 231–244. doi:10.1207/s15324818ame0303_2

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Taylor & Francis.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement, 7*, 75–82. doi:10.1111/j.1745-3984.1970.tb00698.x

Hendrickson, G. F. (1971). The effect of differential option weighting on multiple-choice objective tests. *Journal of Educational Measurement, 8*, 291–296. doi:10.1111/j.1745-3984.1971.tb00941.x

Kane, M. & Moloney, J. (1978). The effect of guessing on item reliability under answer-until-correct scoring. *Applied Psychological Measurement, 2*, 41–49. doi:10.1177/014662167800200104

Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Educational Technology & Society, 14*, 99–110. doi:10.1037/e683312011-118

Liddell, F. D. K. (1983). Simplified exact analysis of case-referent studies: Matched pairs; dichotomous exposure. *Journal of Epidemiology and Community Health, 37*, 82–84. doi:10.1136/jech.37.1.82

Lukas, J. (1997). Modellierung von Fehlkonzepten in einer algebraischen Wissensstruktur [Modeling misconceptions in an algebraic knowledge structure]. *Kognitionswissenschaft, 6*, 196–204. doi:10.1007/s001970050042

Poizner, S. B., Nicewander, W. A., & Gettys, C. F. (1978). Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. *Applied Psychological Measurement, 2*, 83–96. doi:10.1177/014662167800200109

QuestBack. (2013). Unipark EFS Survey 9.1. Retrieved from http://www.unipark.de

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Raffeld, P. (1975). The effects of Guttman weights on the reliability and predictive validity of objective tests when omissions are not differentially weighted. *Journal of Educational Measurement, 12*, 179–185. doi:10.1111/j.1745-3984.1975.tb01020.x

Reilly, R. R. & Jackson, R. (1973). Effects of empirical option weighting on reliability and validity of an academic aptitude test. *Journal of Educational Measurement, 10*, 185–193. doi:10.1111/j.1745-3984.1973.tb00796.x

Rosenthal, R. & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166–169. doi:10.1037/0022-0663.74.2.166

Sabers, D. L. & White, G. W. (1969). The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement, 6*, 93–96. doi:10.1111/j.1745-3984.1969.tb00664.x

Stanley, J. C. & Wang, M. W. (1970). Weighting test items and test-item options, an overview of the analytical and empirical literature. *Educational and Psychological Measurement, 30*, 21–35. doi:10.1177/001316447003000102

Wang, M. W. & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40*, 663–705. doi:10.3102/00346543040005663

Waters, B. K. (1976). The measurement of partial knowledge: A comparison between two empirical option-weighting methods and rights-only scoring. *The Journal of Educational Research, 69*, 256–260. doi:10.1080/00220671.1976.10884892

Wilcox, R. R. (1981). Solving measurement problems with an answer-until-correct scoring procedure. *Applied Psychological Measurement, 5*, 399–414. doi:10.1177/014662168100500313

Table 1

An example item of the Arawaken knowledge test with the points awarded by number-right scoring (NR) and by empirical option weighting (EOW) for each answer option in samples 1 and 2, and the knowledge that is necessary to exclude the respective option from the set of potential solutions.

		EOW		
Multiple-choice options	**NR**	**Sample 1**	**Sample 2**	**Knowledge required for exclusion**
Potatoes, sweet potatoes, beans	0	−.38	−.35	Partial knowledge or complete knowledge
Sweet potatoes, tomatoes, cocoa	0	−.22	−.18	Complete knowledge
Sweet potatoes, beans, peanuts	1	.50	.42	(Correct answer, no exclusion possible)

Notes. The table lists the three answer options to the question "Which foods were known to the Arawaken?" The no-knowledge group received no information about the foods known to the Arawaken. The partial-knowledge group was provided with: "They knew the so-called sweet potatoes and ate them frequently; but the potatoes that are popular today were unknown to them." In addition to this sentence, the group with complete knowledge also learned: "Moreover, they ate beans and peanuts. They knew neither the tomatoes with which other tribes were familiar nor the cocoa plant, which is widely distributed in some regions of South America." The points awarded by empirical option weighting (EOW) were option-total correlations.

Table 2

Item difficulty, item discriminatory power, and empirical option weights for each answer option of all 7 experimental items (sorted by difficulty) separately for samples 1 and 2.

			Discrimatory power		Option weights		
					Answer option		
Sample	**Item**	**Difficulty**	**NR**	**EOW**	**1**	**2**	**3**
Sample 1	7	0.71	0.35	0.47	0.35	–0.22	–0.26
	6	0.63	0.45	0.49	0.40	–0.21	–0.31
	4	0.65	0.48	0.54	0.46	–0.13	–0.42
	2	0.57	0.42	0.54	0.50	–0.22	–0.38
	1	0.49	0.43	0.74	0.51	–0.17	–0.44
	5	0.36	0.68	0.47	0.67	–0.06	–0.62
	3	0.26	0.53	0.34	0.50	–0.20	–0.28
Sample 2	7	0.73	0.35	0.55	0.35	–0.25	–0.21
	6	0.67	0.40	0.56	0.45	–0.22	–0.34
	4	0.61	0.46	0.52	0.48	–0.18	–0.40
	2	0.57	0.50	0.54	0.42	–0.18	–0.35
	1	0.47	0.51	0.74	0.43	–0.16	–0.37
	5	0.37	0.67	0.41	0.68	–0.09	–0.59
	3	0.27	0.50	0.35	0.53	–0.18	–0.32

Notes. NR = number-right scoring; EOW = empirical option weighting. Option weights for the correct answer (1), for the option that could be excluded based on complete information but not on partial information (2), and for the option that could be excluded by participants in the partial and full-knowledge conditions (3).

Table 3

Mean (M) and standard deviation (SD) of the difficulty and discriminatory power of the seven knowledge test items for samples 1 and 2, and the mean scores achieved in the three knowledge conditions under number-right scoring.

	Difficulty		**Discriminatory power**		**Mean score**		
					Knowledge condition		
Sample	***M***	***SD***	***M***	***SD***	**None**	**Partial**	**Complete**
Sample 1	.52	.17	.48	.11	1.84	3.06	6.30
Sample 2	.53	.17	.49	.10	1.84	3.17	6.26

Table 4

Mean (M), standard deviation (SD), minimum (min), maximum (max), reliability, and validity of the scores achieved by participants in the two samples under number-right scoring and empirical option weighting.

Sample	**Scoring method**	***M***	***SD***	**min**	**max**	**α**	***h***	η_g^2	**Hit rate**
Sample 1	Number-right scoring	3.67	2.13	0.00	7.00	.758	–	.77	78.52%
	Empirical option weighting	0.75	1.78	–2.70	3.40	.782	1.146	.82	87.32%
Sample 2	Number-right scoring	3.69	2.15	0.00	7.00	.766	–	.73	74.56%
	Empirical option weighting	0.75	1.76	–2.56	3.34	.790	1.151	.77	84.10%

Notes. The reliability was calculated as Cronbach's coefficient α. The Spearman-Brown statistic *h* quantifies by which factor the test would have to be lengthened under number-right scoring to achieve the same α as under empirical option weighting. As a measure of validity, the generalized eta-squared (η_g^2) represents the proportion of explained variance in an ANOVA with test scores as dependent variables and knowledge conditions as independent variables. As a second measure of validity, the hit rate indicates the proportion of correct classifications of participants' knowledge conditions in a one-dimensional linear discriminant analysis based on test scores.

Table 5

Number of correct and incorrect classifications of participants to their known knowledge conditions based on a one-dimensional linear discriminant analysis under number-right scoring (NR) and empirical option weighting (EOW).

			EOW					
Sample			**Correct**		**Incorrect**		**Σ**	
Sample 1	**NR**	**Correct**	222	(78.17%)	1	(0.35%)	223	(78.52%)
		Incorrect	26	(9.15%)	35	(12.32%)	61	(21.48%)
		Σ	248	(87.32%)	36	(12.68%)	284	(100.00%)
Sample 2	**NR**	**Correct**	209	(73.85%)	2	(0.71%)	211	(74.56%)
		Incorrect	29	(10.25%)	43	(15.19%)	72	(25.44%)
		Σ	238	(84.10%)	45	(15.90%)	283	(100.00%)

Option weights should be determined empirically and not by experts when assessing knowledge with multiple-choice items

Birk Diedenhofen and Jochen Musch

University of Düsseldorf

Author Note

Birk Diedenhofen and Jochen Musch, Department of Experimental Psychology, University of Düsseldorf. Correspondence concerning this article should be addressed to Birk Diedenhofen, Department of Experimental Psychology, University of Düsseldorf, Universitätsstrasse 1, Building 23.03, 40225 Düsseldorf, Germany. Tel.: +49 211 81-12065, Fax: +49 211 81-11753, E-mail: birk.diedenhofen@uni-duesseldorf.de

We are grateful for the help of Jana Sommer in collecting the expert options weights.

Abstract

Option weighting is an alternative multiple-choice scoring procedure that awards partial credit for incomplete knowledge reflected by certain distractor choices. Option weights may be determined either empirically or by expert judgment. We identify a major weakness in the method used by previous studies to investigate empirical option weights. When splitting the sample into two halves for cross-validation, the assessment of the usefulness of empirical option weighting critically depends on how the sample is divided. To obtain more reliable and generalizable results, we therefore used repeated random sub-sampling validation and found that empirical option weighting, but not expert option weighting, increased the reliability of a knowledge test. Neither option weighting procedure improved test validity.

Keywords: option weighting, multiple-choice test, partial knowledge, reliability, validity

Option weights should be determined empirically and not by experts when assessing knowledge with multiple-choice items

Multiple-choice tests are frequently administered in educational and personnel selection contexts to assess knowledge. The multiple-choice answer format allows for the construction of objective tests with convincing psychometric properties and the employment of efficient scoring procedures (Downing & Haladyna, 2006). Multiple-choice items are usually scored by awarding 1 point when the correct answer is chosen and 0 points when one of the distractors is chosen. This dichotomous scoring procedure is called *number-right scoring* because the total test score equals the number of correctly answered items (Lord, 1975). By awarding 0 points regardless of which distractor is chosen, number-right scoring takes an all-or-nothing approach and implies that the choice of certain distractor provides no valuable information apart from the fact that the test taker did not succeed in identifying the correct answer. Thus, number-right scoring is not able to capture partial knowledge and may therefore miss valuable information about the test taker's state of knowledge (Frary, 1989). Consider, for example, a soccer knowledge item asking "In 1992, Denmark surprisingly won the European Championship. Against whom did the team win the final with a score of 2:0?" with "Germany" as the correct answer and "Netherlands" and "Sweden" as the two distractors (see Table 1). Under number-right scoring, the test taker would be awarded 1 point for choosing Germany and 0 points for choosing the Netherlands or Sweden. It may be argued, however, that the two incorrect answers are differentially informative with regard to the domain under question. Choosing Sweden may indicate a serious lack of soccer expertise because Sweden never won a European Championship—the greatest achievement of the team in a European tournament was the participation in a semifinal. Choosing the Netherlands, a team that has already won the European Championship and participated in several

semifinals, reflects at least partial soccer knowledge. By scoring items dichotomously, number-right scoring ignores the partial knowledge that may be reflected by such distractor choices.

Table 1

An example item ("In 1992, Denmark surprisingly won the European Championship. Against whom did the team win the final with a score of 2:0?") from the soccer knowledge test and the corresponding scoring schemes for number-right scoring (NR), empirical option weighting (EOW), and expert option weighting (XOW).

Multiple-choice options	**NR**	**EOW**	**XOW**
Germany (correct answer)	1	.41	100.00
Netherlands	0	-.20	33.50
Sweden	0	-.28	23.50

Notes. Empirical option weights were calculated as the average point-biserial correlation between the choice of the respective answer option and the total test score. Expert option weights were provided on a scale ranging from 0 to 100 points.

Option weighting

Option weighting is an alternative multiple-choice scoring procedure that captures partial knowledge by assigning individual weights to each answer option (Downey, 1979). This polychotomous scoring approach is based on the assumption that test takers choose rationally between distractors and do not simply guess if they cannot identify the solution. Distractor choices are therefore assumed to be useful for discriminating between test takers of varying ability. Assigning higher weights to distractors that reflect partial knowledge may help to extract additional information contained in distractor choices that would be neglected under number-right scoring (Haladyna, 1990). It is unlikely that two test takers will receive the same scores under empirical option weighting, and empirical option weighting scores thereby approximate a continuous rather than a binomial distribution. Whereas number-right scoring numerically differentiates between only 11 knowledge levels on a test containing 10 three-option multiple-choice items, option weighting distinguishes between up to $3^{10} = 59{,}049$ different knowledge levels (cf. Haladyna, 1990).

Unlike other procedures that claim to capture partial knowledge, such as subset-selection methods or the answer-until-correct format (Frary, 1989), option weighting is based on the standard multiple-choice format. Thus, test takers do not have to familiarize themselves with a new answer format that requires a more complex and time-consuming response than the format they are most used to. However, the pivotal question in option weighting is: What score best reflects the level of knowledge that can be inferred from the choice of a particular answer option?

The amount of partial credit that should be granted for a distractor choice can be determined either empirically (empirical option weighting) or subjectively by expert judgment (expert option weighting; Stanley & Wang, 1970). The general idea behind empirical option weighting is to reward test takers' choice of a distractor that is popular among high (low) scorers with a high (low) item score. A common way to determine option weights empirically is to calculate option-total correlations. The option-total correlation is the point-biserial correlation between an option choice (i.e., 1 or 0 depending on whether the test taker chooses the option or not, respectively) and the total test score according to number-right scoring (Haladyna, 1990). In previous studies, the higher computational effort associated with empirical option weighting in comparison with conventional scoring has been viewed as a major drawback (e.g., Haladyna, 1990; Raffeld, 1975; Wang & Stanley, 1970). Today, however, many tests are scored by computers; hence, once a weighting scoring algorithm has been implemented, there is effectively no difference in effort between empirical option weighting and conventional number-right scoring.

Another drawback of empirical option weighting is that the weights that will be applied to determine a participant's score depend on the responses of other participants. Expert option

weighting avoids this problem by collecting the judgments of one or more experts in the domain under study. Experts who are qualified to estimate how much partial knowledge is reflected in distractor choices should be able to assign option weights that reward partial knowledge adequately. If more than one expert is consulted, average option weights are usually computed. Because a person's score does not depend on the performance of other test takers, test results determined under expert option weighting may enjoy higher acceptance among test takers than results determined under empirical option weighting.

The goal of the present study was to scrutinize which of three scoring procedures—empirical option weighting, expert option weighting, and conventional number-right scoring—performs best in terms of reliability and validity when applied to score a knowledge test.

Empirical vs. expert option weighting

For a meaningful comparison between empirical option weighting, expert option weighting, and number-right scoring, the three scoring procedures need to be examined under identical conditions. To this end, the three procedures must be applied to one set of answers that were provided by a single sample in response to the same multiple-choice test. Only three previous studies conducted by Echternacht (1976), Downey (1979) and Cross, Ross, and Geller (1980) met the criterion of directly contrasting all three scoring procedures using the same set of responses. Unfortunately, however, in none of these studies were the differences in reliability and validity tested for statistical significance.

Echternacht (1976) used a speeded test consisting of 30 items. Because only about 17% of the participants managed to respond to all of the items within the given time frame, the last 12 items of the test had to be removed to achieve a completion rate that surpassed 90%. Empirical option weights were then calculated using formula scores rather than number-right scores.

Formula scores slightly differ from number-right scores because they are corrected for guessing by subtracting the expected number of points that can be achieved using simple guessing from the observed total score (Lord, 1975). The empirical option weights that were based on formula scoring resulted in a descriptively higher reliability and convergent validity than number-right scoring and formula scoring. Empirical option weights based on number-right scoring were not examined. In the expert weighting condition, Echternacht (1976) did not allow expert option weights to differ between items. Rather, for each item, 6 points had to be awarded for the correct answer, 1 point for a distractor indicating partial knowledge, and -4 points for a distractor indicating no partial knowledge. Expert option weights assigned in this manner did not improve test reliability or validity. A potential explanation for these null findings is that the scoring scheme that was used was too simple and not specific enough to adequately capture the partial knowledge reflected by the distractor choices.

In a second study, which was by Downey (1979), empirical option weighting descriptively increased reliability but not predictive or concurrent validity in comparison with number-right scoring. However, the test material that was used offers a potential explanation for this lack of improvement in validity. For 9 out of 30 items, the option weight applied to the correct answer was actually smaller than the option weight applied to one of the distractors. Thus, for almost one third of the items, more able test takers chose an answer option other than the option that was identified as the correct solution. Expert option weighting was examined by asking 7 experts to rate the amount of knowledge reflected by all answer options on a 7-point scale. When the weights were averaged across the experts, the test reliability increased; however, this improvement over number-right scoring was descriptively smaller than the improvement found with empirical option weighting. Downey (1979) is the only study that also reported a

"moderate increase in predictive validity" (p. 460) under expert option weighting. However, this conclusion held for only half of the sample, and neither the increase in reliability nor the increase in validity was tested for statistical significance. In neither half of the sample did expert option weighting increase concurrent validity.

In the third study, which was by Cross et al. (1980), empirical option weighting improved the reliability but not the validity of various multiple-choice tests. The developer of the test also served as the expert who determined the expert option weights. These weights increased neither reliability nor validity. However, for many items, experts assigned a weight of zero to all distractors. Assigning weights of zero to all distractors cannot possibly lead to an improvement over number-right scoring because such a pattern is equivalent to the pattern applied in dichotomous scoring. It is possible that the experts' judgments were simply inaccurate, or that the test material was unsuitable for option weighting because the distractors did not reflect any partial knowledge.

The inconsistent results regarding validity improvements under empirical option weighting in the three studies discussed above mirror the findings of studies that did not include experts and were limited to an examination of empirical option weighting. Some of these other studies found an improvement in validity for empirical option weighting over number-right scoring (Cross & Frary, 1978; Diedenhofen & Musch, in press; Haladyna, 1990; Raffeld, 1975). Other studies, however, found no such improvement (Reilly & Jackson, 1973; Waters, 1976). Studies that were limited to the examination of expert option weighting also provided inconsistent results. Whereas Patnaik and Traub (1973) found an improvement in reliability, Hambleton, Roberts, and Traub (1970) did not obtain evidence for the superiority of expert option weighting over empirical option weighting. In a study by Kansup and Hakstian (1975),

expert option weighting led to an improvement in internal consistency reliability but not in test-retest reliability. Regarding validity, none of the three studies (Hambleton et al., 1970; Kansup & Hakstian, 1975; Patnaik & Traub, 1973) found any improvement under expert option weighting.

To summarize, no clear pattern of results could be obtained from previous comparisons of empirical and expert option weighting. It was therefore not possible to derive a clear recommendation for whether option weights should better be determined empirically or by asking for the judgments of experts.

The present study

In the present study, we compared the test reliability and validity results of empirical option weighting, expert option weighting, and number-right scoring. To compare the psychometric properties of a soccer knowledge test that was scored according to the competing procedures, we calculated coefficient α as a measure of internal consistency reliability. To measure validity, we correlated the test scores with the participants' self-rated soccer expertise.

The potential advantages of option-weighted scores over conventional dichotomous scores should be most prominent if test items are difficult (Haladyna, 1990). For easy items, respondents usually choose the correct solution, and weighted distractor options therefore cannot provide additional information. This may explain why some previous studies found no improvement in test validity when using easy items under empirical or expert option weighting. For example, in the studies by Cross et al. (1980) and Downey (1979), participants answered 68% and 67% of the items correctly, respectively. In the present study, we therefore employed a challenging test with a mean item difficulty below .5.

The method chosen to determine expert option weights may have been the cause of the poor performance of expert option weighting reported in some previous studies. When a rank

order approach was used to determine expert option weights, no improvements in test validity were found (Hambleton et al., 1970; Patnaik & Traub, 1973), and Hambleton et al. (1970) did not even observe an improvement in reliability. To gain more information per expert using a less troublesome procedure, we employed ratio scales to determine the expert option weights. Our procedure thus followed the only study that reported an improvement in reliability and validity through expert option weighting (Downey, 1979). With the exception of Echternacht (1976), previous studies involving more than one expert determined option weights by aggregating the weights across all experts. However, it is conceivable that some experts provide better option weights than others, and aggregating weights across experts disregards such potential individual differences. In addition to providing a summary analysis across all experts and unlike previous studies, we therefore repeated all analyses separately for each expert.

Repeated random sub-sampling validation

When investigating the reliability and validity of scores based on empirical option weighting, "cross-validation is a must" (Stanley & Wang, 1970, p. 27). On the one hand, if the calculation of empirical option weights and the scoring of the test takers' responses are both based on the same sample, overfitting may lead to an overestimation of the usefulness of empirical option weighting. On the other hand, when using empirical option weights computed in one sample to score responses in another, the use of independent samples may underestimate the usefulness of empirical option weighting because weights necessarily model sampling error at least in part. With the exception of Cross et al. (1980) and Waters (1976), samples from previous studies were therefore randomly split into two halves to score the responses of one half using empirical option weights that were based on the responses of the other half of the sample. Unfortunately, all previous studies that performed a cross-validation reported only the results of

one of the very large number of possible random splits that could be conducted, and did not consider how this particular split may itself have influenced the outcome. If random splits are performed repeatedly, however, the results may differ considerably from one split to another. Moreover, because the choice of the random sample split that is reported is at the researcher's discretion, choosing a particular way to split a sample has to be considered an example of *researcher degrees of freedom* according to Simmons, Nelson, and Simonsohn (2011). These authors offered the criticism that any flexibility in data analysis and reporting may increase the proportion of false positive findings. As a remedy for addressing this problem, Simmons et al. (2011) urged researchers to demonstrate that their results do not hinge on arbitrary analytical decisions. In the present study, we therefore considered the split of the sample to be a potential moderator of the results, and unlike all previous studies, we conducted a repeated random sub-sampling validation to ensure that the influence of the sample split would be visible when we evaluated empirical option weighting. Arguably, results that are based on a large number of cross-validations instead of only a single random split should be far more robust and generalizable. We therefore applied 10,000 rounds of cross-validation to calculate means and confidence intervals and compared the results thus obtained with the results we would have obtained using the conventional methods that had been employed in previous studies.

Method

Participants

We recruited a sample of 772 participants via an invitation e-mail sent to members of a panel consisting of persons who took part in previous studies conducted by the Department of XXXX (removed for blind review) at the University of XXXX. None of the participants had previously taken a test containing any of the materials used in the present study. Participants who

did not complete the questionnaire (n = 73) or who took part repeatedly using the same IP address (n = 24; Aust, Diedenhofen, Ullrich, & Musch, 2013) were excluded, leaving 675 cases (57.5% female) for analysis. The average age of the participants was 33.3 years (SD = 13.4).

Ten well-known German soccer journalists participated as experts and provided expert option weights for all answer alternatives for all items on the soccer knowledge test. Experts were invited via mail and e-mail and had the opportunity to answer either a paper or online version of the questionnaire. Six experts decided to fill out the paper version; the remaining experts participated online. All experts had previously reported on German professional soccer in the Bundesliga—the highest German soccer league—in television or radio broadcasts.

Design

All participants completed the same soccer knowledge test consisting of 27 multiple-choice items offering three answer options each. To score the test, the sample was split into two halves, and the responses of only half of the sample were scored. We scored the tests using number-right scoring, option weights devised by experts, and empirical option weights determined for each answer alternative on the basis of the responses of the other half of the sample. The reliability and validity of the test scores were determined as a function of the scoring procedure, which varied within subjects. For each scoring procedure, we calculated Cronbach's α as an index of internal consistency reliability. As an index of convergent validity, we computed the correlation between the test scores and the test taker's self-reported soccer expertise as an external criterion.

To ensure that the outcome of the experiment would not be determined solely by one particular way of splitting the sample, we conducted a repeated random sub-sampling validation by repeating all analyses 10,000 times and averaging the indices computed for each round. To

test the differences in reliability and validity between the three scoring procedures for statistical significance, we conducted two different analyses. On the one hand, we conducted a conventional analysis by statistically comparing the reliability and validity of the three scoring procedures separately for each round. To compare the dependent α coefficients, we employed the significance test by Feldt, Woodruff, and Salih (1987) implemented in the R package *cocron* (version 1.0-0; Diedenhofen, 2013). To compare the validities reflected in the correlation between the test scores and the criterion, we used the test for dependent correlations by Steiger (1980) as implemented in the R package *cocor* (Diedenhofen & Musch, 2015). On the other hand, we determined the distribution of the α coefficients and the distribution of the validity coefficients across all 10,000 rounds of random sub-sampling and computed 95% confidence intervals by determining the 2.5% and 97.5% percentiles of these distributions. Differences in reliability or validity between two scoring procedures were considered statistically significant if the respective confidence intervals did not overlap. For all of our analyses, we used the R statistical computing environment (version 3.2.2; R Core Team, 2015).

Scoring

The responses to the soccer knowledge test were scored by number-right scoring, expert option weighting, and empirical option weighting. Under number-right scoring, test takers were awarded 1 point for a correct answer and 0 points for an incorrect answer. Under both empirical and expert option weighting, the choice of each answer option was awarded with its respective weight. Under empirical option weighting, an option weight was computed as the point-biserial correlation between the choice of this particular option—1 if it was chosen and 0 if it was not—and the part-whole-corrected total test score under number-right scoring. Under expert option weighting, the weight of the correct answer option was always fixed at 100 points. In doing so,

we followed procedures by Cross et al. (1980), Hambleton et al. (1970), and Kansup and Hakstian (1975) who also held the weights of the correct answer options constant. This procedure makes the experts' task easier because it allows them to provide only two weights (on a scale ranging from 0 to 100 points) for the two distractors that can both be quantified relative to the weight of the correct solution. Apart from the experts' individual option weights, we also determined the mean of the weights of all experts for each distractor. Table 1 displays an item from the soccer knowledge test and illustrates the three scoring procedures.

Material and procedure

The online questionnaire first welcomed all participants and asked them to provide their gender and age. Next, to assess participants' soccer expertise, they answered a scale that consisted of the following 11 items. First, participants self-rated their soccer knowledge (1) on a 7-point scale that ranged from "very little" to "very much." Next, using the same scale, participants were asked to estimate how friends who were not soccer enthusiasts would probably rate the participants' level of soccer knowledge (2). Participants were then asked to estimate the percentage of people in the general population who presumably knew more about soccer (3). Then, participants were asked to indicate how often they engage in conversations about soccer (4) on a 7-point scale ranging from "very rarely" to "very often" and to report how many hours per week on average they watch soccer on TV (5). Next, participants were asked to indicate how frequently they read sports magazines (6) including "Sportbild", "Kicker", "Bravo Sport", "11 Freunde", and the sports section of their daily newspaper. Responses were again provided on a 7-point scale ranging from "never" to "very often," and the item score was calculated as the sum across all five examples. The same answer format and the same method for computing the item score were also used to ask participants how often they view or listen to sports programs on the

TV or radio (7). The programs mentioned were “Sportschau (ARD)”, “das aktuelle Sportstudio (ZDF)”, “Doppelpass (DSF/Sport1)”, “Bundesliga pur (DSF/Sport1)”, “ran (Sat1)”, “Sportsendungen im Dritten (z.B. Sport im Westen/WDR, Sport aktuell/BR, Flutlicht/SWR“), and the radio transmission of the Bundesliga matches. Next and again using the same answer format and the same calculation of the item score as in the two previous items, participants reported how often they obtain information about soccer from six popular German soccer websites (“kicker.de”, “sport1.de”, “ran.de”, “transfermarkt.de”, “spox.com”, “bundesliga.de”) and the online sports section of four major online news websites (“bild.de”, “spiegel.de”, “focus.de”, “sueddeutsche.de”) (8). Then, participants indicated the number of soccer books they had read over the last five years (9). The last two items on the soccer expertise scale asked the participants to rate how intensively they kept up with the German national team’s matches (10) and with the European Championship and the World Cup (11) on a 7-point scale ranging from “not at all” to “very intensively.” The participants’ score on the soccer expertise scale was computed as the sum of the z-standardized scores across all 11 items.

Next, all participants completed the soccer knowledge test consisting of 27 three-option multiple-choice items. The untimed test consisted of questions about facts and events in German and international soccer history. To obtain expert option weights, the same questions were also presented to the 10 soccer experts. The weight of the correct answer was always preset with a fixed value of 100 points on the expert questionnaire, and the experts were asked to indicate for each distractor the number of points that should be granted when participants chose the respective answer option. The paper version of the questionnaire was identical to the online version.

Results

To account for random effects introduced by random sample splits, we repeated all analyses 10,000 times. All of the subsequent statistics were averaged across all 10,000 rounds unless stated otherwise.

The empirical and expert option weights determined for the soccer knowledge test are listed in Table 2. As intended, the soccer knowledge test was challenging with a mean item difficulty of .48 (*SD* = .14). The mean item discriminatory power was .29 (*SD* = .16) under number-right scoring, .31 (*SD* = .18) under empirical option weighting, and .29 (*SD* = .16) under expert option weighting. The mean score on the soccer expertise scale was *M* = 0.00 (*SD* = 8.44) due to the standardization.

To compare the internal consistency reliability of the scores obtained under number-right scoring and under empirical and expert option weighting, we calculated Cronbach's α. All scoring methods achieved satisfactory α coefficients (Table 3). Descriptively, empirical option weighting yielded the highest reliability (α = .82), whereas there was no difference between expert option weighting when the weights were averaged across all experts (α = .77) and number-right scoring (α = .77). When each expert was analyzed separately, none of the experts provided weights that enabled the scores to surpass the reliability that could be achieved under number-right scoring. Conversely, the increase in α achieved by empirical option weighting was equivalent to the increase that would be expected from extending the length of the test by 36%. This was shown by calculating the *h* statistic (Haladyna, 1990, p. 236), which indicates the factor by which a number-right scored test would have to be lengthened to achieve the same increase in reliability that resulted from the use of empirical option weighting.

Table 2

Empirical option weights (computed as the average point-biserial correlation between the choice of the respective answer option and the total test score) and expert option weights (provided on a scale ranging from 0 to 100 points) for each answer option for the 27 items on the soccer knowledge test.

	Empirical option weighting			**Expert option weighting**		
	Answer options			**Answer options**		
Item	**1**	**2**	**3**	**1**	**2**	**3**
1	.17	-.04	-.16	100.00	28.30	16.50
2	.33	-.10	-.31	100.00	25.20	35.90
3	.17	.11	-.32	100.00	45.20	34.70
4	.55	-.24	-.34	100.00	26.50	36.80
5	.59	-.24	-.40	100.00	23.60	28.40
6	.09	-.02	-.10	100.00	49.00	45.30
7	.38	-.12	-.23	100.00	40.10	34.50
8	.11	-.01	-.10	100.00	40.40	31.30
9	.42	-.18	-.31	100.00	26.90	27.70
10	.59	-.27	-.39	100.00	7.20	38.00
11	.06	.02	-.10	100.00	35.30	21.90
12	.22	-.10	-.17	100.00	19.80	13.90
13	.36	-.16	-.23	100.00	33.60	66.40
14	.41	-.20	-.28	100.00	33.50	23.50
15	.38	-.16	-.30	100.00	22.50	31.20
16	.36	-.12	-.27	100.00	17.20	53.00
17	.38	-.23	-.25	100.00	11.20	25.30
18	.36	-.20	-.26	100.00	9.90	35.50
19	.48	-.22	-.36	100.00	11.60	32.40
20	.16	-.09	-.13	100.00	11.40	29.40
21	.14	.07	-.20	100.00	34.40	49.50
22	.11	-.04	-.08	100.00	32.40	31.40
23	.16	.03	-.21	100.00	16.70	50.60
24	.34	-.09	-.20	100.00	53.20	32.30
25	.33	-.11	-.24	100.00	45.00	30.00
26	.04	.02	-.12	100.00	43.80	28.30
27	.22	-.12	-.15	100.00	20.30	50.60

Notes. Option weights are given for the correct answer (1) and for the two distractors (2 and 3). Empirical option weights were sorted in decreasing order of magnitude for each item. Expert option weights were then sorted in the same order as the empirical option weights to allow for easy comparison.

Table 3

Cronbach's α coefficients and convergent validity (r_c) calculated from the participants' responses to the 27 items on the soccer knowledge test scored under number-right scoring, empirical option weighting, and expert option weighting.

	Reliability				Validity		
			Proportion of $p < .05$			Proportion of $p < .05$	
Scoring method	**α [95% CI]**	***h***	**< NR**	**> NR**	**r_c [95% CI]**	**< NR**	**> NR**
Number-right scoring [NR]	.77 [.74, .79]	–	–	–	.71 [.67, .75]	–	–
Empirical option weighting	.82 [.80, .83]	1.36	.00	1.00	.73 [.69, .76]	.00	.46
Expert option weighting							
Average across all experts	.77 [.74, .79]	1.00	.04	.01	.71 [.67, .75]	.00	.05
Expert 1	.76 [.73, .79]	0.97	.40	.00	.71 [.67, .75]	.00	.02
Expert 2	.75 [.72, .78]	0.90	.99	.00	.70 [.66, .74]	.01	.00
Expert 3	.75 [.71, .77]	0.88	1.00	.00	.70 [.66, .74]	.39	.00
Expert 4	.75 [.72, .78]	0.92	1.00	.00	.71 [.67, .75]	.00	.00
Expert 5	.75 [.72, .78]	0.90	1.00	.00	.70 [.67, .74]	.01	.00
Expert 6	.77 [.74, .79]	1.00	.03	.01	.71 [.67, .75]	.00	.07
Expert 7	.77 [.74, .79]	1.00	.05	.00	.71 [.67, .74]	.00	.00
Expert 8	.75 [.72, .78]	0.90	.98	.00	.70 [.65, .74]	.19	.00
Expert 9	.76 [.73, .79]	0.95	.81	.00	.71 [.67, .75]	.00	.02
Expert 10	.68 [.64, .71]	0.63	1.00	.00	.68 [.64, .72]	.42	.00

Notes. All values are means across 10,000 rounds of replication; 95% confidence intervals are reported in brackets. Also shown is the proportion of rounds in which empirical and expert option weighting achieved a significantly lower (< NR) or higher (> NR) reliability or validity than number-right scoring (NR). The h statistic quantifies the factor by which the test would have to be lengthened under number-right scoring to achieve the same α as under empirical and expert option weighting.

To evaluate whether the differences in reliability were statistically significant, we analyzed the data in two different ways. On the one hand, we conducted a conventional analysis and statistically compared the α coefficients of the three scoring procedures separately for each round of replication. On the other hand, we analyzed the distribution of α coefficients across all 10,000 rounds of random sub-sampling validation. The results showed that in every one of the 10,000 replications, empirical option weighting led to a statistically significantly higher α coefficient than number-right scoring. However, expert option weighting improved internal consistency reliability in only 1% of all replications (Table 3). Means and 95% confidence intervals were computed for the α coefficients in the repeated random sub-sampling analysis (Table 3). A difference in α between two scoring procedures was considered statistically significant if the respective confidence intervals did not overlap. Empirical option weighting, 95% CI for α: [.80, .83], was thus shown to be superior to expert option weighting, [.74, .79], when the weights were averaged across all experts. Moreover, empirical option weighting was also found to be superior to number-right scoring, [.74, .79]. Expert 10 was the only expert who provided expert weights that led to a significantly lower reliability, [.64, .71], than number-right scoring, [.74, .79]. The confidence intervals for expert option weighting, [.74, .79], and number-right scoring, [.74, .79], overlapped when the weights were averaged across all experts; thus, these methods were not statistically significantly different from each other.

To assess validity, we used the soccer expertise scale as an external criterion. Justifying the aggregation of the items into a composite score, the internal consistency of the participants' responses on this scale was satisfactory (α = .73). The correlations of the test scores based on the three different scoring methods and the soccer expertise scale are reported in Table 3. Descriptively, we found that empirical option weighting but not expert option weighting

improved the validity of the soccer knowledge test when number-right scoring was used as the baseline. When the experts were analyzed separately, all experts provided weights that resulted in validities that were the same or lower than the validity obtained with number-right scoring.

To test for statistically significant differences in validity between the scoring procedures, we again pursued both a conventional approach and an approach based on repeated random sub-sampling. In 46% of the 10,000 rounds of cross-validation, the correlation between the test scores and the criterion was statistically significantly higher for empirical option weighting than for number-right scoring. By contrast, expert option weights achieved a statistically significantly higher correlation with the criterion than number-right scoring in only 5% of the 10,000 rounds of cross-validation. In the complementary analysis that was based on repeated random sub-sampling validation, the means and 95% confidence intervals were computed for the validity coefficients in the repeated random sub-sampling analysis (Table 3). Because the confidence intervals for all scoring procedures overlapped, none of the observed differences in validity could be considered statistically significant.

Discussion

The present study investigated whether option weights that were based on either empirical data or expert judgments would improve the reliability and validity of a multiple-choice knowledge test in comparison with traditional number-right scoring. To avoid overfitting and capitalizing on chance, previous studies randomly split the sample and performed the calculation and application of empirical option weights on two independent subsamples. The choice of a random split, however, is an example of researcher degrees of freedom that may have a significant impact on the outcome. We therefore conducted repeated random sub-sampling validation by repeating all analyses 10,000 times and averaging the reliability and validity

indices across all rounds of cross-validation. Using this procedure, we found that empirical option weighting increased the reliability of a knowledge test compared with number-right scoring. This increase in reliability was equivalent to extending the length of the test by 36%. The computation of empirical option weights was thus shown to have the potential to save test creators from the need to create additional items. The use of option weights provided by experts was not found to improve reliability, and in one case, it even diminished reliability. Averaging weights across multiple experts neither improved nor worsened test reliability. Regarding validity, neither empirical option weights nor expert option weights were successful at improving the validity of the test. This was true regardless of whether expert option weights were averaged across all experts or employed separately for each of the 10 experts.

In general, the extent to which multiple-choice tests may benefit from option weighting is influenced by the properties of the test. First, option weighting can improve a test only by granting partial credit for incorrect answers if a sufficient number of test takers have incomplete knowledge and therefore decide to choose a distractor. Difficult tests are therefore more likely to profit from option weighting than easy tests. However, although we met this precondition in the present study by employing a test with a mean difficulty of .48, we observed no improvement in validity as a result of employing option weights. Second, option weighting may lead to an improvement over number-right scoring only if option weights actually differ from the dichotomous scoring pattern employed for number-right scoring. In the present study, the empirical and expert weights of some items turned out to be almost equivalent to number-right scoring, that is, both distractors received essentially the same weight and could therefore not distinguish between different levels of knowledge. This may explain why the test validity of the two option weighting procedures failed to surpass that of number-right scoring.

Another reason why expert option weighting performed poorly may be a lack of expertise on the part of the experts. However, given that all of the experts who participated in the current study were professional soccer journalists, they arguably should all have been able to judge the answer options given on the soccer test we employed. The soccer experts may however have performed poorly because they were unable to view the test answers from a novice's perspective. If an expert's judgment of which answer options are likely to be selected by test takers with much less knowledge is inaccurate, the expert's option weights will most likely not capture the test takers' knowledge level well. The improvements obtained by using expert option weighting that were reported by Downey (1979) may be due to the fact that the experts in his study were teachers who were experienced at anticipating which answers are attractive to students with different knowledge levels. Finally, it is conceivable that, in addition to a thorough knowledge of a subject domain, a certain degree of expertise in test construction is necessary to assign suitable option weights. No such knowledge was documented for the soccer experts who participated in the present study. On the other hand, the test developers consulted by Cross et al. (1980) did not achieve any improvements either.

The results of the conventional analysis that was performed separately for each expert suggest that if expert option weighting is employed, the weights should be aggregated across multiple experts. The option weights of more than half of the experts led to a significant decrease in reliability in nearly all rounds of replication. The option weights of experts 3, 8, and 10 even significantly diminished the validity in a considerable number of replication rounds. When option weights were averaged across all experts, however, we observed almost no decrements in reliability or validity.

Previous studies that evaluated empirical option weighting performed only a single

sample split for cross-validation. The outcome of an evaluation may however well depend on how a sample is split. When samples are split randomly, only some splits may result in an improvement. In fact, we observed an instability of results across multiple random splits in the present study, suggesting that the results of previous studies may also have depended on the particular split that was chosen. Arguably, the results of previous studies should therefore be considered with caution. In the present analysis, the chance of observing a significant improvement in test validity on the basis of a single sample split was 46%. The outcome of our analysis would therefore have resembled that of a coin toss if it had been based on the method used in previous studies. By contrast, calculating confidence intervals across a large number of sample splits provided clear evidence against an improvement in validity with the use of empirical option weights. We argue that whenever a random process is part of the research design (e.g., when random selections are made for the purpose of cross-validation), it is important to conduct a sensitivity analysis to assess the degree to which the results depend on the outcome of this random process. Variations between repeated random draws should always be reported, and results should be aggregated across a large number of draws. In view of steadily decreasing computational costs, some additional programming is usually the only obstacle to running the same analysis repeatedly. Our recommendation is therefore to employ a repeated random sub-sampling validation in all future studies investigating empirical option weighting.

To summarize, our findings suggest that rather than asking experts, test makers should rely on the empirical determination of option weights to award partial credit on multiple-choice tests. However, we found that empirical option weights improved only reliability and not validity. Even though the present approach should be repeated and validated using different tests and samples, our results tentatively suggest that the benefits of employing empirical option

weights may be rather limited. Moreover and in contrast to Downey's (1979) findings, our results also suggest that there may be no benefits to using expert option weights at all.

References

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavioral Research Methods*, *45*, 527–535. doi:10.3758/s13428-012-0265-2

Cross, L. H. & Frary, R. B. (1978). Empirical choice weighting under "guess" and "do not guess" directions. *Educational and Psychological Measurement*, *38*, 613–620. doi:10.1177/001316447803800302

Cross, L. H., Ross, F. K., & Geller, E. S. (1980). Using choice-weighted scoring of multiple-choice tests for determination of grades in college courses. *The Journal of Experimental Education*, *48*, 296–301. doi:10.1080/00220973.1980.11011747

Diedenhofen, B. (2013). *cocron: Statistical comparisons of two or more alpha coefficients*. (Version 1.0-0). Retrieved November 3, 2015, from http://comparingcronbachalphas.org/

Diedenhofen, B. & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*, e0121945. doi:10.1371/journal.pone.0121945

Diedenhofen, B. & Musch, J. (in press). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment*. doi:10.1027/1015-5759/a000295

Downey, R. G. (1979). Item-option weighting of achievement tests: Comparative study of methods. *Applied Psychological Measurement*, *3*, 453–461. doi:10.1177/014662167900300403

Downing, S. M. & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Echternacht, G. (1976). Reliability and validity of item option weighting schemes. *Educational and Psychological Measurement*, *36*, 301–309. doi:10.1177/001316447603600208

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*, 93–103. doi:10.1177/014662168701100107

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, *2*, 79–96. doi:10.1207/s15324818ame0201_5

Haladyna, T. M. (1990). Effects of empirical option weighting on estimating domain scores and making Pass/Fail decisions. *Applied Measurement in Education*, *3*, 231–244. doi:10.1207/s15324818ame0303_2

Hambleton, R. K., Roberts, D. M., & Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement*, *7*, 75–82. doi:10.1111/j.1745-3984.1970.tb00698.x

Kansup, W. & Hakstian, A. R. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. *Journal of Educational Measurement*, *12*, 219–230. doi:10.1111/j.1745-3984.1975.tb01023.x

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, *12*, 7–11. doi:10.1111/j.1745-3984.1975.tb01003.x

Patnaik, D. & Traub, R. E. (1973). Differential weighting by judged degree of correctness.

Journal of Educational Measurement, *10*, 281–286. doi:10.1111/j.1745-3984.1973.tb00805.x

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved November 3, 2015 from http://www.R-project.org/

Raffeld, P. (1975). The effects of Guttman weights on the reliability and predictive validity of objective tests when omissions are not differentially weighted. *Journal of Educational Measurement*, *12*, 179–185. doi:10.1111/j.1745-3984.1975.tb01020.x

Reilly, R. R. & Jackson, R. (1973). Effects of empirical option weighting on reliability and validity of an academic aptitude test. *Journal of Educational Measurement*, *10*, 185–193. doi:10.1111/j.1745-3984.1973.tb00796.x

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Stanley, J. C. & Wang, M. W. (1970). Weighting test items and test-item options, an overview of the analytical and empirical literature. *Educational and Psychological Measurement*, *30*, 21–35. doi:10.1177/001316447003000102

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251. doi:10.1037/0033-2909.87.2.245

Wang, M. W. & Stanley, J. C. (1970). Differential weighting: A review of methods and

empirical studies. *Review of Educational Research*, *40*, 663–705. doi:10.3102/00346543040005663

Waters, B. K. (1976). The measurement of partial knowledge: A comparison between two empirical option-weighting methods and rights-only scoring. *The Journal of Educational Research*, *69*, 256–260. doi:10.1080/00220671.1976.10884892

PLOS ONE

RESEARCH ARTICLE

cocor: A Comprehensive Solution for the Statistical Comparison of Correlations

Birk Diedenhofen[1]*, **Jochen Musch**[1]

1 Department of Experimental Psychology, University of Duesseldorf, Duesseldorf, Germany

* birk.diedenhofen@uni-duesseldorf.de (BD)

Abstract

A valid comparison of the magnitude of two correlations requires researchers to directly contrast the correlations using an appropriate statistical test. In many popular statistics packages, however, tests for the significance of the difference between correlations are missing. To close this gap, we introduce `cocor`, a free software package for the R programming language. The `cocor` package covers a broad range of tests including the comparisons of independent and dependent correlations with either overlapping or nonoverlapping variables. The package also includes an implementation of Zou's confidence interval for all of these comparisons. The platform independent `cocor` package enhances the R statistical computing environment and is available for scripting. Two different graphical user interfaces —a plugin for RKWard and a web interface—make `cocor` a convenient and user-friendly tool.

Citation: Diedenhofen B, Musch J (2015) cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. PLoS ONE 10(4): e0121945. doi:10.1371/journal.pone.0121945

Academic Editor: Jake Olivier, University of New South Wales, AUSTRALIA

Received: October 30, 2014

Accepted: February 5, 2015

Published: April 2, 2015

Data Availability Statement: The cocor R package can be downloaded from http://cran.r-project.org/package = cocor. A web front-end to conveniently access the functionality of the cocor package is available at http://comparingcorrelations.org.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Determining the relationship between two variables is at the heart of many research endeavours. In the social sciences, the most popular statistical method to quantify the magnitude of an association between two numeric variables is the Pearson product-moment correlation. It indicates the strength of a linear relationship between two variables, which may be either positive, negative, or zero. In many research contexts, it is necessary to compare the magnitude of two such correlations, for example, if a researcher wants to know whether an association changed after a treatment, or whether it differs between two groups of interest. When comparing correlations, a test of significance is necessary to control for the possibility of an observed difference occurring simply by chance. However, many introductory statistics textbooks [1–5] do not even mention significance tests for correlations. Also in research practice, the necessity of conducting a proper statistical test when comparing the magnitude of correlations is often ignored. For example, in neuroscientific investigations, correlations between behavioral measures and brain areas are often determined to identify the brain area that is most strongly involved in a given task. Rousselet and Pernet [6] criticized that such studies rarely provide quantitative tests of the difference between correlations. Instead, many authors fall prey to a statistical fallacy, and wrongly consider the existence of a significant and a nonsignificant correlation as providing sufficient evidence for a significant difference between these two

correlations. Nieuwenhuis, Forstmann, and Wagenmakers [7] also found that, when making a comparison between correlations, researchers frequently interpreted a significant correlation in one condition and a nonsignificant correlation in another condition as providing evidence for different correlations in the two conditions. Such an interpretation, however, is fallacious. As pointed out by Rosnow and Rosenthal [8], "God loves the .06 nearly as much as the .05". To make a valid, meaningful, and interpretable comparison between two correlations, it is necessary to directly contrast the two correlations under investigation using an appropriate statistical test [7].

Even when recognizing the importance of a formal statistical test of the difference between correlations, the researcher has many different significance tests to choose from, and the choice of the correct method is vital. Before picking a test, researchers have to distinguish between the following three cases: (1) The correlations were measured in two independent groups A and B. This case applies, for example, if a researcher wants to compare the correlations between anxiety and extraversion in two different groups A and B ($\rho_A = \rho_B$). If the two groups are dependent, the relationship between them needs further differentiation: (2) The two correlations can be overlapping ($\rho_{A12} = \rho_{A23}$), i.e., the correlations have one variable in common. ρ_{A12} and ρ_{A23} refer to the population correlations in group A between variables 1 and 2 and variables 2 and 3, respectively. For instance, a researcher may be interested in determining whether the correlation between anxiety and extraversion is smaller than between anxiety and diligence within the same group A. (3) In the case of two dependent correlations, the two correlations can also be nonoverlapping ($\rho_{A12} = \rho_{A34}$), i.e., they have no variable in common. This case applies, for example, if a researcher wants to determine whether the correlation between anxiety and extraversion is higher than the correlation between intelligence and creativity within the same group. A researcher also faces nonoverlapping dependent correlations when investigating whether the correlation between two variables is higher before rather than after a treatment provided to the same group.

For each of these three cases, various tests have been proposed. An overview of the tests for comparing independent correlations is provided in Table 1, and for comparing dependent correlations—overlapping and nonoverlapping—in Tables 2 and 3, respectively. May and Hittner [9] compared the statistical power and Type I error rate of several tests for dependent overlapping correlations, and found no test to be uniformly preferable. Instead, they concluded that the best choice is influenced by sample size, predictor intercorrelation, effect size, and predictor-criterion correlation. Because no clear recommendation for any of these tests can be formulated that applies under all circumstances, and because different methods may be optimal for a research question at hand, it is important that researchers are provided with a tool that allows them to choose freely between all available options. Detailed discussions of the competing tests for comparing dependent overlapping correlations are given in Dunn and Clark [10], Hittner, May, and Silver [11], May and Hittner [9], Neill and Dunn [12], and Steiger [13]. For the case of dependent nonoverlapping correlations, the pros and cons of various tests are discussed in Raghunathan, Rosenthal, and Rubin [14], Silver, Hittner, and May [15], and Steiger [13]. In contrast to most other approaches, Zou [16] has advocated a test that is based on the

Table 1. Software implementing tests for comparing two correlations based on independent groups.

Test	psych	multilevel	Weaver & Wuensch	cocor
Fisher's [20] *z*	•	•	•	•
Zou's [16] confidence interval			•	•

doi:10.1371/journal.pone.0121945.t001

Table 2. Software implementing tests for comparing two correlations based on dependent groups with overlapping variables.

Test	psych	multilevel	DEPCORR	DEPCOR	Weaver & Wuensch	cocor
Pearson and Filon's [21] *z*						•
Hotelling's [22] *t*			•			•
Williams' [23] *t*	•	•	•	•	•	•
Olkin's [24] *z*			•			•
Dunn and Clark's [25] *z*			•	•		•
Hendrickson et al.'s [26] modification of Williams' [23] *t*			•			•
Steiger's [13] modification of Dunn and Clark's [25] *z*			•	•		•
Meng, Rosenthal, and Rubin's [27] *z*			•	•		•
Hittner et al.'s [11] modification of Dunn and Clark's [25] *z*				•		•
Zou's [16] confidence interval					•	•

doi:10.1371/journal.pone.0121945.t002

computation of confidence intervals, which are often regarded as superior to significance testing because they separately indicate the magnitude and the precision of an estimated effect [17, 18]. Confidence intervals can be used to test whether a correlation significantly differs from zero or from some constant, and whether the difference between two correlations exceeds a predefined threshold. Zou's confidence interval [16] is available for comparisons of independent and dependent correlations with either overlapping or nonoverlapping variables. The tests proposed by Zou [16] have been compared to other confidence interval procedures by Wilcox [19].

Existing Software

Many popular statistics programs do not provide any, or only a subset of the significance tests described above. Moreover, existing programs that allow for statistical comparisons between correlations are isolated stand-alone applications and do not come with a graphical user interface (GUI). For example, DEPCOR [28] is a program that is limited to comparisons of two dependent correlations—either overlapping or nonoverlapping. The program is written in Fortran and runs in a DOS command prompt console under the Windows platform. Another available package, DEPCORR [29], is an SAS macro [30] for comparing two dependent overlapping correlations. The latest release of SAS/STAT software (version 9.4) runs on Windows and Linux systems. However, DEPCORR has no GUI and covers only one of the three cases described above. The two packages psych [31] and multilevel [32] for the R programming language [33] also offer functions to compare two dependent or independent correlations. However, each of these functions covers only one or two of the many different available tests of comparison, and there is no GUI available to access the functions of the packages. Weaver and

Table 3. Software implementing tests for comparing two correlations based on dependent groups with nonoverlapping variables.

Test	psych	DEPCOR	Weaver & Wuensch	cocor
Pearson and Filon's [21] *z*			•	•
Dunn and Clark's [25] *z*	•	•		•
Steiger's [13] modification of Dunn and Clark's [25] *z*	•	•		•
Raghunathan, Rosenthal, and Rubin's [14] modification of Pearson and Filon's [21] *z*			•	•
Silver, Hittner, and May's [15] modification of Dunn and Clark's [25] *z*		•		•
Zou's [16] confidence interval			•	•

doi:10.1371/journal.pone.0121945.t003

Wuensch [34] recently published thoroughly documented scripts for comparing dependent or independent correlations in SPSS and SAS.

cocor

With `cocor` (version 1.1-0), we provide a comprehensive solution to compare two correlations based on either dependent or independent groups. The `cocor` package enhances the R programming environment [33], which is freely available for Windows, Mac, and Linux systems and can be downloaded from CRAN (http://cran.r-project.org/package = cocor). All that is needed to install the `cocor` package is to type `install.packages("cocor")` in the R console, and the functionality of the package is made available by typing `library("cocor")`. The function `cocor()` calculates and compares correlations from raw data. The underlying variables are specified via a formula interface (see Fig. 1). If raw data are not available, `cocor` offers three functions to compare correlation coefficients that have already

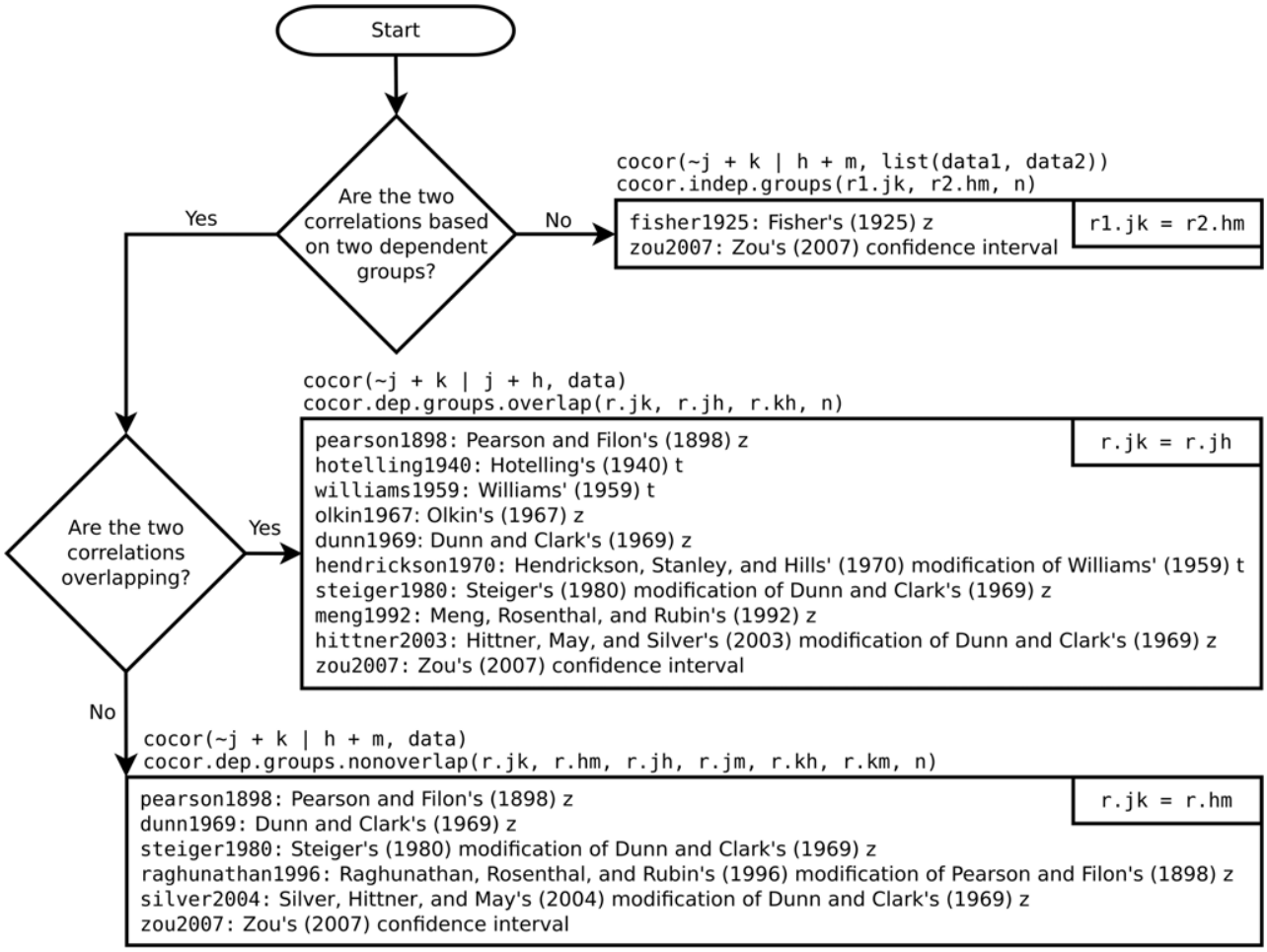

Fig 1. A flowchart of how to use the four main functions of cocor, displaying all available tests. For each case, an example of the formula passed as an argument to the `cocor()` function and the required correlation coefficients for the functions `cocor.indep.groups()`, `cocor.dep.groups.overlap()`, and `cocor.dep.groups.nonoverlap()` are given. The test label before the colon may be passed as a function argument to calculate specific tests only.

doi:10.1371/journal.pone.0121945.g001

been determined. The function `cocor.indep.groups()` compares two independent correlations, whereas the functions `cocor.dep.groups.overlap()` and `cocor.dep.groups.nonoverlap()` compare two dependent overlapping or nonoverlapping correlations, respectively. Internally, `cocor()` passes the calculated correlations coefficients to one of these three functions. All functions allow to specify the argument `null.value` to test whether the difference between the correlations exceeds a given threshold using the confidence intervals by Zou [16]. The results are either returned as an S4 object of class `cocor` whose input and result parameters can be obtained using the `get.cocor.input()` and `get.cocor.results()` functions, respectively. Optionally, results may also be returned as a list of class `htest`. By default, all tests available are calculated. Specific tests can be selected by passing a test label to the function using the `test` argument. The flowchart in Fig. 1 shows how to access the available tests and lists them with their individual test label (e.g., `zou2007`). The formulae of all implemented tests are detailed in S1 Appendix.

A comparison of `cocor` with competing software can be found in Tables 1–3. These tables show that `cocor` offers a larger variety of tests and a more comprehensive approach than all previous solutions. In particular, `cocor` is the first R package to implement the tests by Zou [16]. Further unique features of the `cocor` package are the formula interface for comparing correlations that extracts the correlations from data, and the unified function for statistical tests capable of comparing both, independent and dependent correlations with either overlapping or nonoverlapping variables.

Some limitations of `cocor` should be acknowledged, however. First, `cocor` is limited to the comparison of two correlations. The simultaneous comparison of more than two correlations needs tests that go beyond the scope of the present contribution [35–37]. Second, `cocor` does not allow one to employ structural equation models that are needed for more advanced, but also more complex approaches to the statistical comparison of correlations [38, 39].

GUIs for cocor

There are two convenient ways to use `cocor` via a GUI. First, the package includes a plugin for the platform independent R front-end RKWard [40] (Fig. 2). Second, for those unfamiliar with R, a web interface is also available at http://comparingcorrelations.org (Fig. 3).

Thus, `cocor` offers the best of two worlds: On the one hand, it has the power of a scripting language with the possibility of automation. On the other hand, the two available GUIs allow even inexperienced users to use `cocor` in a convenient way. As `cocor` is embedded in the R environment for statistical computing, it allows for a seamless integration into R analyses. R code can be generated via the GUIs and used for subsequent batch analyses. Since `cocor` is published under the GNU General Public License (GPL; version 3 or higher), all users are invited to inspect, use, copy, modify, and redistribute the code under the same license.

Code Examples

In the following, using fictional data, examples are given for all three cases that may occur when comparing correlations.

Comparison of Two Correlations Based on Independent Groups

The first example presents code for the comparison of the correlations between a score achieved on a logic task (`logic`) and an intelligence measure A (`intelligence.a`) in two different groups. Note that the underlying data set (`aptitude`) is a list that contains two separate data sets.

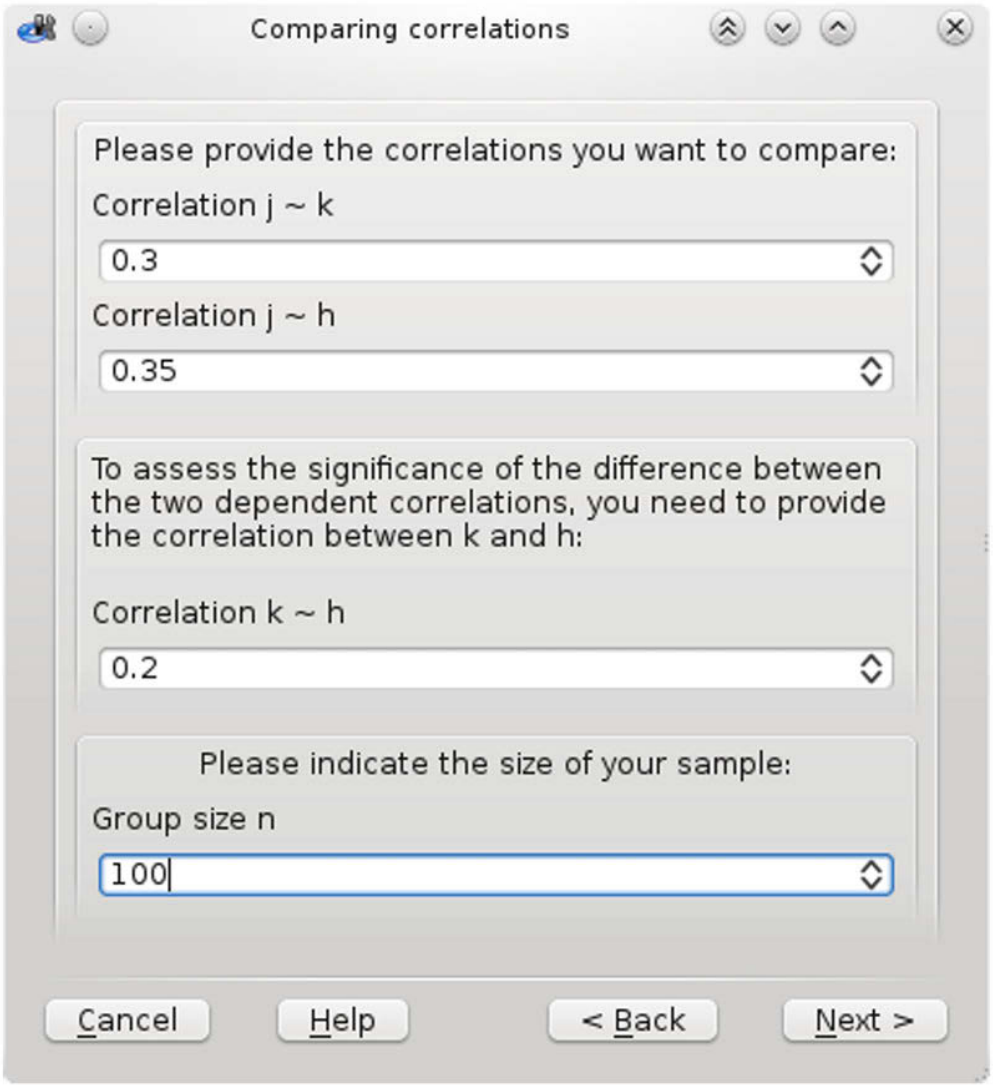

Fig 2. Screenshot of the cocor GUI plugin for RKWard.

doi:10.1371/journal.pone.0121945.g002

```
R> require ("cocor")
R> data ("aptitude")
R> cocor (~logic + intelligence.a | logic + intelligence.a,
+ aptitude)
  Results of a comparison of two correlations based on independent
  groups
Comparison between r1.jk (logic, intelligence.a) = 0.3213 and r2.
hm (logic, intelligence.a) = 0.2024
Difference: r1.jk—r2.hm = 0.1189
Data: sample1: j = logic, k = intelligence.a; sample2: h = logic,
m = intelligence.a
Group sizes: n1 = 291, n2 = 334
Null hypothesis: r1.jk is equal to r2.hm
```

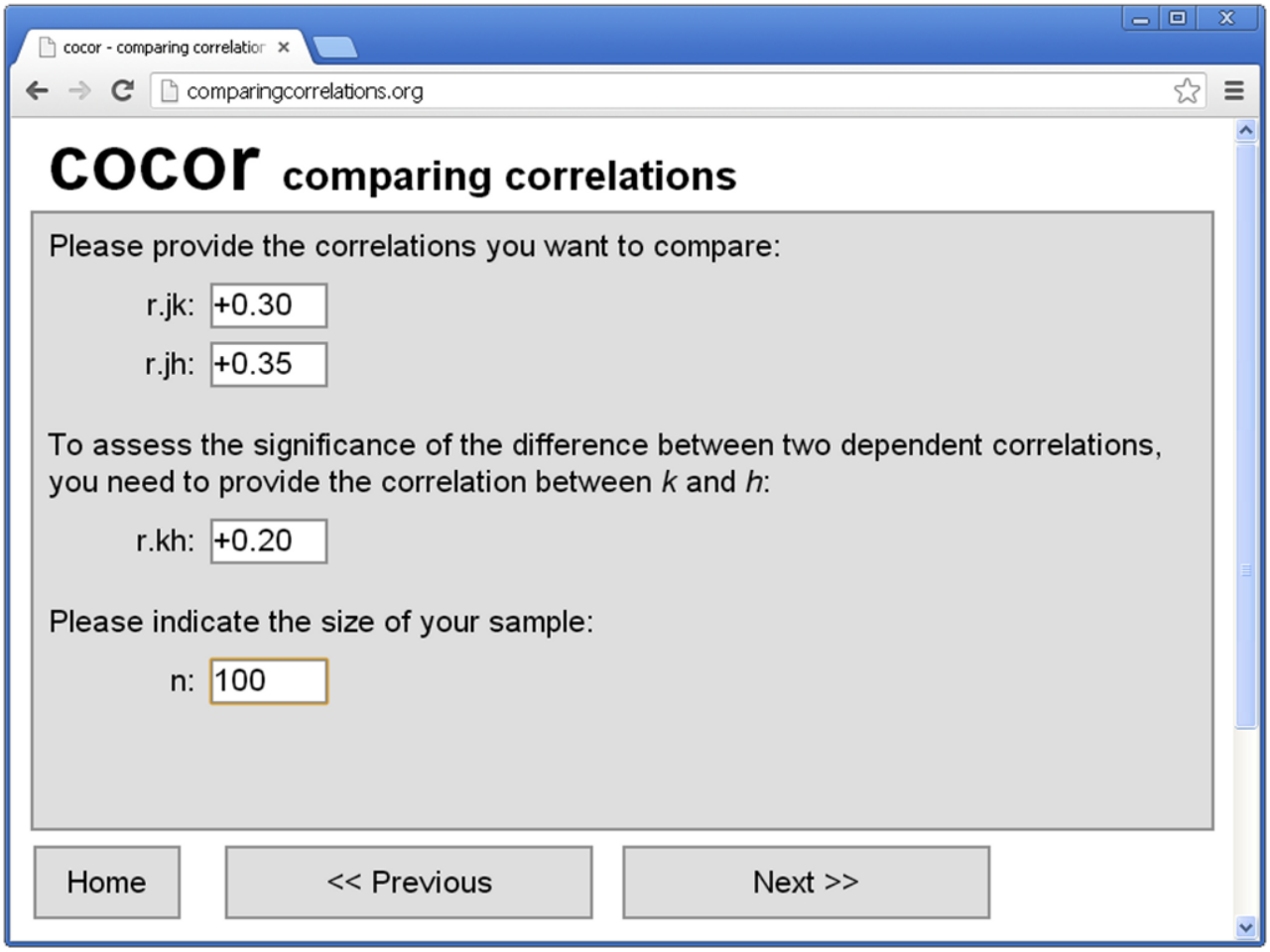

Fig 3. Screenshot of the cocor web interface on http://comparingcorrelations.org.

doi:10.1371/journal.pone.0121945.g003

```
Alternative hypothesis: r1.jk is not equal to r2.hm (two-sided)
Alpha: 0.05
fisher1925: Fisher's z (1925)
  z = 1.5869, p-value = 0.1125
  Null hypothesis retained
zou2007: Zou's (2007) confidence interval
  95% confidence interval for r1.jk-r2.hm: -0.0281 0.2637
  Null hypothesis retained (Interval includes 0)
```

In this example, the test result indicates that the difference between the two correlations `r1.jk` and `r2.hm` is not significant, and the null hypothesis cannot be rejected. Alternatively, the same comparison can also be conducted based on the correlation coefficients and the group sizes using the function `cocor.indep.groups()`.

```
R> cocor.indep.groups(r1.jk = 0.3213, r2.hm = 0.2024, n1 = 291,
+ n2 = 334)
```

Comparison of Two Overlapping Correlations Based on Dependent Groups

The second example code determines whether the correlation between a score achieved on general knowledge questions (`knowledge`) and an intelligence measure A (`intelligence.a`) differs from the correlation between a score achieved on a logic task (`logic`) and the same intelligence measure A (`intelligence.a`) within a group of $n = 291$ persons.

```
R> cocor (~knowledge + intelligence.a | logic + + intelligence.a,
aptitude[[ "sample1"]] )
  Results of a comparison of two overlapping correlations based on
  dependent groups
Comparison between r.jk (intelligence.a, knowledge) = 0.1038 and
r.jh (intelligence.a, logic) = 0.3213
Difference: r.jk—r.jh = -0.2175
Related correlation: r.kh = 0.0257
Data: aptitude[[ "sample1"]] : j = intelligence.a, k = knowledge,
h = logic
Group size: n = 291
Null hypothesis: r.jk is equal to r.jh
Alternative hypothesis: r.jk is not equal to r.jh (two-sided)
Alpha: 0.05
pearson1898: Pearson and Filon' s z (1898)
  z = -2.7914, p-value = 0.0052
  Null hypothesis rejected
hotelling1940: Hotelling' s t (1940)
  t = -2.8066, df = 288, p-value = 0.0053
  Null hypothesis rejected
williams1959: Williams' t (1959)
  t = -2.7743, df = 288, p-value = 0.0059
  Null hypothesis rejected
olkin1967: Olkin' s z (1967)
  z = -2.7914, p-value = 0.0052
  Null hypothesis rejected
dunn1969: Dunn and Clark' s z (1969)
  z = -2.7595, p-value = 0.0058
  Null hypothesis rejected
hendrickson1970: Hendrickson, Stanley, and Hills' (1970) modifi-
cation of Williams' t (1959)
  t = -2.8065, df = 288, p-value = 0.0053
  Null hypothesis rejected
steiger1980: Steiger' s (1980) modification of Dunn and Clark' s z
(1969) using average correlations
  z = -2.7513, p-value = 0.0059
  Null hypothesis rejected
meng1992: Meng, Rosenthal, and Rubin' s z (1992)
  z = -2.7432, p-value = 0.0061
  Null hypothesis rejected
  95% confidence interval for r.jk—r.jh: -0.3925 -0.0654
```

```
  Null hypothesis rejected (Interval does not include 0)
hittner2003: Hittner, May, and Silver's (2003) modification of
Dunn and Clark's z (1969) using a backtransformed average Fisher's
(1921) Z procedure
  z = -2.7505, p-value = 0.0059
  Null hypothesis rejected
zou2007: Zou's (2007) confidence interval
  95% confidence interval for r.jk—r.jh: -0.3689 -0.0630
  Null hypothesis rejected (Interval does not include 0)
```

The results of all tests lead to the convergent conclusion that the difference between the two correlations `r.jk` and `r.jh` is significant, and the null hypothesis should be rejected. Alternatively, the same comparison can also be conducted based on the correlation coefficients and the group size using the function `cocor.dep.groups.overlap()`.

```
R> cocor.dep.groups.overlap (r.jk = 0.1038, r.jh = 0.3213, + r.
kh = 0.0257, n = 291)
```

Comparison of Two Nonoverlapping Correlations Based on Dependent Groups

The third example code tests whether the correlation between a score achieved on general knowledge questions (`knowledge`) and an intelligence measure A (`intelligence.a`) differs from the correlation between a score achieved on a logic task (`logic`) and an intelligence measure B (`intelligence.b`) within the same group of $n = 291$ persons.

```
R> cocor (~knowledge + intelligence.a | logic + + intelligence.b,
aptitude[[ "sample1"]])
  Results of a comparison of two nonoverlapping correlations based
  on dependent groups
Comparison between r.jk (knowledge, intelligence.a) = 0.1038 and
r.hm (logic, intelligence.b) = 0.2679
Difference: r.jk—r.hm = -0.164
Related correlations: r.jh = 0.0257, r.jm = 0.1713, r.kh = 0.3213,
r.km = 0.4731
Data: aptitude[[ "sample1"]] : j = knowledge, k = intelligence.a,
h = logic, m = intelligence.b
Group size: n = 291
Null hypothesis: r.jk is equal to r.hm
Alternative hypothesis: r.jk is not equal to r.hm (two-sided)
Alpha: 0.05
pearson1898: Pearson and Filon's z (1898)
  z = -2.0998, p-value = 0.0357
  Null hypothesis rejected
dunn1969: Dunn and Clark's z (1969)
  z = -2.0811, p-value = 0.0374
  Null hypothesis rejected
steiger1980: Steiger's (1980) modification of Dunn and Clark's z
(1969) using average correlations
```

```
  z = -2.0755, p-value = 0.0379
  Null hypothesis rejected
raghunathan1996: Raghunathan, Rosenthal, and Rubin's (1996) modi-
fication of Pearson and Filon's z (1898)
  z = -2.0811, p-value = 0.0374
  Null hypothesis rejected
silver2004: Silver, Hittner, and May's (2004) modification of Dunn
and Clark's z (1969) using a backtransformed average Fisher's
(1921) Z procedure
  z = -2.0753, p-value = 0.0380
  Null hypothesis rejected
zou2007: Zou's (2007) confidence interval
  95% confidence interval for r.jk—r.hm: -0.3162 -0.0095
  Null hypothesis rejected (Interval does not include 0)
```

Also in this example, the test results converge in showing that the difference between the two correlations `r.jk` and `r.hm` is significant, and the null hypothesis should be rejected. Alternatively, the same comparison can also be conducted based on the correlation coefficients and the group size using the function `cocor.dep.groups.nonoverlap()`.

```
R> cocor.dep.groups.nonoverlap (r.jk = 0.1038, r.hm = 0.2679, + r.
jh = 0.0257, r.jm = 0.1713, r.kh = 0.3213, + r.km = 0.4731, n = 291)
```

Discussion and Summary

In this article, we introduced `cocor`, a free software package for the R programming language [33]. The `cocor` package provides a wide range of tests for comparisons of independent and dependent correlations with either overlapping or nonoverlapping variables. Unlike existing solutions, `cocor` is available for scripting within the R environment, while offering two convenient GUIs: a plugin for RKWard [40] and a web interface. Thus, `cocor` enables users of all knowledge levels to access a large variety of tests for comparing correlations in a convenient and user-friendly way.

Supporting Information

S1 Appendix. Documentation of All Tests Implemented in cocor.
(PDF)

Acknowledgments

We would like to thank Meik Michalke for his valuable and constructive suggestions during the development of the `cocor` package.

Author Contributions

Wrote the paper: BD JM. Software development: BD.

References

1. Baguley T.Serious Stats: A Guide to Advanced Statistics for the Behavioral Sciences. Basingstoke, UK: Palgrave Macmillan; 2012.

2. Bakeman R, Robinson BF. Understanding Statistics in the Behavioral Sciences. Mahwah, NJ: Lawrence Erlbaum; 2005.
3. Freund JE, Simon GA. Modern Elementary Statistics. 9th ed. London, UK: Prentice-Hall; 1997.
4. Larson R, Farber B. Elementary Statistics: Picturing the World. 5th ed. Boston, MA: Pearson; 2011.
5. Wright DB, London K. First (and Second) Steps in Statistics. 2nd ed. London, UK: Sage; 2009.
6. Rousselet GA, Pernet CR. Improving Standards in Brain-Behavior Correlation Analyses. Front Hum Neurosci. 2012; 6: 1–11. doi: 10.3389/fnhum.2012.00119
7. Nieuwenhuis S, Forstmann BU, Wagenmakers EJ. Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance. Nat Neurosci. 2011; 14: 1105–1107. doi: 10.1038/nn.2886 PMID: 21878926
8. Rosnow RL, Rosenthal R. Statistical Procedures and the Justification of Knowledge in Psychological Science. Am Psychol. 1989; 44: 1276–1284. doi: 10.1037/0003-066X.44.10.1276
9. May K, Hittner JB. Tests for Comparing Dependent Correlations Revisited: A Monte Carlo Study. J Exp Educ. 1997; 65: 257–269. doi: 10.1080/00220973.1997.9943458
10. Dunn OJ, Clark VA. Comparison of Tests of the Equality of Dependent Correlation Coefficients. J Am Stat Assoc. 1971; 66: 904–908. doi: 10.1080/01621459.1971.10482369
11. Hittner JB, May K, Silver NC. A Monte Carlo Evaluation of Tests for Comparing Dependent Correlations. J Gen Psychol. 2003; 130: 149–168. doi: 10.1080/00221300309601282 PMID: 12773018
12. Neill JJ, Dunn OJ. Equality of Dependent Correlation Coefficients. Biometrics. 1975; 31: 531–543. doi: 10.2307/2529435
13. Steiger JH.Tests for Comparing Elements of a Correlation Matrix. Psychol Bull. 1980; 87: 245–251. doi: 10.1037/0033-2909.87.2.245
14. Raghunathan TE, Rosenthal R, Rubin DB. Comparing Correlated but Nonoverlapping Correlations. Psychol Methods. 1996; 1: 178–183. doi: 10.1037/1082-989X.1.2.178
15. Silver NC, Hittner JB, May K. Testing Dependent Correlations with Nonoverlapping Variables: A Monte Carlo Simulation. J Exp Educ. 2004; 73: 53–69. doi: 10.3200/JEXE.71.1.53-70
16. Zou GY. Toward Using Confidence Intervals to Compare Correlations. Psychol Methods. 2007; 12: 399–413. doi: 10.1037/1082-989X.12.4.399 PMID: 18179351
17. Cohen J.The Earth Is Round (p < .05). Am Psychol. 1994; 49: 997–1003. doi: 10.1037/0003-066X.49.12.997
18. Olkin I, Finn JD. Correlation Redux. Psychol Bull. 1995; 118: 155–164. doi: 10.1037/0033-2909.118.1.155
19. Wilcox RR. Comparing Pearson Correlations: Dealing with Heteroscedasticity and Nonnormality. Commun Stat Simul Comput. 2009; 38: 2220–2234. doi: 10.1080/03610910903289151
20. Fisher RA. Statistical Methods for Research Workers. Edinburgh, Scotland: Oliver and Boyd; 1925. Available: http://psychclassics.yorku.ca. Accessed 21 February 2015.
21. Pearson K, Filon LNG. Mathematical Contributions to Theory of Evolution: IV. On the Probable Errors of Frequency Constants and on the Influence of Random Selection and Correlation. Philos Trans R Soc Lond A. 1898; 191: 229–311.
22. Hotelling H.The Selection of Variates for Use in Prediction, with Some Comments on the General Problem of Nuisance Parameters. Ann Math Stat. 1940; 11: 271–283. doi: 10.1214/aoms/1177731867
23. Williams EJ. The Comparison of Regression Variables. J R Stat Soc B. 1959; 21: 396–399. Available: http://www.jstor.org/stable/2983809. Accessed 21 February 2015.
24. Olkin I.Correlations Revisited. In: Stanley JC, editor. Improving Experimental Design and Statistical Analysis. Chicago, IL: Rand McNally; 1967. pp. 102–128.
25. Dunn OJ, Clark VA. Correlation Coefficients Measured on the Same Individuals. J Am Stat Assoc. 1969; 64: 366–377. doi: 10.1080/01621459.1969.10500981
26. Hendrickson GF, Stanley JC, Hills JR. Olkin's New Formula for Significance of r13 vs. r23 Compared with Hotelling's Method. Am Educ Res J. 1970; 7: 189–195. doi: 10.2307/1162159
27. Meng XL, Rosenthal R, Rubin DB. Comparing Correlated Correlation Coefficients. Psychol Bull. 1992; 111: 172–175. doi: 10.1037/0033-2909.111.1.172
28. Silver NC, Hittner JB, May K. A FORTRAN 77 Program for Comparing Dependent Correlations. Appl Psychol Meas. 2006; 30: 152–153. doi: 10.1177/0146621605277132
29. Hittner JB, May K. DEPCORR: A SAS Program for Comparing Dependent Correlations. Appl Psychol Meas. 1998; 22: 93–94. doi: 10.1177/01466216980221010

30. SI Inc. SAS/STAT Software, Version 9.4. Cary, NC; 2013. Available: http://www.sas.com. Accessed 21 February 2015.
31. Revelle W. psych: Procedures for psychological, psychometric, and personality research; 2014. R package version 1.4.8. Available: http://cran.R-project.org/package = psych. Accessed 21 February 2015.
32. Bliese P. multilevel: Multilevel Functions; 2013. R package version 2.5. Available: http://cran.R-project.org/package = multilevel. Accessed 21 February 2015.
33. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2014. Available: http://www.R-project.org. Accessed 21 February 2015.
34. Weaver B, Wuensch KL. SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. Behav Res Methods. 2013; 45: 880–895. doi: 10.3758/s13428-012-0289-7 PMID: 23344734
35. Levy KJ. A Multiple Range Procedure for Independent Correlations. Educ Psychol Meas. 1976; 36: 27–31. doi: 10.1177/001316447603600103
36. Paul SR. A Multiple Range Procedure for Independent Correlations. Can J Stat. 1989; 17: 217–227. doi: 10.2307/3314850
37. Silver NC, Zaikina H, Hittner JB, May K. INCOR: A Computer Program for Testing Differences among Independent Correlations. Mol Ecol Resour. 2008; 8: 763–764. doi: 10.1111/j.1755-0998.2008.02107.x PMID: 21585885
38. Cheung MWL, Chan W. Testing Dependent Correlations via Structural Equation Modeling. Org Res Methods. 2004; 7: 206–223. doi: 10.1177/1094428104264024
39. Cheung MWL. Constructing Approximate Confidence Intervals for Parameters with Structural Equation Models. Struct Equ Modeling. 2009; 16: 267–294. doi: 10.1080/10705510902751291
40. Rödiger S, Friedrichsmeier T, Kapat P, Michalke M. RKWard: A Comprehensive Graphical User Interface and Integrated Development Environment for Statistical Analysis with R. J Stat Softw. 2012; 49: 1–34. Available: http://www.jstatsoft.org/v49/i09. Accessed 21 February 2015.

PLOS ONE

CORRECTION

Correction: cocor: A Comprehensive Solution for the Statistical Comparison of Correlations

The *PLOS ONE* Staff

The URL in the Data Availability statement for this paper is incorrect. The correct statement is: "Data Availability Statement: The cocor R package can be downloaded from http://cran.r-project.org/package = cocor. A web front-end to conveniently access the functionality of the cocor package is available at http://comparingcorrelations.org." The publisher apologizes for the error.

There is an error in the URL in the second sentence of the subsection "cocor" in the Introduction. The correct sentence should be: "The cocor package enhances the R programming environment [33], which is freely available for Windows, Mac, and Linux systems and can be downloaded from CRAN (http://cran.r-project.org/package = cocor)." The publisher apologizes for the error.

There is an error in the URL in reference 31 of the References. The correct reference should be: "Revelle W. psych: Procedures for psychological, psychometric, and personality research; 2014. R package version 1.4.8. Available: http://cran.R-project.org/package = psych. Accessed 21 February 2015." The publisher apologizes for the error.

There is an error in the URL in reference 32 of the References. The correct reference should be: "Bliese P. multilevel: Multilevel Functions; 2013. R package version 2.5. Available: http://cran.R-project.org/package = multilevel. Accessed 21 February 2015." The publisher apologizes for the error.

There are several errors in the "Comparison of Two Correlations Based on Independent Groups" subsection of the Code Examples section. The publisher apologizes for the error. Please view the correct code here. Figs 4 and 5

Citation: The *PLOS ONE* Staff (2015) Correction: cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. PLoS ONE 10(6): e0131499. doi:10.1371/journal.pone.0131499

Published: June 26, 2015

```
R> require("cocor")
R> data("aptitude")
R> cocor(~logic + intelligence.a | logic + intelligence.a,
+   aptitude)

  Results of a comparison of two correlations based on
  independent groups

Comparison between r1.jk (logic, intelligence.a) = 0.3213
  and r2.hm (logic, intelligence.a) = 0.2024
Difference: r1.jk - r2.hm = 0.1189
Data: sample1: j = logic, k = intelligence.a; sample2:
  h = logic, m = intelligence.a
Group sizes: n1 = 291, n2 = 334
Null hypothesis: r1.jk is equal to r2.hm
Alternative hypothesis: r1.jk is not equal to r2.hm
  (two-sided)
Alpha: 0.05

fisher1925: Fisher's z (1925)
  z = 1.5869, p-value = 0.1125
  Null hypothesis retained

zou2007: Zou's (2007) confidence interval
  95% confidence interval for r1.jk - r2.hm: -0.0281 0.2637
  Null hypothesis retained (Interval includes 0)
```

Fig 4.

doi:10.1371/journal.pone.0131499.g001

```
R> cocor.indep.groups(r1.jk=0.3213, r2.hm=0.2024, n1=291,
+   n2=334)
```

Fig 5.

doi:10.1371/journal.pone.0131499.g002

There are several errors in the "Comparison of Two Overlapping Correlation Based on Dependent Groups" subsection of the Code Examples section. The publisher apologizes for the error. Please view the correct code here. Figs 6 and 7

```
R> cocor(~knowledge + intelligence.a | logic +
+  intelligence.a, aptitude[["sample1"]])

  Results of a comparison of two overlapping correlations
  based on dependent groups

Comparison between r.jk (intelligence.a, knowledge) = 0.1038
  and r.jh (intelligence.a, logic) = 0.3213
Difference: r.jk - r.jh = -0.2175
Related correlation: r.kh = 0.0257
Data: aptitude[["sample1"]]: j = intelligence.a,
  k = knowledge, h = logic
Group size: n = 291
Null hypothesis: r.jk is equal to r.jh
Alternative hypothesis: r.jk is not equal to r.jh (two-sided)
Alpha: 0.05

pearson1898: Pearson and Filon's z (1898)
  z = -2.7914, p-value = 0.0052
  Null hypothesis rejected

hotelling1940: Hotelling's t (1940)
  t = -2.8066, df = 288, p-value = 0.0053
  Null hypothesis rejected

williams1959: Williams' t (1959)
  t = -2.7743, df = 288, p-value = 0.0059
  Null hypothesis rejected

olkin1967: Olkin's z (1967)
  z = -2.7914, p-value = 0.0052
  Null hypothesis rejected

dunn1969: Dunn and Clark's z (1969)
  z = -2.7595, p-value = 0.0058
  Null hypothesis rejected

hendrickson1970: Hendrickson, Stanley, and Hills' (1970)
modification of Williams' t (1959)
  t = -2.8065, df = 288, p-value = 0.0053
  Null hypothesis rejected

steiger1980: Steiger's (1980) modification of Dunn and
Clark's z (1969) using average correlations
  z = -2.7513, p-value = 0.0059
  Null hypothesis rejected

meng1992: Meng, Rosenthal, and Rubin's z (1992)
  z = -2.7432, p-value = 0.0061
  Null hypothesis rejected
  95% confidence interval for r.jk - r.jh: -0.3925 -0.0654
  Null hypothesis rejected (Interval does not include 0)

hittner2003: Hittner, May, and Silver's (2003) modification
of Dunn and Clark's z (1969) using a backtransformed average
Fisher's (1921) Z procedure
  z = -2.7505, p-value = 0.0059
  Null hypothesis rejected

zou2007: Zou's (2007) confidence interval
  95% confidence interval for r.jk - r.jh: -0.3689 -0.0630
  Null hypothesis rejected (Interval does not include 0)
```

Fig 6.

doi:10.1371/journal.pone.0131499.g003

```
R> cocor.dep.groups.overlap(r.jk=0.1038, r.jh=0.3213,
+   r.kh=0.0257, n=291)
```

Fig 7.

doi:10.1371/journal.pone.0131499.g004

There are several errors in the "Comparison of Two Nonoverlapping Correlations Based on Dependent Groups" subsection of the Code Examples section. The publisher apologizes for the error. Please view the correct code here. Figs 8 and 9

```
R> cocor(~knowledge + intelligence.a | logic +
+  intelligence.b, aptitude[["sample1"]])

  Results of a comparison of two nonoverlapping correlations
  based on dependent groups

Comparison between r.jk (knowledge, intelligence.a) = 0.1038
  and r.hm (logic, intelligence.b) = 0.2679
Difference: r.jk - r.hm = -0.164
Related correlations: r.jh = 0.0257, r.jm = 0.1713,
  r.kh = 0.3213, r.km = 0.4731
Data: aptitude[["sample1"]]: j = knowledge,
  k = intelligence.a, h = logic, m = intelligence.b
Group size: n = 291
Null hypothesis: r.jk is equal to r.hm
Alternative hypothesis: r.jk is not equal to r.hm (two-sided)
Alpha: 0.05

pearson1898: Pearson and Filon's z (1898)
  z = -2.0998, p-value = 0.0357
  Null hypothesis rejected

dunn1969: Dunn and Clark's z (1969)
  z = -2.0811, p-value = 0.0374
  Null hypothesis rejected

steiger1980: Steiger's (1980) modification of Dunn and
Clark's z (1969) using average correlations
  z = -2.0755, p-value = 0.0379
  Null hypothesis rejected

raghunathan1996: Raghunathan, Rosenthal, and Rubin's (1996)
modification of Pearson and Filon's z (1898)
  z = -2.0811, p-value = 0.0374
  Null hypothesis rejected

silver2004: Silver, Hittner, and May's (2004) modification
of Dunn and Clark's z (1969) using a backtransformed average
Fisher's (1921) Z procedure
  z = -2.0753, p-value = 0.0380
  Null hypothesis rejected

zou2007: Zou's (2007) confidence interval
  95% confidence interval for r.jk - r.hm: -0.3162 -0.0095
  Null hypothesis rejected (Interval does not include 0)
```

Fig 8.

doi:10.1371/journal.pone.0131499.g005

```
R> cocor.dep.groups.nonoverlap(r.jk = 0.1038, r.hm = 0.2679,
+  r.jh = 0.0257, r.jm = 0.1713, r.kh = 0.3213,
+  r.km = 0.4731, n=291)
```

Fig 9.

doi:10.1371/journal.pone.0131499.g006

In the Supporting Information file S1 Appendix, there are errors in Equations 4, 32, and 51. These equations should contain a "+" sign before the square root sign instead of a "-" sign. The publisher apologizes for the error. Please view the correct S1 Appendix below.

Supporting Information

S1 Appendix. Documentation of All Tests Implemented in cocor.
(PDF)

Reference

1. Diedenhofen B, Musch J (2015) cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. PLoS ONE 10(4): e0121945. doi: 10.1371/journal.pone.0121945

S1 Appendix. Documentation of All Tests Implemented in cocor

This Appendix is part of the article *cocor: A Comprehensive Solution for the Statistical Comparison of Correlations* by Birk Diedenhofen[1] and Jochen Musch published in PLOS ONE. In the following, the formulae of all tests implemented in the R package [1] `cocor` (version 1.1-0) are provided. z statistics are based on a normal distribution, whereas t statistics rely on a Student's t-distribution with given degrees of freedom. Some tests make use of Fisher's [2, p 26] r-to-Z transformation:

$$Z = \frac{1}{2}(ln(1+r) - ln(1-r)). \tag{1}$$

Tests for Comparison of Two Correlations Based on Independent Groups

The function `cocor.indep.groups()` implements tests for the comparison of two correlations based on independent groups.

fisher1925: Fisher's [3] z

This significance test was first described by Fisher [3, pp 161–168] and its test statistic z is calculated as

$$z = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}. \tag{2}$$

Z_1 and Z_2 are the two Z transformed correlations that are being compared. n_1 and n_2 specify the size of the two groups the correlations are based on. Equation 2 is also given for example in Peters and van Voorhis [4, p 188] and Cohen, Cohen, West, and Aiken [5, p 49, formula 2.8.11].

zou2007: Zou's [6] confidence interval

This test calculates the confidence interval of the difference between the two correlation coefficients r_1 and r_2. If the confidence interval includes zero, the null hypothesis that the two correlations are equal must be retained. If the confidence interval does not include zero, the null hypothesis has to be rejected. A lower and upper bound for the interval (L and U, respectively) is given by

$$L = r_1 - r_2 - \sqrt{(r_1 - l_1)^2 + (u_2 - r_2)^2} \tag{3}$$

[1]corresponding author, e-mail: birk.diedenhofen@uni-duesseldorf.de

and

$$U = r_1 - r_2 + \sqrt{(u_1 - r_1)^2 + (r_2 - l_2)^2} \tag{4}$$

[6, p 409]. A lower and upper bound for the confidence interval of r_1 (l_1 and u_1) and r_2 (l_2 and u_2) are calculated as

$$l = \frac{exp(2l') - 1}{exp(2l') + 1}, \tag{5}$$

$$u = \frac{exp(2u') - 1}{exp(2u') + 1} \tag{6}$$

[6, p 406], where

$$l', u' = Z \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n-3}} \tag{7}$$

[6, p 406]. α denotes the desired alpha level of the confidence interval, whereas n specifies the size of the group the correlation is based on.

Tests for Comparison of Two Overlapping Correlations Based on Dependent Groups

The function `cocor.dep.groups.overlap()` implements tests for the comparison of two overlapping correlations based on dependent groups. In the following, r_{jk} and r_{jh} are the two correlations that are being compared; Z_{jk} and Z_{jh} are their Z transformed equivalents. r_{kh} is the related correlation that is additionally required. n specifies the size of the group the two correlations are based on.

pearson1898: Pearson and Filon's [7] z

This test was proposed by Pearson and Filon [7, p 259, formula xxxvii]. The test statistic z is computed as

$$z = \frac{\sqrt{n}(r_{jk} - r_{jh})}{\sqrt{(1 - r_{jk}^2)^2 + (1 - r_{jh}^2)^2 - 2k}} \tag{8}$$

[8, p 246, formula 4], where

$$k = r_{kh}(1 - r_{jk}^2 - r_{jh}^2) - \frac{1}{2}(r_{jk} r_{jh})(1 - r_{jk}^2 - r_{jh}^2 - r_{kh}^2) \tag{9}$$

[8, p 245, formula 3].

hotelling1940: Hotelling's [9] t

The test statistic t is given by

$$t = \frac{(r_{jk} - r_{jh})\sqrt{(n-3)(1+r_{kh})}}{\sqrt{2|R|}} \tag{10}$$

[9, p 278, formula 7] with $df = n - 3$, where

$$|R| = 1 + 2r_{jk}r_{jh}r_{kh} - r_{jk}^2 - r_{jh}^2 - r_{kh}^2 \tag{11}$$

[9, p 278]. Equation 10 is also given in Steiger [8, p 246], Glass and Stanley [10, p 311, formula 15.7], and Hittner et al. [11, p 152].

williams1959: Williams' [12] t

This test is a modification of Hotelling's [9] t and was suggested by Williams [12]. Two mathematically different formulae for Williams' t can be found in the literature [11, p 152]. This is the version that Hittner et al. [11, p 152] labeled as "standard Williams' t":

$$t = (r_{jk} - r_{jh})\sqrt{\frac{(n-1)(1+r_{kh})}{2(\frac{n-1}{n-3})|R| + \bar{r}^2(1-r_{kh})^3}} \tag{12}$$

with $df = n - 3$, where

$$\bar{r} = \frac{r_{jk} + r_{jh}}{2} \tag{13}$$

and

$$|R| = 1 + 2r_{jk}r_{jh}r_{kh} - r_{jk}^2 - r_{jh}^2 - r_{kh}^2. \tag{14}$$

An alternative formula for Williams' t – termed as "Williams' modified t per Hendrickson, Stanley, and Hills" [13] by Hittner et al. [11, p 152] – is implemented in `cocor` as `hendrickson1970` (see Equation 18 below). Equation 12 is also given in Steiger [8, p 246, formula 7] and Neill and Dunn [14, p 533].

Results from Equation 12 are in accordance with the results of DEPCORR [15] and DEPCOR [16]. However, we found several typographical errors in formulae that also claim to compute Williams' t. For example, the formula reported by Boyer, Palachek, and Schucany [17, p 76] contains an error because

the term $(1 - r_{rk})$ is not being cubed. There are also typographical errors in the formula described by Hittner et al. [11, p 152]. For example, $r_{jk} - r_{jh}$ is divided instead of being multiplied by the square root term, and in the denominator of the fraction in the square root term, there are additional parentheses so that the whole denominator is multiplied by 2. These same errors can also be found in Wilcox and Tian [18, p 107, formula 1].

olkin1967: Olkin's [19] z

In the original article by Olkin [19, p 112] and in Hendrickson et al. [13, p 190, formula 2], the reported formula contains a typographical error. Hendrickson and Collins [20, p 639] provide a corrected version. In the revised version, however, n in the enumerator is decreased by 1. The `cocor` package implements the corrected formula without the decrement. The formula implemented in `cocor` is used by Glass and Stanley [21, p 313, formula 14.19], Hittner et al. [11, p 152], and May and Hittner [22, p 259] [23, p 480]:

$$z = \frac{(r_{jk} - r_{jh})\sqrt{n}}{\sqrt{(1 - r_{jk}^2)^2 + (1 - r_{jh}^2)^2 - 2r_{kh}^3 - (2r_{kh} - r_{jk}r_{jh})(1 - r_{kh}^2 - r_{jk}^2 - r_{jh}^2)}}. \tag{15}$$

dunn1969: Dunn and Clark's [24] z

The test statistic z of this test is calculated as

$$z = \frac{(Z_{jk} - Z_{jh})\sqrt{n - 3}}{\sqrt{2 - 2c}} \tag{16}$$

[24, p 370, formula 15], where

$$c = \frac{r_{kh}(1 - r_{jk}^2 - r_{jh}^2) - \frac{1}{2}r_{jk}r_{jh}(1 - r_{jk}^2 - r_{jh}^2 - r_{kh}^2)}{(1 - r_{jk}^2)(1 - r_{jh}^2)} \tag{17}$$

[24, p 368, formula 8].

hendrickson1970: Hendrickson, Stanley, and Hills [13] modification of Williams' [12] t

This test is a modification of Hotelling's [9] t and was suggested by Williams [12]. Two mathematically different formulae of Williams' [12] t can be found in the literature. `hendrickson1970` is the version that Hittner et al. [11, p 152] labeled as "Williams' modified t per Hendrickson, Stanley, and Hills" [13].

An alternative formula termed as "standard Williams' t" by Hittner et al. [11, p 152] is implemented as `williams1959` (see Equation 12 above). The `hendrickson1970` formula can be found in Hendrickson et al. [13, p 193], May and Hittner [22, p 259] [23, p 480], and Hittner et al. [11, p 152]:

$$t = \frac{(r_{jk} - r_{jh})\sqrt{(n-3)(1+r_{kh})}}{\sqrt{2|R| + \frac{(r_{jk}-r_{jh})^2(1-r_{kh})^3}{4(n-1)}}}, \tag{18}$$

with $df = n - 3$. A slightly changed version of this formula was provided by Dunn and Clark [25, p 905, formula 1.2], but seems to be erroneous, due to an error in the denominator.

steiger1980: Steiger's [8] modification of Dunn and Clark's [24] z using average correlations

This test was proposed by Steiger [8] and is a modification of Dunn and Clark's [24] z. Instead of r_{jk} and r_{jh}, the mean of the two is used. The test statistic z is defined as

$$z = \frac{(Z_{jk} - Z_{jh})\sqrt{n-3}}{\sqrt{2-2c}} \tag{19}$$

[8, p 247, formula 14], where

$$\bar{r} = \frac{r_{jk} + r_{jh}}{2} \tag{20}$$

[8, p 247] and

$$c = \frac{r_{kh}(1-2\bar{r}^2) - \frac{1}{2}\bar{r}^2(1-2\bar{r}^2-r_{kh}^2)}{(1-\bar{r}^2)^2} \tag{21}$$

[8, p 247, formula 10; in the original article, there are brackets missing around the divisor].

meng1992: Meng, Rosenthal, and Rubin's [26] z

This test is based on the test statistic z,

$$z = (Z_{jk} - Z_{jh})\sqrt{\frac{n-3}{2(1-r_{kh})h}}, \tag{22}$$

[26, p 173, formula 1], where

$$h = \frac{1 - f\overline{r^2}}{1 - \overline{r^2}} \tag{23}$$

[26, p 173, formula 2],

$$f = \frac{1 - r_{kh}}{2(1 - \overline{r^2})} \tag{24}$$

(f must be ≤ 1) [26, p 173, formula 3], and

$$\overline{r^2} = \frac{r_{jk}^2 + r_{jh}^2}{2} \tag{25}$$

[26, p 173]. This test also constructs a confidence interval of the difference between the two correlation coefficients r_{jk} and r_{jh}:

$$L, U = Z_{jk} - Z_{jk} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{2(1 - r_{kh})h}{n - 3}} \tag{26}$$

[26, p 173, formula 4]. α denotes the desired alpha level of the confidence interval. If the confidence interval includes zero, the null hypothesis that the two correlations are equal must be retained. If the confidence interval does not include zero, the null hypothesis has to be rejected.

hittner2003: Hittner, May, and Silver's [11] modification of Dunn and Clark's [24] z using a backtransformed average Fisher's [2] Z procedure

The approach to backtransform averaged Fisher's [2] Zs was first proposed by Silver and Dunlap [27] and was applied to the comparison of overlapping correlations by Hittner et al. [11]. The test is based on Steiger's [8] approach. The test statistic z is calculated as

$$z = \frac{(Z_{jk} - Z_{jh})\sqrt{n - 3}}{\sqrt{2 - 2c}} \tag{27}$$

[11, p 153], where

$$c = \frac{r_{kh}(1 - 2\bar{r}_z^2) - \frac{1}{2}\bar{r}_z^2(1 - 2\bar{r}_z^2 - r_{kh}^2)}{(1 - \bar{r}_z^2)^2} \tag{28}$$

[11, p 153],

$$\bar{r}_z = \frac{exp(2\bar{Z} - 1)}{exp(2\bar{Z} + 1)} \tag{29}$$

[27, p 146, formula 4], and

$$\bar{Z} = \frac{Z_{jk} + Z_{jh}}{2} \tag{30}$$

[27, p 146].

zou2007: Zou's [6] confidence interval

This test calculates the confidence interval of the difference between the two correlation coefficients r_{jk} and r_{jh}. If the confidence interval includes zero, the null hypothesis that the two correlations are equal must be retained. If zero is outside the confidence interval, the null hypothesis has to be rejected. A lower and upper bound for the interval (L and U, respectively) is given by

$$L = r_{jk} - r_{jh} - \sqrt{(r_{jk} - l_1)^2 + (u_2 - r_{jh})^2 - 2c(r_{jk} - l_1)(u_2 - r_{jh})} \tag{31}$$

and

$$U = r_{jk} - r_{jh} + \sqrt{(u_1 - r_{jk})^2 + (r_{jh} - l_2)^2 - 2c(u_1 - r_{jk})(r_{jh} - l_2)} \tag{32}$$

[6, p 409], where

$$l = \frac{exp(2l') - 1}{exp(2l') + 1}, \tag{33}$$

$$u = \frac{exp(2u') - 1}{exp(2u') + 1} \tag{34}$$

[6, p 406],

$$c = \frac{(r_{kh} - \frac{1}{2} r_{jk} r_{jh})(1 - r_{jk}^2 - r_{jh}^2 - r_{kh}^2) + r_{kh}^3}{(1 - r_{jk}^2)(1 - r_{jh}^2)} \tag{35}$$

[6, p 409], and

$$l', u' = Z \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n-3}} \tag{36}$$

[6, p 406]. α denotes the desired alpha level of the confidence interval.

Tests for Comparison of Two Nonoverlapping Correlations Based on Dependent Groups

The function `cocor.dep.groups.nonoverlap()` implements tests for the comparison of two nonoverlapping correlations based on dependent groups. In the following, r_{jk} and r_{hm} are the two correlations that are being compared; Z_{jk} and Z_{hm} are their Z transformed equivalents. r_{jh}, r_{kh}, r_{jm}, and r_{km} are the related correlations that are also required. n specifies the size of the group the two correlations are based on.

pearson1898: Pearson and Filon's [7] z

This test was proposed by Pearson and Filon [7, p 262, formula xl]. The formula for the test statistic z is computed as

$$z = \frac{\sqrt{n}(r_{jk} - r_{hm})}{\sqrt{(1 - r_{jk}^2)^2 + (1 - r_{hm}^2)^2 - k}} \tag{37}$$

[28, p 179, formula 1], where

$$\begin{aligned} k = {} & (r_{jh} - r_{jk}r_{kh})(r_{km} - r_{kh}r_{hm}) + (r_{jm} - r_{jh}r_{hm})(r_{kh} - r_{jk}r_{jh}) \\ & + (r_{jh} - r_{jm}r_{hm})(r_{km} - r_{jk}r_{jm}) + (r_{jm} - r_{jk}r_{km})(r_{kh} - r_{km}r_{hm}) \end{aligned} \tag{38}$$

[28, p 179, formula 2]. The two formulae can also be found in Steiger [8, p 245, formula 2 and p. 246, formula 5].

dunn1969: Dunn and Clark's [24] z

The test statistic z of this test is calculated as

$$z = \frac{(Z_{jk} - Z_{hm})\sqrt{n-3}}{\sqrt{2 - 2c}} \tag{39}$$

[24, p 370, formula 15], where

$$\begin{aligned} c = \Big(& \frac{1}{2} r_{jk}r_{hm}(r_{jh}^2 + r_{jm}^2 + r_{kh}^2 + r_{km}^2) + r_{jh}r_{km} + r_{jm}r_{kh} \\ & - (r_{jk}r_{jh}r_{jm} + r_{jk}r_{kh}r_{km} + r_{jh}r_{kh}r_{hm} + r_{jm}r_{km}r_{hm}) \Big) \\ & \Big/ \Big((1 - r_{jk}^2)(1 - r_{hm}^2) \Big) \end{aligned} \tag{40}$$

[24, p 368, formula 9].

steiger1980: Steiger's [8] modification of Dunn and Clark's [24] z using average correlations

This test was proposed by Steiger [8] and is a modification of Dunn and Clark's [24] z. Instead of r_{jk} and r_{hm} the mean of the two is being used. The test statistic z is given by

$$z = \frac{(Z_{jk} - Z_{hm})\sqrt{n-3}}{\sqrt{2-2c}} \tag{41}$$

[8, p 247, formula 15], where

$$\bar{r} = \frac{r_{jk} + r_{hm}}{2} \tag{42}$$

[8, p 247] and

$$\begin{aligned} c = \Big(\frac{1}{2}\bar{r}^2(r_{jh}^2 + r_{jm}^2 + r_{kh}^2 + r_{km}^2) + r_{jh}r_{km} + r_{jm}r_{kh} \\ -(\bar{r}r_{jh}r_{jm} + \bar{r}r_{kh}r_{km} + r_{jh}r_{kh}\bar{r} + r_{jm}r_{km}\bar{r})\Big) \\ \Big/(1-\bar{r}^2)^2 \end{aligned} \tag{43}$$

[8, p 247, formula 11; in the original article, there are brackets missing around the divisor].

raghunathan1996: Raghunathan, Rosenthal, and Rubin's [28] modification of Pearson and Filon's [7] z

This test of Raghunathan et al. [28] is based on Pearson and Filon's [7] z. Unlike Pearson and Filon [7], Raghunathan et al. [28] use Z transformed correlation coefficients. The test statistic z is computed as

$$z = \sqrt{\frac{n-3}{2}}\frac{Z_{jk} - Z_{hm}}{\sqrt{1 - \frac{k}{2(1-r_{jk}^2)(1-r_{hm}^2)}}} \tag{44}$$

[28, p 179, formula 3], where

$$\begin{aligned} k = (r_{jh} - r_{jk}r_{kh})(r_{km} - r_{kh}r_{hm}) + (r_{jm} - r_{jh}r_{hm})(r_{kh} - r_{jk}r_{jh}) \\ +(r_{jh} - r_{jm}r_{hm})(r_{km} - r_{jk}r_{jm}) + (r_{jm} - r_{jk}r_{km})(r_{kh} - r_{km}r_{hm}) \end{aligned} \tag{45}$$

[28, p 179, formula 2].

silver2004: Silver, Hittner, and May's [29] modification of Dunn and Clark's [24] z using a backtransformed average Fisher's [2] Z procedure

The approach to backtransform averaged Fisher's [2] Zs was first proposed in Silver and Dunlap [27] and was applied to the comparison of nonoverlapping correlations by Silver et al. [29]. The test is based on Steiger's [8] approach. The formula of the test statistic z is given by

$$z = \frac{(Z_{jk} - Z_{hm})\sqrt{n-3}}{\sqrt{2-2c}} \tag{46}$$

[29, p 55, formula 5], where

$$\begin{aligned} c = \Big(\frac{1}{2}\bar{r}_z^2(r_{jh}^2 + r_{jm}^2 + r_{kh}^2 + r_{km}^2) + r_{jh}r_{km} + r_{jm}r_{kh} \\ -(\bar{r}_z r_{jh} r_{jm} + \bar{r}_z r_{kh} r_{km} + r_{jh} r_{kh} \bar{r}_z + r_{jm} r_{km} \bar{r}_z)\Big) \\ \Big/(1 - \bar{r}_z^2)^2 \end{aligned} \tag{47}$$

[29, p 56],

$$\bar{r}_z = \frac{exp(2\bar{Z} - 1)}{exp(2\bar{Z} + 1)} \tag{48}$$

[27, p 146, formula 4], and

$$\bar{Z} = \frac{Z_{jk} + Z_{hm}}{2} \tag{49}$$

[29, p 55].

zou2007: Zou's [6] confidence interval

This test calculates the confidence interval of the difference between the two correlations r_{jk} and r_{hm}. If the confidence interval includes zero, the null hypothesis that the two correlations are equal must be retained. If the confidence interval does not include zero, the null hypothesis has to be rejected. A lower and upper bound for the interval (L and U, respectively) is given by

$$L = r_{jk} - r_{hm} - \sqrt{(r_{jk} - l_1)^2 + (u_2 - r_{hm})^2 - 2c(r_{jk} - l_1)(u_2 - r_{hm})} \tag{50}$$

and

$$U = r_{jk} - r_{hm} + \sqrt{(u_1 - r_{jk})^2 + (r_{hm} - l_2)^2 - 2c(u_1 - r_{jk})(r_{hm} - l_2)} \tag{51}$$

[6, pp 409–410], where

$$l = \frac{exp(2l') - 1}{exp(2l') + 1}, \tag{52}$$

$$u = \frac{exp(2u') - 1}{exp(2u') + 1} \tag{53}$$

[6, p 406],

$$\begin{aligned} c = \Big(\frac{1}{2} r_{jk} r_{hm} (r_{jh}^2 + r_{jm}^2 + r_{kh}^2 + r_{km}^2) + r_{jh} r_{km} + r_{jm} r_{kh} \\ -(r_{jk} r_{jh} r_{jm} + r_{jk} r_{kh} r_{km} + r_{jh} r_{kh} r_{hm} + r_{jm} r_{km} r_{hm})\Big) \\ \Big/ \Big((1 - r_{jk}^2)(1 - r_{hm}^2)\Big) \end{aligned} \tag{54}$$

[6, p 409], and

$$l', u' = Z \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n-3}} \tag{55}$$

[6, p 406]. α denotes the desired alpha level of the confidence interval.

References

1. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2014. Available: `http://www.R-project.org`. Accessed 21 February 2015.

2. Fisher RA. On the Probable Error of a Coefficient of Correlation Deduced From a Small Sample. Metron. 1921;1: 3–32. Available: `http://hdl.handle.net/2440/15169`. Accessed 21 February 2015.

3. Fisher RA. Statistical Methods for Research Workers. Edinburgh, Scotland: Oliver and Boyd; 1925. Available: `http://psychclassics.yorku.ca`. Accessed 21 February 2015.

4. Peters CC, van Voorhis WR. Statistical Procedures and Their Mathematical Bases. New York, NJ: McGraw-Hill; 1940.

5. Cohen J, Cohen P, West SG, Aiken LS. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. 3rd ed. Mahwah, NJ: Erlbaum; 2003.

6. Zou GY. Toward Using Confidence Intervals to Compare Correlations. Psychol Methods. 2007;12: 399–413. doi: 10.1037/1082-989X.12.4.399

7. Pearson K, Filon LNG. Mathematical Contributions to Theory of Evolution: IV. On the Probable Errors of Frequency Constants and on the Influence of Random Selection and Correlation. Philos Trans R Soc Lond A. 1898;191: 229–311. doi: 10.1098/rsta.1898.0007

8. Steiger JH. Tests for Comparing Elements of a Correlation Matrix. Psychol Bull. 1980;87: 245–251. doi: 10.1037//0033-2909.87.2.245

9. Hotelling H. The Selection of Variates for Use in Prediction, with Some Comments on the General Problem of Nuisance Parameters. Ann Math Stat. 1940;11: 271–283. doi: 10.1214/aoms/1177731867

10. Glass GV, Stanley JC. Statistical Methods in Education and Psychology. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall; 1984.

11. Hittner JB, May K, Silver NC. A Monte Carlo Evaluation of Tests for Comparing Dependent Correlations. J Gen Psychol. 2003;130: 149–168. doi: 10.1080/00221300309601282

12. Williams EJ. The Comparison of Regression Variables. J R Stat Soc B. 1959;21: 396–399. Available: http://www.jstor.org/stable/2983809. Accessed 21 February 2015.

13. Hendrickson GF, Stanley JC, Hills JR. Olkin's New Formula for Significance of r13 vs. r23 Compared with Hotelling's Method. Am Educ Res J. 1970;7: 189–195. doi: 10.2307/1162159

14. Neill JJ, Dunn OJ. Equality of Dependent Correlation Coefficients. Biometrics. 1975;31: 531–543. doi: 10.2307/2529435

15. Hittner JB, May K. DEPCORR: A SAS Program for Comparing Dependent Correlations. Appl Psychol Meas. 1998;22: 93–94. doi: 10.1177/01466216980221010

16. Silver NC, Hittner JB, May K. A FORTRAN 77 Program for Comparing Dependent Correlations. Appl Psychol Meas. 2006;30: 152–153. doi: 10.1177/0146621605277132

17. Boyer IE, Palachek AD, Schucany WR. An Empirical Study of Related Correlation Coefficients. J Educ Stat. 1983;8: 75–86. doi: 10.2307/1164871

18. Wilcox RR, Tian T. Comparing Dependent Correlations. J Gen Psychol. 2008;135: 105–112. doi: 10.3200/GENP.135.1.105-112

19. Olkin I. Correlations Revisited. In: Stanley JC, editor. Improving Experimental Design and Statistical Analysis. Chicago, IL: Rand McNally; 1967. pp. 102–128.

20. Hendrickson GF, Collins JR. Note Correcting the Results in 'Olkin's New Formula for the Significance of r13 vs. r23 Compared with Hotelling's Method'. Am Educ Res J. 1970;7: 639–641. doi: 10.2307/1161847

21. Glass GV, Stanley JC. Statistical Methods in Education and Psychology. Englewood Cliffs, NJ: Prentice-Hall; 1970.

22. May K, Hittner JB. A Note on Statistics for Comparing Dependent Correlations. Psychol Rep. 1997;80: 475–480. doi: 10.2466/pr0.1997.80.2.475

23. May K, Hittner JB. Tests for Comparing Dependent Correlations Revisited: A Monte Carlo Study. J Exp Educ. 1997;65: 257–269. doi: 10.1080/00220973.1997.9943458

24. Dunn OJ, Clark VA. Correlation Coefficients Measured on the Same Individuals. J Am Stat Assoc. 1969;64: 366–377. doi: 10.2307/2283746

25. Dunn OJ, Clark VA. Comparison of Tests of the Equality of Dependent Correlation Coefficients. J Am Stat Assoc. 1971;66: 904–908. doi: 10.2307/2284252

26. Meng XL, Rosenthal R, Rubin DB. Comparing Correlated Correlation Coefficients. Psychol Bull. 1992;111: 172–175. doi: 10.1037//0033-2909.111.1.172

27. Silver NC, Dunlap WP. Averaging Correlation Coefficients: Should Fisher's Z Transformation Be Used? J Appl Psychol. 1987;72: 146–148. doi: 10.1037//0021-9010.72.1.146

28. Raghunathan TE, Rosenthal R, Rubin DB. Comparing Correlated but Nonoverlapping Correlations. Psychol Methods. 1996;1: 178–183. doi: 10.1037//1082-989X.1.2.178

29. Silver NC, Hittner JB, May K. Testing Dependent Correlations with Nonoverlapping Variables: A Monte Carlo Simulation. J Exp Educ. 2004;73: 53–69. doi: 10.3200/JEXE.71.1.53-70

Erklärung zum eigenen Beitrag

Artikel 1

Diedenhofen, B. & Musch, J. (2015). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment*. Advance online publication. doi: 10.1027/1015-5759/a000295

Ich war maßgeblich für die Planung, Durchführung und Auswertung der Studie sowie für das Schreiben des Manuskriptes verantwortlich.

Artikel 2

Diedenhofen, B. & Musch, J. (2015). Option weights should be determined empirically and not by experts when assessing knowledge using multiple-choice items. Manuscript submitted for publication.

Ich war maßgeblich für die Planung, Durchführung und Auswertung der Studie sowie für das Schreiben des Manuskriptes verantwortlich.

Artikel 3

Diedenhofen, B. & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*(4), e0121945. doi: 10.1371/journal.pone.0121945

Ich war maßgeblich für die Programmierung der Software und das Schreiben des Manuskriptes verantwortlich.

Düsseldorf, den 18.12.2015	Düsseldorf, den 18.12.2015
Birk Diedenhofen	Prof. Dr. Jochen Musch (Bestätigung des Betreuers)

Versicherung an Eides Statt

Hiermit versichere ich an Eides Statt, dass die Dissertation mit dem Titel „Untersuchungen zur Optionsgewichtung als Methode für die Erfassung von Teilwissen in Multiple-Choice-Tests" von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf erstellt worden ist. Ferner versichere ich, dass die Arbeit in der vorgelegten oder in ähnlicher Form bisher bei keiner anderen Fakultät als Dissertation eingereicht wurde und dass ich bisher keine erfolglosen Promotionsversuche unternommen habe.

Düsseldorf, den 18.12.2015

Birk Diedenhofen