GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Sozial vermittelte Lernprozesse bei quantitativen Schätzaufgaben

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
„Doctor rerum naturalium“
der Georg-August-Universität Göttingen

im Promotionsprogramm Biologie
der Georg-August University School of Science (GAUSS)

vorgelegt von
Alexander Stern
aus Norderney

Göttingen, März 2017

Betreuungsausschuss

Prof. Dr. Stefan Schulz-Hardt, Abteilung Wirtschafts- und Sozialpsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Margarete Boos, Abteilung Sozial- und Kommunikationspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Mitglieder der Prüfungskommission

Referent: Prof. Dr. Stefan Schulz-Hardt, Abteilung Wirtschafts- und Sozialpsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Korreferentin: Prof. Dr. Margarete Boos, Abteilung Sozial- und Kommunikationspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Nivedita Mani, Abteilung Psychologie der Sprache, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Lars Penke, Abteilung Biologische Persönlichkeitspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Annekathrin Schacht, Abteilung Affektive Neurowissenschaft und Psychophysiologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

PD. Dr. Micha Strack, Abteilung Sozial- und Kommunikationspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 11. April 2017

Danksagung:

Mein aufrichtiger Dank gilt zunächst dem Erstbetreuer meiner Promotion, Stefan Schulz-Hardt, von dessen großem Sachverstand, Rat und Kritik ich stets sehr profitieren konnte. Er hat mich nicht nur bei der Abfassung der hier eingereichten Schriften unterstützt, sondern mir auch durch sein Vertrauen Freiheiten eröffnet, die zu meiner beruflichen und persönlichen Entwicklung beigetragen haben.

Ein ganz besonderer Dank gilt auch meiner Zweitbetreuerin, Margarete Boos, die kurzfristig und unbürokratisch dieses Amt übernommen hat und mir damit die Möglichkeit gegeben hat, mich voll auf den Abschluss meiner Promotion zu konzentrieren.

Bei meinen Kolleginnen und Kollegen möchte ich mich für die anregenden Diskussionen und das stets angenehme und freundschaftliche Arbeitsverhältnis bedanken. Besonders hervorheben möchte ich an dieser Stelle Thomas Schultze, dessen fachliche Expertise mich sehr geprägt hat und dessen Tür mir immer offen stand.

Mein tiefer Dank gilt natürlich auch meiner Familie. Meine Eltern haben mir immer vertraut und mich in meinen Entscheidungen unterstützt. Auch in anstrengenden Phasen habe ich bei ihnen stets ein Umfeld der Ruhe gefunden und konnte mich in jeder Situation voll auf sie verlassen.

Natürlich muss ich mich an dieser Stelle auch bei meinem gesamten Freundeskreis bedanken. Hier konnte ich immer wieder den nötigen Abstand von der Arbeit gewinnen, um mich im Anschluss wieder auf meine beruflichen Aufgaben konzentrieren zu können.

Abschließend ein paar Worte zu Julia. Sie ist stets ohne jegliche Bedingung für mich da und hat mich auch in schwierigeren Situationen motivierend begleitet. Ihr gebührt mein besonderer Dank.

Inhaltsverzeichnis

Einleitung

Die zwei im Rahmen dieser Dissertation eingereichten Schriften befassen sich mit sozialen Lernprozessen im Gruppenkontext sowie in Ratschlagssituationen bei quantitativen Schätzaufgaben. Es handelt sich hierbei nicht um eine publikationsbasierte Dissertation, ist aber einer solchem im Aufbau nachempfunden. Im Fokus beider Arbeiten steht die Frage, ob Individuen in der Lage sind, durch die Interaktion mit anderen Personen aufgabenrelevante Fertigkeiten zu erlernen, welche ihre individuelle Leistung verbessern. Die Manuskripte leisten einerseits eine systematische, kritische Analyse und Bewertung aktueller empirischer Befunde zu sozialen Lernprozessen und andererseits eine laborexperimentelle Überprüfung daraus hervorgehender Hypothesen.

Stern, A., Schultze, T & Schulz-Hardt, S. (2017a). How much group is necessary? Group-to-individual transfer in estimation tasks. Unpublished manuscript.

Stern, A., Schultze, T & Schulz-Hardt, S. (2017b). Social learning in the judge-advisor-system: A neglected advantage of advice-taking. Under review at: *Journal of Personality and Social Psychology*.

Bei besonders wichtig erscheinenden Aufgaben sichern sich Menschen gerne sozial ab. So werden vor allem folgenreiche Entscheidungen in Politik und Wirtschaft häufig von Gremien getroffen, Finanz- oder Klimaprognosen von Expertinnen- und Expertengruppen erstellt und beim Treffen einer wichtigen Investitionsentscheidung Ratschläge eingeholt. Ein naheliegender Grund hierfür könnte sein, dass mehreren Menschen auch mehr intellektuelle Ressourcen zur Verfügung stehen, was sie in die Lage versetzt, besonders gute Urteile abzugeben oder Entscheidungen zu fällen (vgl. Schulz-Hardt, Greitemeyer, Brodbeck & Frey, 2002). Des Weiteren könnte diese soziale Absicherung erfolgen, um die Verantwortung für das eigene Urteil zu teilen (vgl. Harvey & Fischer, 1997). Neben diesen Motiven für das Arbeiten in Gruppen oder das Einholen von Ratschlägen können Menschen allerdings auch noch auf einer weiteren Ebene von diesen sozialen Interaktionen profitieren. Die Zusammenarbeit kann dazu führen, dass aufgabenrelevante Fertigkeiten vermittelt werden, von denen das Individuum bei zukünftiger Aufgabenausführung profitieren kann. Im Folgenden sollen solche Lerngewinne im Rahmen von Gruppenarbeit und Ratschlagssituationen diskutiert und laborexperimentell untersucht werden. Dazu wird mit *diskretionären Aufgaben* (Steiner, 1972) ein – in der Gruppenforschung – bisher eher wenig beforschter Aufgabentyp verwendet. Beispiele für diesen Aufgabentyp sind quantitative Schätz- und Prognoseaufgaben. Diskretionäre Aufgaben zeichnen sich dadurch aus, dass es im Ermessen der Gruppenmitglieder liegt, wie sie die individuellen Beiträge ihrer Mitglieder zu einem Gruppenprodukt kombinieren. So kann eine Gruppe, die die zukünftige Entwicklung des Bruttosozialprodukts vorhersagen soll, zum Beispiel den Mittelwert aller individuellen Prognosen bilden, die Vorhersage eines einzelnen Gruppenmitglieds wählen oder die Einzelbeiträge in Abhängigkeit der subjektiven Expertise der Gruppenmitglieder unterschiedlich stark für das Gruppenurteil gewichten. Dieser Aufgabentyp eignet sich genau deshalb für die Untersuchung von Lerneffekten, weil bei diskretionären Aufgaben zum einen die Qualität eines Urteils nicht in jedem Fall gut demonstrierbar ist und zum anderen Veränderungen in der individuellen Leistung graduell bestimmt werden können. Ziel solcher quantitativer Schätzaufgaben ist dementsprechend nicht eine distinkte Unterscheidung zwischen richtig und falsch, sondern vielmehr, mit einer Schätzung dem wahren Wert möglichst nahe zu kommen. Im Folgenden wird zunächst dargelegt, unter welchen Umständen es zu Lernprozessen in Gruppen kommen kann und wie diese Prozesse neben den individuellen Ressourcen der Gruppenmitglieder auch die Koordination innerhalb der Gruppe fördern können. Anschließend werden die daraus abgeleiteten Hypothesen laborexperimentell überprüft. Im zweiten Teil des Forschungsprogramms wird die Möglichkeit von vergleichbaren Lerngewinnen als Folge von Ratschlägen diskutiert und ebenfalls empirisch untersucht. In anderen Worten soll überprüft werden, ob es auch durch den reinen Austausch numerischer Informationen zu individuellen Leistungsverbesserungen kommen kann und inwiefern diese von der Ratschlagsqualität abhängen. Abschließend werden die Ergebnisse beider Forschungsbereiche zusammenfassend bewertet und diskutiert.

Forschungsprogramm Teil I

Bei kollektiven quantitativen Schätzaufgaben gelingt es Gruppen häufig, Vergleichsindividuen zu übertreffen (vgl. Bonner & Baumann, 2012; Bonner, Sillito & Baumann, 2007; Laughlin, Bonner, Miner & Carnevale, 1999) oder sogar die Leistung ihres besten Gruppenmitglieds zu erreichen (vgl. Einhorn, Hogarth & Klempner, 1977; Laughlin, Gonzalez & Sommer, 2003). Gruppen können also eine höhere Leistung erreichen, als es die reine Ausmittelung individueller Fehler ihrer Gruppenmitglieder vermuten lässt. Nach Brodbeck und Greitemeyer (2000) können diese Leistungsverbesserungen in zwei Kategorien des Gruppenlernens unterteilt werden. Einerseits können die Gruppenmitglieder im Laufe der Zeit lernen, besser in einer Gruppe zu kooperieren (learning to collaborate, Brodbeck & Greitemeyer, 2000), andererseits können Gruppenmitglieder durch die Interaktion mit anderen Handlungsstrategien erlernen, welche ihre individuellen Leistungen verbessern (learning to perform the task). Diese Steigerung der individuellen Ressourcen wird in der Gruppenleistungsforschung als *group-to-individual Transfer* (G-I Transfer, vgl. Laughlin & Barth, 1981; Laughlin & Sweeney, 1977) bezeichnet und kann sich wiederum positiv auf die Gruppenleistung auswirken. In anderen Worten tritt G-I Transfer dann auf wenn Gruppeninteraktionen dazu führen, dass die Produktivität oder Effizienz ihrer Gruppenmitglieder bei anschließenden Aufgaben steigen. In diesem Zusammenhang ist es wichtig, zwei Arten von G-I Transfer zu unterscheiden: zum einen *spezifischer* Transfer bei dem der positive Effekt der Gruppeninteraktion auf *dieselbe Aufgabe* beschränkt ist, und zum anderen *genereller* Transfer von einer Aufgabe zu einer *anderen Aufgabe der gleichen Art* (Stasson, Kameda, Parks, Zimmerman & Davis, 1991).

Während die meiste Forschung zu G-I Transfer bei Problemlöseaufgaben wie mathematischen Problemen durchgeführt worden ist und dabei sowohl Evidenz für spezifischen G-I Transfer (z.B. Laughlin & Ellis, 1986) als auch generellen Transfer (z.B. Laughlin, Carey & Kerr, 2008; Stasson et al., 1991) gefunden werden konnte, ist die wissenschaftliche Evidenz bei quantitativen Schätzaufgaben eher spärlich. Dem Autor ist nur eine Studie bekannt, die solche Aufgaben mit einem Design verbindet, welches die Überprüfung von individuellen Lerneffekten als Folge vorheriger Gruppeninteraktion zulässt. Schultze, Mojzisch und Schulz-Hardt (2012) konnten dabei nachweisen, dass sich die individuelle Leistung schlechterer Gruppenmitglieder bereits nach der ersten Gruppeninteraktion stark verbessert, was als Nachweis für einen generellen G-I Transfer gesehen werden kann, da die Autoren mit unterschiedlichen Aufgaben der gleichen Art arbeiteten. Darüber hinaus zeigte sich, dass die Gruppenmitglieder von weiteren Gruppeninteraktionen nicht zusätzlich profitieren, was darauf hindeutet, dass bei der verwendeten Schätzaufgabe eine einmalige Gruppeninteraktion für das Auftreten eines G-I Transfers ausreicht.

Obwohl die Befunde von Schultze, Mojzisch und Schulz-Hardt (2012) als erster sauberer Nachweis eines generellen G-I Transfers bei quantitativen Schätzaufgaben gesehen werden können, lassen die Ergebnisse zwei wichtige Fragen unbeantwortet. Erstens bleibt unklar, *was* Gruppenmitglieder durch die Interaktion miteinander lernen. Aufbauend auf der Argumentation von Brown und Siegler (1993) kann das Wissen bei quantitativen

Schätzaufgaben in zwei Komponenten unterteilt werden. Zum einen spiegelt das sogenannte *metric knowledge* ein grundsätzliches Verständnis für die Skalierung der Aufgabe wider. In anderen Worten gibt das metric knowledge an, ob die betreffende Person eine genaue Vorstellung davon hat, welcher Schätzbereich plausibel ist. Zum anderen bestimmt das *mapping knowledge*, ob Menschen Verständnis für die Rangordnung verschiedener zu schätzender Objekte haben. Mapping knowledge erlaubt also, verschiedene Schätzwerte in die richtige Reihenfolge zu bringen (z.B. die Entfernung zwischen London und Rom ist größer als die Entfernung zwischen London und Paris). Zweitens sollte geklärt werden, ob die nachgewiesenen individuellen Leistungsverbesserungen auch stabil sind, wenn die Gruppe nach einmaliger Interaktion wieder aufgelöst wird. Somit sollte überprüft werden, ob es zur Aufrechterhaltung dieses Wissenstransfers der Gruppe bedarf, oder ob die volle Leistungsverbesserung auch ohne weitere Gruppenunterstützung erhalten bleibt. Über diese beiden zentralen Fragestellungen hinaus sollte explorativ überprüft werden, ob es zusätzlich zu differenziellen Gewichtungsstrategien, also einer stärkeren Gewichtung von kompetenteren Gruppenmitgliedern bei den Gruppenurteilen, kommt. In anderen Worten sollte zwischen einer Steigerung der individuellen Leistung (G-I Transfer) und einer besseren Kooperationsfähigkeit der Gruppenmitglieder unterschieden werden, die stärkere Gewichtung leistungsstärkerer Mitglieder bzw. akkuraterer Schätzungen zulässt. Auch wenn der Nachweis solcher Gewichtungsstrategien unter Kontrolle von individuellen Lerneffekten bisher fehlgeschlagen ist (vgl. Schultze et al., 2012), kann dieser Befund nicht zwingend generalisiert werden, da er nur bei einem Aufgabentyp erfolgt ist. Das Ziel der empirischen Studien des ersten Teils des Forschungsprogramms bestand darin, diese offenen Fragen zu klären.

Eine laborexperimentelle Überprüfung von Gruppenlernprozessen bei quantitativen Schätzaufgaben

Methode

In zwei Laborexperimenten wurde die Anzahl der Gruppeninteraktionen manipuliert, um gezielt die Stabilität des G-I Transfers zu überprüfen. Zu diesem Zweck arbeiteten die Gruppen entweder fortwährend oder nur einmalig miteinander und wurden einer nicht interagierenden Nominalgruppenbedingung gegenübergestellt. Im ersten Experiment sollten die Versuchspersonen die Luftlinienentfernung zwischen verschiedenen europäischen Hauptstädten schätzen und im zweiten das Gewicht verschiedener Gegenstände. Als wichtigste abhängige Variable wurde jeweils die gemittelte absolute prozentuale Abweichung vom wahren Wert jedes Trials erfasst. Dieser sogenannte MAPE ist ein gängiges Leistungsmaß in der Gruppenurteilsforschung (vgl. Sniezek & Henry, 1989, 1990).

Im ersten Experiment wurden 183 und im zweiten Experiment 252 Versuchspersonen randomisiert drei Versuchsbedingungen zugeteilt. Nach einer individuellen Übungsphase von 10 Aufgaben, die als Baselinemessung für spätere Leistungsversänderungen diente, arbeiteten die Versuchspersonen in Abhängigkeit der Experimentalbedingung entweder an 10 weiteren Aufgaben als Dreiergruppe, oder an nur einer Aufgabe als Dreiergruppe und bei den verbleibenden 9 Aufgaben wieder individuell, oder analog zur Übungsphase ohne Gruppeninteraktion. Der Ablauf der einzelnen Gruppeninteraktionen sah vor, dass die einzelnen Gruppenmitglieder zunächst individuelle Schätzungen abgeben sollten, danach ihre Individualschätzungen diskutieren und abschließend zu einem gemeinsamen Gruppenergebnis kommen sollten. Auf diese Weise war es möglich, individuelle Fertigkeitsgewinne im Laufe der Zeit zu bestimmen und von zusätzlichen Leistungsverbesserungen durch differenzielle Gewichtungsstrategien, also z.B. einer stärkeren Gewichtung besserer Gruppenmitglieder, abzugrenzen.

Ergebnisse

In beiden Experimenten trat eine starke Leistungsverbesserung schwächerer Gruppenmitglieder bereits nach der ersten Gruppeninteraktion auf, womit die generellen Befunde von Schultze et al. (2012) hinsichtlich individueller Fertigkeitsverbesserungen repliziert werden konnten. Ein Vergleich von Mitgliedern fortwährend interagierender Gruppen mit Mitgliedern einmalig interagierenden Gruppen zeigte, dass die individuelle Schätzgenauigkeit in gleichem Maße zunahm und diese gesteigerte individuelle Leistung auf diesem hohen Niveau blieb, auch wenn die Gruppe nach nur einer Interaktion aufgelöst wurde und die Versuchspersonen wieder individuell weiterarbeiteten. Diese Leistungsveränderung war ausschließlich auf verringerte metric errors zurückzuführen, während sich die mapping errors der Versuchspersonen nicht systematisch in Abhängigkeit der Versuchsbedingung veränderten. In der Nominalgruppenbedingung veränderte sich die Leistung der Versuchspersonen im Laufe der Zeit nicht. Diese Befunde traten unabhängig vom Aufgabentyp auf.

Bezüglich differenzieller Gewichtungsstrategien unterschieden sich die Ergebnisse zwischen den Experimenten 1 und 2. Während im ersten Experiment die Versuchspersonen ihre (z.T. verbesserten) individuellen Urteile im Durchschnitt relativ gleichmäßig gewichteten, um auf ein Gruppenurteil zu kommen, ergab sich im zweiten Experiment ein anderes Bild. Hier gelang der Nachweis einer differenziellen Gewichtung der Einzelbeiträge, also einer stärkeren Gewichtung leistungsstärkerer Mitglieder bzw. akkuraterer Schätzungen. Die Gruppenurteile waren akkurater als es eine reine Mittelung der verbesserten individuellen Schätzungen vorhersagen würde.

Fazit

Im ersten Teil des Forschungsprogramms testeten wir zum einen die zeitliche Stabilität von G-I Transfers und zum anderen, was Gruppenmitglieder durch die Interaktion miteinander lernen. Der in beiden Experimenten nachgewiesene Befund, dass sich die individuelle Urteilsgenauigkeit der Versuchspersonen bereits nach einmaliger Gruppeninteraktion verbessert und danach auf diesem hohen Niveau verbleibt, selbst wenn die Gruppe danach aufgelöst wird, kann vor dem Hintergrund von fehlenden individuellen Leistungsverbesserungen in der Nominalgruppenbedingung tatsächlich als Vermittlung von Handlungsstrategien im Sinne eines generellen G-I Transfers interpretiert werden. Die Tatsache, dass die Anzahl der Gruppeninteraktionen keinen Einfluss auf die Fertigkeitsgewinne hatte, spricht darüber hinaus für die Einfachheit dieses Transfers. Die wahrscheinlichste Erklärung für dieses Phänomen ist, dass Gruppenmitglieder während der ersten Gruppeninteraktion Referenzwerte austauschen (z.B. Deutschland von Nord nach Süd ist 900 Kilometer lang), was sich auch in dem Ergebnis widerspiegelt, dass Gruppeninteraktionen ausschließlich zu einer Reduzierung des metric error führt, ohne den mapping error systematisch zu verändern. Vorangegangene Befunde haben bereits die Wichtigkeit von Referenzwerten für die Verbesserung der Schätzgenauigkeit nachgewiesen (z.B. Bonner & Baumann, 2008; Bonner et al., 2007; Laughlin et al., 1999; Laughlin et al., 2003). Solche Referenzwerte sollten einfach zu kommunizieren sein und Informationen enthalten, welche den metric error und dadurch auch die Schätzgenauigkeit nachhaltig verbessern können. Anders ausgedrückt kann ein guter Maßstab den individuellen Schätzfehler reduzieren, was insbesondere bei schwächeren Gruppenmitgliedern zu einer sprunghaften Leistungsverbesserung führt. Da das Datenmuster bezüglich individueller Fertigkeitsgewinne und einer Reduzierung des metric error bei beiden Aufgabentypen gleich ausfiel, kann davon ausgegangen werden, dass diese Befunde auf andere Schätzaufgaben von ähnlichem Schwierigkeitsgrad generalisierbar sind.

Im Gegensatz dazu konnten nur im zweiten Experiment differenzielle Gewichtungsstrategien nachgewiesen werden. Vermutlich ist dieser Befund auf Unterschiede im Aufgabentyp zurückzuführen. Beim ersten Experiment wurden Aufgaben ohne populationsstereotypen Bias verwendet, d.h. die Versuchspersonen haben die wahren Werte nicht systematisch über- oder unterschätzt, sondern unsystematische Zufallsfehler produziert. Diese Situation könnte den Nachweis von differenziellen Gewichtungsstrategien dadurch erschweren, dass bei einer Aufgabe ohne Bias der Durchschnitt der individuellen Schätzungen der Gruppenmitglieder („average"-model) durch die Ausmittelung individueller Fehler bereits eine hohe Genauigkeit aufweist und der potenzielle Leistungszuwachs, der aus differenzieller Gewichtung resultieren könnte, deutlich eingeschränkt ist. Es könnten also theoretisch durchaus, auch bei einer Aufgabe ohne Bias, differenzielle Gewichtungsstrategien auftreten. Sie wären nur schwerer nachweisbar. Die Tatsache, dass in Experiment 1 die Genauigkeit des Gruppenurteils dem Durchschnitt der individuellen Schätzungen der Gruppenmitglieder aber sogar deskriptiv leicht unterlegen war, spricht zumindest dafür, dass falls eine differenzielle Gewichtung stattgefunden haben sollte, diese zumindest nicht funktional war. Bei Aufgaben mit Populationsbias können Gruppen hingegen stärker von differenziellen Gewichtungsstrategien profitieren, da in einem solchen Fall die besten Einzelschätzungen systematisch näher am wahren Wert liegen als das Mittelwertsmodell (vgl. Davis-Stober, Budescu, Dana & Bromwell, 2014; Einhorn et al., 1977). Folglich sollte eine größere Gewichtung besserer Einzelurteile das Gruppenergebnis stärker verbessern als bei einer Aufgabe ohne einen solchen Bias. Da die Aufgabe, mit der in Experiment 2 gearbeitet wurde, durch einen solchen Populationsbias charakterisiert war, könnte dies das Auftreten von funktionalen differenziellen Gewichtungsstrategien begünstigt haben. Zusammenfassend konnte gezeigt werden, dass Gruppenmitglieder durch die Interaktion mit Anderen Informationen austauschen, die ihren metric error reduzieren und damit ihre Schätzgenauigkeit sprunghaft und stabil nach nur einmaliger Gruppeninteraktion verbessern. Des Weiteren konnte nachgewiesen werden, dass Gruppen unter bestimmten Umständen funktionale, differenzielle Gewichtungsstrategien anwenden, die dazu führen, dass die Gruppenurteile besser sind als die einfache Mittelung der verbesserten individuellen Schätzungen.

Forschungsprogramm Teil II

Neben Gruppenarbeit spielt auch das Einholen von Ratschlägen eine wichtige Rolle in vielen Urteils- und Entscheidungsprozessen. Das Hauptmotiv dieser Suche nach Ratschlägen ist die Verbesserung der Urteilsqualität, weswegen Entscheidungsträger_innen auch am meisten Interesse an Ratschlägen von Expertinnen und Experten haben (vgl. Van Swol & Sniezek, 2005). Ratschlagssituation werden in der sozialpsychologischen Forschung häufig im

Rahmen des sogenannten „Judge Advisor System“ untersucht (JAS; Sniezek & Buckley, 1995; Sniezek & Van Swol, 2001), wobei zwischen einer Ratgeberin oder einem Ratgeber (Advisor), die oder der Ratschläge abgibt, und einer Entscheidungsträgerin oder einem Entscheidungsträger (Judge), die oder der die Verantwortung für ein Urteil hat, unterschieden wird. Der klassische Ablauf eines solchen JAS sieht vor, dass der Judge ein erstes Initialurteil abgibt, anschließend einen Ratschlag erhält, und abschließend ein finales, möglicherweise revidiertes Urteil fällt (z.B. Yaniv, 2004b; Sniezek, Schrah & Dalal, 2004). Dabei liegt der Fokus der Judge Advisor Forschung auf der Genauigkeit des finalen Urteils und der Ratschlagsgewichtung (als Überblick siehe Bonaccio & Dalal, 2006). Hierbei konnte zum einen nachgewiesen werden, dass Ratschläge die Genauigkeit der Finalurteile erhöhen können (z.B. Gardner & Berry, 1995; Sniezek et al., 2004; Yaniv, 2004a), was wiederum, wenig überraschend, mit der Qualität der Ratschläge zusammenhängt (Harvey & Fischer, 1997; Sniezek et al., 2004). Zum anderen kommt es zu egozentrischer Ratschlagsnutzung, also einer stärkeren Gewichtung des eigenen Initialurteils im Vergleich zum Ratschlag (z.B. Yaniv, 2004b; Yaniv & Kleinberger, 2000). Die Stärke der Ratschlagsgewichtung hängt dabei entscheidend von der wahrgenommenen Expertise der Ratgeber_innen (z.B. Harvey & Fischer, 1997; Sniezek et al., 2004) und der Qualität der Ratschläge ab (z.B. Gardner & Berry, 1995; Yaniv & Kleinberger, 2000). In anderen Worten können Ratschläge zwar zu besseren Finalurteilen führen, ihr Nutzen wird aber nicht vollständig ausgeschöpft, weil das eigene Initialurteil zu hoch gewichtet wird.

Mit dem Fokus auf die Ratschlagsgewichtung und die Akkuratheit von auf einen Ratschlag folgenden Finalurteilen ist allerdings eine entscheidende Frage bisher vollständig unbeachtet geblieben: Können Entscheidungsträger_innen, analog zu den Fertigkeitsverbesserungen im Gruppenkontext, auch von Ratschlägen lernen, bei anschließenden Aufgaben des gleichen Typs ihre Schätzgenauigkeit zu erhöhen? Zwar hat sich die vorangegangene Forschung bereits mit der Frage von sozialem Lernen in Ratschlagssituationen auseinandergesetzt (z.B. Biele, Rieskamp & Gonzalez, 2009; Çelen, Kariv & Schotter, 2010; Chaudhuri, Graziano & Maitra, 2006; Kocher, Sutter & Wakolbinger, 2014); die untersuchten Leistungsverbesserungen bezogen sich hierbei aber stets auf Finalurteile, die auf einen konkreten Ratschlag folgten. In anderen Worten wurde überprüft, ob ein Ratschlag die darauffolgende Leistung bei derselben Aufgabe verbessert. Die Befunde von Biele et al. (2009) zeigen zum Beispiel, dass ein einzelner Ratschlag, der vor dem ersten von mehreren Durchgängen einer Entscheidungsaufgabe gegeben wird, die anschließende Leistung erhöhen kann. Diese Leistungsveränderungen können also ausschließlich als *spezifischer*

Transfer innerhalb einer konkreten Aufgabe interpretiert werden. Mit diesen Befunden übereinstimmend kann also die Ratschlagsgewichtung in JAS Experimenten, die die Genauigkeit der Finalurteile erhöhen kann (z.B. Sniezek et al., 2004), ebenfalls als spezifischer Transfer interpretiert werden. Über diese Leistungsverbesserungen der Finalurteile hinaus könnten Ratschläge aber einen weiteren positiven Effekt haben. Ratschläge könnten auch die Akkuratheit folgender unbeeinflusster Initialurteile positiv beeinflussen, was im zweiten Teil des Forschungsprogramms überprüft werden sollte. In anderen Worten bestand die Forschungsfrage darin, ob ein einfacher Austausch numerischer Informationen in Ratschlagssituationen, analog zum ersten Teil des Forschungsprogramms, zu einem *generellen* Transfer von einer Aufgabe zu einer anderen Aufgabe der gleichen Art führen kann.

Die Befunde aus dem ersten Teil des Forschungsprogramms sowie vorangegangener Studien zu Schätzaufgaben (Schultze et al., 2012) implizieren, dass eine einmalige Gruppeninteraktion bereits ausreicht, um einen stabilen Lerngewinn zu erzielen. Aber was ist es, was diesen Lerngewinn auslöst? Basierend auf den Befunden des ersten Teils des Forschungsprogramms, dass der simple Austausch von numerischen Informationen bei nur einmaliger Gruppeninteraktion als Referenzwert ausreicht, um folgende Urteile zu verbessern, könnte eine solche Verbesserung der Schätzgenauigkeit auch durch einen einfachen Ratschlag erfolgen. Auch ohne zusätzliche Erläuterung sollten Entscheidungsträger_innen in der Lage sein, Unterschiede zwischen ihren eigenen Schätzungen und den Ratschlägen, insbesondere im metric error, zu erkennen und potenziell zu reduzieren. Wenn Entscheidungsträger_innen zum Beispiel die Entfernungen zwischen europäischen Hauptstädten stets größer als 20.000 Kilometer schätzen und in der Folge Ratschläge erhalten, die 3.000 Kilometer niemals überschreiten, sollte diese systematische Diskrepanz relativ leicht zu erkennen sein und folglich versucht werden, diese zu reduzieren. Auch hier sollte, analog zu den Lerneffekten im Gruppenkontext, der mapping error unberührt bleiben.

Es ist naheliegend, dass die beschriebenen Veränderungen der Schätzgenauigkeit eng mit der Qualität des Ratschlags zusammenhängen sollten. Zum einen ist es zielführender, sich einem guten Ratschlag anzupassen, als eine ungenaue Ratschlagstendenz zu übernehmen. Zum anderen sollten Handlungsweisen, mit denen Andere Erfolg haben, eher nachgeahmt werden als solche, die sich nicht bewährt haben (vgl. Bandura, 1986). Entscheidungsträger_innen sollten sich in ihren Initialurteilen also insbesondere an gute Ratschläge anpassen. Da Entscheidungsträger_innen in der JAS Forschung in der Regel die Gedankengänge ihrer Ratgeber_innen nicht nachvollziehen können (vgl. Bonaccio & Dalal, 2006), ist es mitunter schwierig, die Ratschlagsakkuratheit zu erkennen. Aus diesem Grund sollten mögliche

Lerneffekte insbesondere dann auftreten, wenn die Entscheidungsträger_innen Feedback über die Ratschlagsqualität erhalten. Nichtsdestotrotz sollte es auch ohne zusätzliche Rückmeldung über die Qualität des Ratschlags zu Verbesserungen der Schätzgenauigkeit der Initialurteile kommen, da Entscheidungsträger_innen, bis zu einem gewissen Grad, dennoch für deren Qualität sensitiv sein sollten (Biele et al., 2009; Yaniv & Kleinberger, 2000).

Zusammenfassend sollte im zweiten Teil des Forschungsprogramms überprüft werden, ob gute Ratschläge den metric error reduzieren und dadurch die Akkuratheit von nachfolgenden Initialurteilen verbessern, sowie welche Rolle die Salienz der Ratschlagsqualität dabei spielt. Darüber hinaus sollte explorativ analysiert werden, wie stark mögliche soziale Lerneffekte zu dem Befund von verbesserten Finalurteilen nach Erhalt von Ratschlägen (z.B. Soll & Larrick, 2009; Sniezek et al., 2004) beitragen. In anderen Worten sollte zwischen zwei Quellen der Ratschlagsadjustierung unterschieden werden: zum einen Akkuratheitsgewinne der Initialurteile als Folge von sozialem Lernen, und zum anderen Ratschlagsgewichtung im Sinne einer Anpassung des Finalurteils an einen konkreten Ratschlag.

Eine laborexperimentelle Überprüfung von sozialen Lernprozessen in Ratschlagssituationen

Methode

Die Fragen des zweiten Teils des Arbeitsprogramms wurden in drei Laborexperimenten überprüft. Im ersten Experiment sollten die Versuchspersonen (*N* = 197) verschiedene Luftlinie-Distanzen zwischen europäischen Hauptstädten so genau wie möglich schätzen und erhielten dabei Ratschläge von immer derselben zufälligen Ratgeberin oder demselben zufälligen Ratgeber, unterschiedlichen zufälligen Ratgeberinnen und Ratgebern oder keine Ratschläge. Das Experiment bestand dabei aus zwei Phasen. In der ersten Phase von 10 Aufgaben erhielten alle Versuchspersonen keinen Ratschlag. Diese Übungsphase diente als Leistungsbaseline, um Veränderungen in der Schätzgenauigkeit nach den ersten Ratschlägen bestimmen zu können. Die zweite Phase von 20 Aufgaben folgte in den beiden Ratschlagsbedingungen dem klassischen JAS Ablauf, mit Initialurteil, Ratschlag und Finalurteil. Die Ratschläge wurden dabei zufällig von 76 Vortest-Versuchspersonen gezogen. In der Kontrollbedingung wurden auch in der zweiten Phase keine Ratschläge gewährt. Wichtigste abhängige Variable war (in allen Experimenten) die Akkuratheit der unbeeinflussten Initialschätzungen, die analog zum ersten Teil des Forschungsprogramms durch den MAPE erfasst wurde.

Das zweite Experiment arbeitete neben einem anderen Schätzaufgabentyp (Gewicht verschiedener Gegenstände) auch mit einem veränderten Design. Anstatt die Ratgeber_innen zufällig zu ziehen, wurde im zweiten Experiment die Ratschlagsqualität manipuliert. Diese Ratschläge waren entweder die Schätzung der besten, mittleren oder schlechtesten Versuchsperson aus einem Vortest mit 61 Teilnehmern. Des Weiteren erhielten die Versuchspersonen (*N* = 132) am Anfang des Experiments Informationen darüber, welchen Rangplatz ihr_e Ratgeber_in im Vortest erzielt hatte. Auf diese Weise sollte die Salienz der Qualität der Ratschläge erhöht werden, um ein möglichst optimales Umfeld für Lerngewinne zu schaffen. Den drei Ratschlagsbedingungen wurde wiederum eine Kontrollbedingung ohne Ratschläge gegenübergestellt. Darüber hinaus fand im zweiten Experiment keine individuelle Übungsphase statt, weil im ersten Experiment keine Leistungsveränderungen ohne Ratschlag festgestellt werden konnten, sodass die Kontrollbedingung ausreicht, um Unterschiede der Initialurteile in Abhängigkeit der Ratschlagsqualität überprüfen zu können.

Experiment 3 war darauf ausgerichtet, die grundlegenden Befunde des zweiten Experiments mit einigen Veränderungen zu replizieren. Zunächst sollten die Versuchspersonen (*N* = 164), analog zu Experiment 1, Luftlinie-Distanzen zwischen europäischen Hauptstädten schätzen. Des Weiteren wurden im dritten Experiment, basierend auf den Ergebnissen des zweiten Experiments, nur noch die extremen Bedingungen gute_r (beste von 76 Vortest-Versuchspersonen) und schlechte_r (schlechteste von 76 Vortest-Versuchspersonen) Ratgeber_in miteinander verglichen. Weiterhin wurde manipuliert, ob die Versuchspersonen Feedback über den Rangplatz ihrer Ratgeberin oder ihres Ratgebers aus dem Vortest erhielten. Auf diese Weise sollte überprüft werden, ob die Versuchspersonen auch ohne solche Informationen in der Lage sind, von Ratschlägen zu lernen, und inwiefern mögliche Lerngewinne mit Feedback stärker ausfallen als ohne. Schließlich wurde, wie bei Experiment 1, eine individuelle Übungsphase von 10 Aufgaben ergänzt. Dies erlaubt es, Leistungsveränderungen als Folge von Ratschlägen zu überprüfen und nicht nur Leistungsunterschiede, wie beim zweiten Experiment. Darüber hinaus konnte durch diese Übungsphase die Kontrollbedingung ohne Ratschläge eingespart werden, weil während dieser Phase überprüft werden konnte, ob sich die Urteile der Versuchspersonen ohne den Erhalt von Ratschlägen verändern.

Ergebnisse

In Übereinstimmung mit den Hypothesen zeigte sich deutliche Evidenz, dass Ratschläge von mittlerer bis hoher Qualität die Akkuratheit von nachfolgenden Initialurteilen verbesserten. Schlechte Ratschläge hingegen beeinträchtigten die Initialurteile nicht signifikant. Des Weiteren wiesen die Ergebnisse darauf hin, dass gute Ratschläge in erster Linie den metric error der Entscheidungsträger_innen verringerten, während sich keine Veränderungen beim mapping error nachweisen ließen. Darüber hinaus zeigte sich überraschenderweise kein signifikanter Effekt der Salienz der Ratschlagsqualität auf das generelle Befundmuster. Auch ohne Feedback über die Ratschlagsqualität waren die Versuchspersonen in der Lage, diese grundlegend zu erkennen (die Akkuratheit der guten Ratgeberin oder des guten Ratgebers wurde höher eingeschätzt als die der schlechten Ratgeberin oder des schlechten Ratgebers), sodass die Entscheidungsträger_innen bei guten Ratschlägen ihre Initialurteile verbesserten und bei schlechten Ratschlägen ihre Initialurteile nicht veränderten. Abschließend konnte in zusätzlichen explorativen Analysen zwischen zwei unterschiedlichen positiven Effekten von qualitativ hochwertigen Ratschlägen unterschieden werden. Zum einen traten die bereits beschriebenen Lerneffekte auf, die die Akkuratheit der auf den Ratschlag folgenden Initialurteile verbesserten, zum anderen kam es zu zusätzlichen Verbesserungen durch die Integrierung der Ratschläge in die jeweiligen Finalurteile. Die Ergebnisse lieferten jedoch klare Evidenz, dass die in der vorangegangenen Forschung nicht berücksichtigte Leistungsverbesserungen der Initialurteile den größten Anteil der gesamten Leistungsverbesserungen ausmachen.

Fazit

Im zweiten Teil des Forschungsprogramms wurde überprüft, ob individuelle Lerngewinne auch in Ratschlagssituationen ohne direkte Interaktion auftreten können. Die Befunde der drei Experimente zeigen, dass Ratschläge tatsächlich zu Verbesserungen folgender Initialurteile führen können, diese aber stark von der Qualität der Ratschläge abhängen. In anderen Worten haben Ratschläge nicht nur einen positiven Effekt bei einer bestimmten Aufgabe, wie es bereits mehrfach gezeigt worden ist (z.B. Biele et al., 2009; Çelen et al., 2010; Chaudhuri et al., 2006; Kocher et al., 2014; Soll & Larrick, 2009; Sniezek et al., 2004), sondern auch auf andere Aufgaben der gleichen Art. Die vorliegenden Befunde stellen also den ersten Nachweis von generellem sozialen Lernen durch Ratschläge dar. Die Tatsache, dass es auch in solchen Ratschlagssituation, nach Erhalt eines guten Ratschlags, ausschließlich zu Verbesserungen der metric errors ohne systematische Veränderungen der mapping errors kommt, spricht wiederum für die Zentralität von Referenzwerten (z.B. Bonner & Baumann, 2008; Bonner et al., 2007; Laughlin et al., 1999; Laughlin et al., 2003). Nur wenn Versuchspersonen einen guten Referenzwert von der Ratgeberin oder dem Ratgeber erhalten, können sie davon bei zukünftigen unbeeinflussten Initialurteilen profitieren. Darüber hinaus scheinen die Entscheidungsträger_innen empfindsam für die Ratschlagsqualität zu sein, da die Ergebnisse darauf hindeuten, dass der Lernprozess völlig losgelöst vom Erhalt zusätzlicher Informationen über die Ratschlagsqualität ist. Eine mögliche Erklärung für diesen Befund ist, dass Entscheidungsträger_innen besonders schlechte Ratschläge als unplausibel identifizieren, auch wenn sie selbst nicht in der Lage sind, eine gute Schätzung abzugeben (Yaniv & Kleinberger, 2000). Es sollte zum Beispiel schwierig sein, zu entscheiden, ob zwischen London und Rom eher 1.500 oder 2.000 Kilometer liegen. Wenn der Ratschlag dann allerdings bei 30.000 Kilometer liegt, besteht eine hohe Wahrscheinlichkeit, dass Entscheidungsträger_innen erkennen, dass es sich vermutlich um eine schlechte Ratgeberin oder einen schlechten Ratgeber handelt. Dies sollte wiederum dazu führen, dass dieser schlechte Referenzwert eher nicht ins eigene Urteil eingebunden wird, selbst wenn man keine explizite Rückmeldung über die Qualität des Ratschlags erhält.

Zusammenfassend deuten die vorliegenden Ergebnisse darauf hin, dass Ratschläge auf zwei Arten die Schätzgenauigkeit verbessern können. Zum einen führen soziale Lernprozesse dazu, dass sich die Initialurteile von Entscheidungsträgerinnen und Entscheidungsträgern verbessern, insbesondere, wenn sie gute Ratschläge erhalten. Zum anderen adjustieren Entscheidungsträger_innen ihre Schätzungen in Richtung der Ratschläge, wenn sie zu einem Finalurteil kommen. Der Großteil an Leistungsverbesserungen, die durch Ratschläge ermöglicht werden, kommt allerdings durch den als erstes beschriebenen generellen Lerneffekt zustande. Aus diesem Grund sollte die zukünftige JAS Forschung diese bisher unerforschten sozialen Lerneffekte stärker berücksichtigen, um den vollen Nutzen von Ratschlägen besser abschätzen zu können.

Bewertung des Forschungsprogramms und Diskussion

Im Rahmen der vorliegenden Dissertation wurde das Auftreten von sozial vermittelten Lernprozessen im Gruppenkontext sowie in Ratschlagssituation bei quantitativen Schätzaufgaben untersucht. Dazu wurde in beiden Teilen des Forschungsprogramms der Fokus darauf gerichtet, auf welche Weise die Schätzgenauigkeit der Gruppenmitglieder oder Entscheidungsträger_innen steigt.

Integrative Zusammenfassung der beiden Teile des Forschungsprogramms

Die Herangehensweise erfolgte aus zwei – bezüglich des Ausmaßes der sozialen Interaktion – unterschiedlichen Perspektiven: Zunächst wurde mit tatsächlich interagierenden Gruppen Rahmenbedingungen geschaffen, die sozial vermittelten Lernprozessen zuträglich sein sollten. Gruppenmitglieder haben hierbei nicht nur die Möglichkeit, ihre individuellen Schätzungen auszutauschen, sondern diese auch zu erläutern und somit leichter ihre Qualität salient zu machen. Im Gegensatz hierzu wurden im zweiten Teil des Arbeitsprogramms Ratschlagssituationen mit Hilfe des JAS (Sniezek & Buckley, 1995; Sniezek & Van Swol, 2001) untersucht. Hierbei erhält die Entscheidungsträgerin oder der Entscheidungsträger üblicherweise einen Ratschlag in Form eines numerischen Werts ohne weitere Erklärungen, warum die Ratgeberin oder der Ratgeber gerade diesen Ratschlag abgibt. Dies sollte eine ungleich schwerere Situation für das Auftreten von Lernprozessen darstellen, da kein Wissen über den reinen numerischen Ratschlag hinaus vermittelt werden kann und es für Entscheidungsträger_innen eine größere Herausforderung sein sollte, gute von schlechten Ratschlägen zu unterscheiden, als für Mitglieder von interagierenden Gruppen. Über den Inhalt der separaten Forschungsprogramme hinaus, bietet das Arbeitsprogramm als Ganzes entsprechend die Möglichkeit, herauszufinden, ob die Lernprozesse in beiden Situationen ähnlich verlaufen oder sich grundlegend voneinander unterscheiden.

Soziale Lernprozesse bei quantitativen Schätzaufgaben

Wie bereits beschrieben, zeigen vorangegangene Forschungsbefunde, dass bei quantitativen Schätzaufgaben Gruppenarbeit die Leistung ihrer Mitglieder verbessern kann (Schultze et al., 2012) und Ratschläge die Genauigkeit der finalen, auf einen Ratschlag folgenden Schätzungen erhöhen können (Soll & Larrick, 2009; Sniezek et al., 2004). Das vorliegende Arbeitsprogramm leistet darüber hinaus einen Beitrag dazu, aufzuzeigen, wie einfach solche Leistungsverbesserungen von statten gehen können und welche Prozesse ihnen zugrunde liegen. Der erste Teil des Forschungsprogramms macht deutlich, dass bereits eine einfache Gruppeninteraktion ausreicht, um den vollen Nutzen der Gruppenzusammenarbeit auszuschöpfen, wobei die realisierten Leistungsverbesserungen zeitlich stabil sind. Der zweite Teil des Forschungsprogramms zeigt, dass Ratschläge nicht nur die Finalurteile von Entscheidungsträgerinnen und Entscheidungsträgern verbessern können, sondern darüber hinaus auch die Schätzgenauigkeit folgender Initialurteile positiv beeinflussen. Dies gelingt sogar dann, wenn die Qualität des Ratschlags nicht zusätzlich in Form von Feedback salient gemacht wird. Es scheint demnach für das Auftreten individueller Lerngewinne bei quantitativen Schätzaufgaben nicht nötig zu sein, mit anderen Personen zu kommunizieren.

Die einfache Darbietung eines guten Ratschlags kann dementsprechend ebenso individuelle Leistungsverbesserungen hervorbringen wie eine tatsächliche Gruppeninteraktion. Entscheidend scheint bei diesem Prozess, genauso wie bei Lernprozessen in interagierenden Gruppen, der Austausch von numerischen Referenzwerten zu sein (z.B. Bonner & Baumann, 2008; Bonner et al., 2007; Laughlin et al., 1999; Laughlin et al., 2003). Diese Annahme wird vom gemeinsamen Befund verminderter metric errors aus allen im Arbeitsprogramm enthaltenen Experimenten unterstützt. Sowohl Gruppenmitglieder als auch Entscheidungsträger_innen im JAS erfahren durch die Gruppeninteraktion oder den Ratschlag numerische Schätzungen anderer Personen. Wenn diese Schätzungen als hilfreich wahrgenommen werden, kann die eigene Meinung in Richtung der anderen Gruppenmitglieder oder des Ratschlags angepasst werden. In anderen Worten wird die eigene Metrik adjustiert und das unabhängig davon, ob man in einer Gruppe arbeitet und mit Anderen interagiert oder einen numerischen Referenzwert in Form eines Ratschlags erhält. Zusammenfassend zeigt sich also, dass es bei den verwendeten Schätzaufgaben sehr einfach zu sozialen Lernprozessen kommen kann und die Lernprozesse sowohl in Gruppen als auch in Ratschlagssituationen nach demselben Muster ablaufen.

Allerdings kann man dieses Ergebnis, auch wenn es bei zwei unterschiedlichen quantitativen Schätzaufgaben nachgewiesen werden konnte, sicherlich nicht ohne weiteres auf alle solchen Aufgabentypen generalisieren. Zum Beispiel sollte eine komplexere Aufgabe die Geschwindigkeit des Lernens beeinflussen oder könnte sogar sozial vermitteltem Lernen in einer Ratschlagssituation im Wege stehen. Der einfache Austausch von numerischen Referenzwerten könnte bei solchen Aufgaben nicht ausreichen, um die Schätzgenauigkeit zu erhöhen. Darüber hinaus könnten andere Schätzaufgaben auch die Reduzierung von mapping errors ermöglichen. Die im Arbeitsprogramm verwendeten Aufgaben zeichneten sich dadurch aus, dass Versuchspersonen dazu neigten, relativ große metric errors zu haben. Im Gegensatz dazu könnten die mapping errors der Versuchspersonen bei den untersuchten Aufgabentypen weniger ausgeprägt gewesen sein. Vermutlich hatten die meisten Versuchspersonen eine grobe Vorstellung von der geographischen Lage verschiedener europäischer Länder, was es ihnen ermöglichte, längere von kürzeren Distanzen zu unterscheiden. Darüber hinaus sollten die Versuchspersonen ebenso dazu in der Lage gewesen sein abzuschätzen, dass ein Kamm weniger wiegt als ein Hammer. In anderen Worten boten die verwendeten Aufgaben ein großes Potenzial für Verbesserungen im metric, aber nicht im mapping error. Dieses Spezifikum gilt aber mit Sicherheit nicht für alle Arten von Schätzaufgaben. Zum Beispiel sollten Finanzprognosen wie Aktienkurse in erster Linie durch mapping errors charakterisiert sein.

Der aktuelle Aktienkurs stellt einen guten Referenzwert dar, der den metric error von vornherein minimieren sollte. Im Gegensatz dazu sollte Wissen über generelle Marktentwicklungen sowie den Erfolg und zukünftige Pläne eines Unternehmens die Prognosen eines Aktienkurses zu verschiedenen Zeitpunkten verbessern. In anderen Worten sollte solches Wissen also in erster Linie den mapping error reduzieren und dadurch die Akkuratheit der Prognosen verbessern. Es bedarf also weiterer Forschung mit quantitativen Schätzaufgaben unterschiedlicher Charakteristika, um ein vollständigeres Bild davon zu erhalten, wie sozial vermittelte Lernprozesse bei solchen Aufgaben ablaufen können.

Schlussbemerkung

Sozial vermittelte Lernprozesse spielen eine wichtige Rolle bei quantitativen Schätzaufgaben. Das im Rahmen dieser Promotion durchgeführte Forschungsprogramm erweitert die empirische Forschung zu diesem Thema um die Frage, was bei typischen, in der Forschung verwendeten Aufgaben gelernt wird und ob ein genereller Transfer – von einer Aufgabe zu einer anderen Aufgabe der gleichen Art – auch bei JAS Experimenten ohne direkte Interaktion zwischen Entscheidungsträger_in und Ratgeber_in stattfindet. Die Ergebnisse zeigen zum einen die Wichtigkeit von numerischen Referenzwerten, was durch die starke Reduktion der metric errors, sowohl durch Gruppeninteraktion als auch durch Ratschläge, verdeutlicht wurde. Zum anderen gelang der erste Nachweis von generellem Lernen in Ratschlagssituation. Die Ergebnisse deuten darüber hinaus sogar an, dass dieser Lernprozess stärker ist als der bereits aus der Forschung bekannte positive Effekt der reinen Ratschlagsgewichtung auf die Qualität von Finalurteilen.

Literaturverzeichnis

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Biele, G., Rieskamp, J. & Gonzalez, R. (2009). Computational models for the combination of advice and individual learning. *Cognitive Science*, *33*, 206-242.

Bonaccio, S. & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*, 127-151.

Bonner, B. L. & Baumann, M. R. (2008). Informational intra-group influence: the effects of time pressure and group size. *European Journal of Social Psychology, 38,* 46-66.

Bonner, B. L. & Baumann, M. R. (2012). Leveraging member expertise to improve knowledge transfer and demonstrability in groups. *Journal of Personality and Social Psychology*, *102*, 337-350.

Bonner, B. L., Sillito, S. D. & Baumann, M. R. (2007). Collective estimation: Accuracy, expertise, and extroversion as sources of intra-group influence. *Organizational Behavior and Human Decision Processes*, *103*, 121-133.

Brodbeck, F. & Greitemeyer, T. (2000). A dynamic model of group performance: Considering the group members' capacity to learn. *Group Processes & Intergroup Relations*, *3*, 159-182.

Brown, N. R. & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100,* 511-534.

Çelen, B., Kariv, S. & Schotter, A. (2010). An experimental test of advice and social learning. *Management Science, 56,* 1687-1701.

Chaudhuri, A., Graziano, S. & Maitra, P. (2006). Social learning and norms in a public goods experiment with inter-generational advice. *The Review of Economic Studies, 73,* 357-380.

Davis-Stober, C. P., Budescu, D. V., Dana, J. & Broomell, S. B. (2014). When is a crowd wise? *Decision, 1,* 79-101.

Einhorn, H. J., Hogarth, R. M. & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin, 84,* 158-172.

Gardner, P. H. & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology, 9,* 55-79.

Harvey, N. & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70,* 117-133.

Kocher, M., Sutter, M. & Wakolbinger, F. (2014). Social Learning in Beauty-Contest Games. *Southern Economic Journal*, *80*, 586-613.

Laughlin, P. R. & Barth, J. M. (1981). Group-to-individual and individual-to-group problem-solving transfer. *Journal of Personality and Social Psychology, 41*, 1087-1093.

Laughlin, P. R., Bonner, B. L., Miner, A. G. & Carnevale, P. J. (1999). Frames of reference in quantity estimations by groups and individuals. *Organizational Behavior and Human Decision Processes*, *80*, 103-117.

Laughlin, P. R., Carey, H. R. & Kerr, N. L. (2008). Group-to-individual problem-solving transfer. *Group Processes & Intergroup Relations, 11*, 319-330.

Laughlin, P. R. & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*, 177-189.

Laughlin, P. R., Gonzalez, C. M. & Sommer, D. (2003). Quantity estimations by groups and individuals: Effects of known domain boundaries. *Group Dynamics: Theory, Research, and Practice*, *7*, 55-63.

Laughlin, P. R. & Sweeney, J. D. (1977). Individual-to-group and group-to-individual transfer in problem solving. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 246-254.

Schultz-Hardt, S., Greitemeyer, T., Brodbeck, F. E. & Frey, D. (2002). Sozialpsychologische Theorien zu Urteilen, Entscheidungen, Leistung und Lernen in Gruppen. *Theorien der Sozialpsychologie*, *2*, 13-46.

Schultze, T., Mojzisch, A. & Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative to differential weighting. *Organizational Behavior and Human Decision Processes*, *118*, 24-36.

Sniezek, J. A. & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*, 159-174.

Sniezek, J. A. & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, *43*, 1-28.

Sniezek, J. A. & Henry, R. A. (1990). Revision, weighting, and commitment in consensus group judgment. *Organizational Behavior and Human Decision Processes*, *45*, 66-84.

Sniezek, J. A., Schrah, G. E. & Dalal, R. S. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, *17*, 173-190.

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. Organizational behavior and human decision processes, 84, 288-307.

Soll, J. B. & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 780-805.

Stasson, M. F., Kameda, T., Parks, C. D., Zimmerman, S. K. & Davis, J. H. (1991). Effects of assigned group consensus requirement on group problem solving and group members' learning. *Social Psychology Quarterly, 54*, 25-35.

Steiner, I. D. (1972). Group processes and group productivity. *New York: Academic*.

Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology*, *44*, 443-461.

Yaniv, I. (2004a). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*, 75-78.

Yaniv, I. (2004b). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*, 1-13.

Yaniv, I. & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*, 260-281.

Anhang 1:

How much group is necessary?

Group-to-individual transfer in estimation tasks

How much group is necessary?

Group-to-individual transfer in estimation tasks

Alexander Stern, Thomas Schultze, & Stefan Schulz-Hardt,

University of Goettingen

Authors' Note

Alexander Stern, Thomas Schultze, and Stefan Schulz-Hardt, University of Goettingen, Institute of Psychology, Germany

Correspondence concerning this manuscript should be addressed to the first author: Alexander Stern, Institute of Psychology, Georg-August-University Göttingen, Waldweg 26, 37073 Göttingen, Germany.

E-mail: stern@psych.uni-goettingen.de

Telephone: +49 551 39 21114

Fax: +49 551 39 1821111

Abstract

G-I transfer denotes an increase in individual performance due to group interaction, for example, because of acquiring certain skills or knowledge from the other group members. Whereas such G-I transfer has been successfully shown for problem-solving tasks, evidence for G-I transfer on quantitative estimation tasks is scarce. We address this research gap with a focus on how often a group has to interact in order to fully exploit the benefit of this learning effect. Results from two experiments support the idea that a single group interaction is sufficient to induce a stable G-I transfer, which reduces group members' metric error. In contrast to nominal groups, both members of continuously interacting groups and members of groups with only one initial interaction exhibited stable G-I transfer, and the size of this transfer did not significantly differ between the latter two conditions. Furthermore, we found evidence for differential weighting of group members' individual contributions that goes beyond sheer individual capability gains under certain circumstances, namely in tasks with a population bias.

Keywords: group judgment; group performance; group-to-individual transfer; quantitative estimates; group learning; differential weighting

How much group is necessary?

Group-to-individual transfer in estimation tasks

Estimation tasks like forecasting prospective profits for a company, or estimating the expected increase in global temperature, lie at the heart of many far-reaching financial or political decisions. Important quantitative judgments are often made in groups, because groups are considered to be superior to a comparable number of individuals with regard to performance. For example, there is evidence that groups outperform comparison individuals (e.g., Bonner & Baumann, 2012; Bonner, Sillito, & Baumann, 2007; Laughlin, Bonner, Miner, & Carnevale, 1999) or even perform on the level of very challenging baselines like the most accurate group member's estimates (e.g., Einhorn, Hogarth, & Klempner, 1977; Laughlin, Gonzalez, & Sommer, 2003).

Beyond that, and independent of the actual group performance, group interaction might have another beneficial effect: individuals who interact in groups are assumed to exhibit particular social learning effects and thereby solve subsequent similar tasks more accurately than individuals who have no prior group experience. This possible individual capability gain as a consequence of prior collective task performance in a group is called *group-to-individual transfer* (G-I transfer, e.g., Laughlin & Barth, 1981; Laughlin & Sweeney, 1977). Despite the fact that there is robust evidence for such group learning processes in problem-solving tasks (e.g., Laughlin, Carey, & Kerr, 2008; Laughlin & Ellis, 1986; Stasson, Kameda, Parks, Zimmerman, & Davis, 1991), this phenomenon has been mostly neglected in research on quantitative group estimations. To our knowledge, the only exception is a study by Schultze, Mojzisch, and Schulz-Hardt (2012), which found strong individual performance improvements in quantitative estimations after group interaction. However, it is yet unclear what people learn and whether they need ongoing social interaction to maintain this improved performance. Therefore, in the present research we try to find out which knowledge is transferred when interacting with others. Furthermore, we focus on the repetitions of group interaction in order to accomplish stable individual performance enhancements. In other words, we investigate if one group interaction is sufficient to produce a significant increase in individual accuracy and, most importantly, whether it persists after group members leave the group.

Group-to-individual transfer

Building on the dynamic model of group performance by Brodbeck and Greitemeyer (2000a), we understand group learning as a function of two sources of change in individual resources that can improve groups in a complementary way. On the one hand, group members can improve their capabilities to work efficiently with each other (*learning to collaborate*). For example, group members might develop a shared mental model of the task, or they could acquire knowledge about the expertise of the other group members. On the other hand, and this is what we focus on in this paper, group members can improve their individual task-related skills as a consequence of group interaction, independent of purely individual practice effects (*learning to perform the task*). As already mentioned, this socially induced individual learning is known as *G-I transfer*. In other words, collectively performing tasks together can enhance group members' individual resources to perform the task on their own. Examples of such learning processes are vicarious learning, or exchange of basic principles and strategies for effective task performance (e.g., Laughlin & Jaccard, 1975; Brodbeck & Greitemeyer, 2000a, 2000b).

So far, most research on G-I transfer has been conducted in the domain of problem-solving tasks, with ample empirical evidence for its existence. For example, participants who had worked on mathematical problems in a group later solved the same or other, logically related problems better than individuals working alone (e.g., Laughlin & Ellis, 1986; Stasson et al., 1991). Similarly, participants with prior group interaction exhibited better individual performance in rule induction tasks compared to participants without such group interaction (e.g., Brodbeck and Greitemeyer, 2000b). By using multiple training sessions, Laughlin et al. (2008) addressed the necessary repetitions of group interaction in order to achieve G-I transfer, the major result of which was that one group interaction was sufficient for the occurrence of a stable G-I transfer. In other words, multiple group interactions did not affect the strength of the individual performance enhancements. However, all of these studies worked with tasks that are very likely characterized by high levels of demonstrability, which is considered to be a prerequisite for the occurrence of G-I transfer (Brodbeck & Greitemeyer, 2000a). According to Laughlin and Ellis (1986), one of the core conditions of demonstrability is that the member with the correct answer must have the ability, motivation, and time to demonstrate the correct solution to the other group members. On mathematical problems, this should usually be the case. When the task complexity is moderate, the member with the right solution should be able to explain its correctness. In contrast, on quantitative estimation tasks, it might be difficult for the best group member to demonstrate the high quality of his or her estimate, and for inferior group members to understand its quality. This lack of demonstrability might have some consequences on the occurrence of G-I-transfer. For example, it is more difficult to justify an estimation of New York City having around 8.4 million inhabitants than explaining that 5 times 7 equals 35. Nevertheless, as long as people do not simply guess their estimations on world knowledge tasks, it is generally possible to explain why certain estimations are better than others. This should be especially true when it comes to rather poor estimations. For example, one might explain quite easily why the population of New York City cannot be 200 million when taking into account that the whole United States of America has around 320 million citizens. On the other hand, it should be more difficult to judge whether the city has either 6 or 7 million inhabitants. Therefore, G-I transfer might have somewhat higher hurdles in estimation tasks as compared to problem-solving tasks.

Beyond that, differences in task demonstrability should have another important consequence. When tasks are characterized by a high demonstrability, as we often find it on problem-solving tasks like arithmetic problems (e.g. Laughlin & Ellis, 1986; Stasson et al., 1991), the most capable member often determines the group outcome. In other words, one capable group member can be sufficient for solving the task. On this basis, group-level performance often does not benefit from individual capability gains because, even if the less capable group members become more capable over time, it is unlikely that their contributions will add anything beyond that of the most capable member. In addition, it is also unlikely that this most capable member will improve his or her performance in the absence of superior models to learn from. In contrast, on tasks with a somewhat lower demonstrability, like estimation tasks, the accuracy of group estimations usually benefits from taking *all* group members’ opinions into account (e.g., in the form of weighted or unweighted averages), because this can help to eliminate or, at least, reduce idiosyncratic errors. In other words, exclusively relying on the most capable member is usually not the best strategy in quantitative estimation tasks (Bednarik & Schultze, 2015; Soll & Larrick, 2009). Consequently, the group as a whole might benefit from improved individual performances of inferior group members. This fact makes the field of quantitative estimation tasks particularly interesting, because here individual capability gains could actually lead to a performance enhancement on the level of the entire group.

To the best of our knowledge, there is only one study combining estimation tasks with a design that allows to detect increases in individual accuracy as a consequence of prior group interaction. Typically, studies in this field use a so-called *I-G design* (individual-group design, e.g., Bonner et al., 2007; Henry, 1993, 1995; Henry, Strickland, Yorges, & Ladd, 1996; Sniezek & Henry, 1989, 1990), meaning that participants first complete a series of quantitative estimation tasks individually (I) and then work on the same series of tasks as groups (G). Unfortunately, this design cannot account for G-I transfer, because individual performance is not measured *after* one or more group interactions have taken place. To address this limitation, Schultze et al. (2012) used an improved *aI-G design* (alternating-individual-group design). Their experiments were separated into two sections: (a) an individual practice phase and (b) a group phase consisting of alternating individual and group estimates of distances between different European capital cities. Consequently, changes in participants' individual accuracy due to the group interaction could be measured on their subsequent individual estimates. With this modified design, the authors found evidence for strong increases in individual accuracy after the first within-group interaction. In other words, group members already improved their individual performance after the first group discussion. In line with the idea of G-I transfer, inferior group members improved in accuracy while the groups' best members' accuracy remained stable. Furthermore, the improved estimation accuracy was relatively constant after this first major performance enhancement. Hence, it seemed that participants did not substantially benefit from further group interaction. These results are in line with the above-mentioned evidence from group problem-solving research, showing that one group interaction can be sufficient for the occurrence of a stable G-I transfer (Laughlin et al., 2008) in this type of task.

However, the findings of Schultze et al. (2012) leave two important question unanswered. The first one is *what* do people learn when interacting with others? To answer this question, it is useful to differentiate two types of estimation error. As outlined by Brown and Siegler (1993; see also Brown, 2002), one's knowledge in a judgment domain can be decomposed into two components: *metric knowledge* and *mapping knowledge*. Metric knowledge is a general understanding of the appropriate scaling, that is, whether people have an accurate representation of the correct upper and lower boundaries, or what range of values is plausible. For example, knowing that Germany has a length of approx. 900 kilometers, and that the equator has a length of roughly 40,000 kilometers, helps when estimating distances, and prevents us from making judgments that are completely implausible. In contrast to that, mapping knowledge is an accurate representation of the relative magnitude of possible target values. In other words, mapping knowledge allows us to put different target values of the same kind in the correct order. Most people know that the distance between London and Paris is shorter than the distance between London and New York, without necessarily having a good guess about the actual distances.

Interacting with others when working on estimation tasks should affect these two sources of estimation error differently, and to a different extent. Previous studies imply that providing people with frames of reference can strongly increase their estimation accuracy (e.g., Bonner & Baumann, 2008; Bonner et al., 2007; Laughlin et al., 1999; Laughlin et al., 2003). Collaborating with others during quantitative estimations could have exactly this effect: During their task-related communication, group members provide the reasoning for their individual estimates and illustrate the validity of certain benchmarks (Schultze et al., 2012). With regard to the two above-mentioned sources of error, such reference values should mainly improve metric knowledge quite rapidly and thereby diminish particularly implausible estimates. In contrast, reducing one’s mapping error during social interaction should be more difficult than understanding differences in group members’ metric error. When estimating distances between European cities one can only recognize that another group member has a different mapping error, for example, when realizing that he or she always overestimates distances between southern European cities and always underestimates distances between northern European cities. In other words, one need to precisely remember multiple estimates of other group members in order to recognize such differences. Consequently, the process of reducing one’s mapping error should be very slow and only possible after a long period of cooperation.

The second open question has to do with the stability of the learning process. Specifically, we do not yet know whether the G-I transfer is stable even if the group is completely disbanded, or whether continuous social interaction is needed for its maintenance. In other words, it is crucial to find out how the individual performance develops after the last group interaction. In the experiments of Schultze et al. (2012), participants continuously alternated between working on the estimation tasks individually and in groups, that is, they remained in a group context until the end of the experiment. Hence, so far there is no research about the temporal stability of G-I transfer in quantitative estimation tasks after members have left the group, and whether, under these conditions, one single group interaction leads to an equally strong individual performance enhancement compared to continuous group interaction.

Answering this question is not only relevant to gain a more conclusive theoretical understanding of the mechanisms that underlie G-I transfer. Rather, it is also a crucial question for practical purposes: The results of the Schultze et al. (2012) study suggest that it might be sufficient to have just one group interaction to fully exploit the benefits of having groups work on quantitative estimation tasks. As bringing group members together and having them discuss and decide on an issue costs more effort than just collecting and averaging individual judgments, truncating group interaction right after the first group judgment would, obviously, save a lot of resources. However, this would only pay off if the benefits of this interaction (i.e., the G-I transfer) do not fade away relatively soon after this first interaction. As Schultze et al. (2012) do not provide an empirical test for a sustainable beneficial G-I transfer after one single group interaction, we want to address this research gap in the current study.

Therefore, it is crucial to (a) replicate the finding of a strong increase in individual accuracy after just one group interaction, (b) analyze if group members increase their metric knowledge after interacting with others and (c) check whether their individual performance enhancement remains stable even if the first group interaction is also the last, that is, if all subsequent individual trials take place without any further group interactions in between. In other words, our study investigates what people actually learn during a group interaction and whether a single group interaction is sufficient to achieve a stable improvement in individual performance, or whether a robust transfer requires ongoing group interaction. We present two experiments exploring these issues. In each, we report all measures and manipulations. Furthermore, no participants were excluded from analyses.

Hypotheses

Schultze et al. (2012) found evidence that one group interaction might be sufficient for a strong increase in individual accuracy in quantitative estimation tasks. Our first aim is to test whether this effect is replicable by comparing two experimental group conditions (differing in the number of group interactions) to a control condition with nominal groups, that is, an equivalent number of non-interacting individuals. We predict that one group interaction is sufficient to achieve a significant increase in individual accuracy (G-I transfer):

Hypothesis 1: *Group members' individual accuracy will increase after the first group interaction. Both members of continuously interacting groups as well as members of groups with only a single group interaction will manifest these performance enhancements, whereas individual accuracy in the nominal groups will not improve at all.*

More importantly, we aim to answer the question of whether the stability of the G-I transfer requires ongoing group interaction. One single group interaction might be sufficient to achieve a stable performance enhancement. However, it is still possible that permanent group interaction is crucial for the stability of the increased individual estimation accuracy. In other words, the individual performance could deteriorate when the group is disbanded. As a consequence, we formulate two competing hypotheses regarding this research question:

Hypothesis 2a: *The increase in individual accuracy after the first group interaction is stable even when the group is disbanded.*

Hypothesis 2b: *The increase in individual accuracy after the first group interaction deteriorates after the group is disbanded.*

Furthermore, we are interested in whether there are differences between the two experimental conditions. Once again, there are two different possibilities, both of which we consider to be plausible. On the one hand, if the individual estimation accuracy is similarly stable after one as compared to many group interactions, the individual estimation accuracy should be (more or less) equally strong in both conditions. On the other hand, the increase in individual accuracy might deteriorate after the group is disbanded, or continuous group interaction could additionally foster the G-I transfer. This, in turn, should lead to stronger performance enhancements for members of continuously interacting groups. Accordingly, we also formulate two competing hypotheses for this issue:

Hypothesis 3a: *The increase in individual accuracy over the course of the individual trials is equally strong for members of single-interaction groups and continuous-interaction groups.*

Hypothesis 3b: *Members of continuous-interaction groups will manifest a stronger increase in individual accuracy than members of single-interaction groups.*

Finally, we aim to test *what* group members learn when interacting with others. We do not expect transfer of mapping knowledge since the process of reducing one's mapping error should be very slow and only possible after a long period of cooperation. In contrast, if the exchange of reference values underlies individual performance enhancements, group members should mainly reduce their metric error. Hence, we hypothesize:

Hypothesis 4: *Interacting with others will reduce group members' metric error. Both members of continuously interacting groups as well as members of groups with only a single group interaction will manifest this transfer of metric knowledge, whereas non-interacting individuals will not improve their metric error.*

Although the focus of our study is on individual performance after group interaction, for exploratory purposes we will also investigate group performance in comparison to individual performance. As hypothesized, group members might benefit individually from group interaction, which, in turn, could make groups better than an equivalent number of individuals. Hence, we will conduct an exploratory test of whether such surplus at the group level occurs in our study. Furthermore, we will also look at the possible occurrence of differential weighting strategies, that is, groups weighting more competent members more strongly. In addition to G-I transfer, such weighting strategies might also contribute to the quality of group judgments complementarily. So far, the only study that controlled for individual performance enhancements did not find any evidence for differential weighting (Schultze et al., 2012). Nevertheless, this latter finding need not necessarily be generalizable, because Schultze et al. only used one specific type of task. Consequently, we want to analyze if groups engage in differential weighting strategies on different quantitative estimation tasks. However, since individual capability gains are the focus of our study, we refrain from formulating hypotheses and, instead, analyze these two questions in an exploratory manner.

Experiment 1

In Experiment 1, we aimed to investigate whether the previously found individual performance enhancements due to group interaction (Schultze et al., 2012) are replicable, and whether or not this increase in accuracy requires ongoing group interaction. In other words, we wanted to find out whether a single group interaction has the same beneficial effect as multiple interactions. For this purpose, we compared continuously interacting groups with groups that were disbanded after their first within-group interaction, and with nominal groups. Members of both continuous-interaction and single-interaction groups should provide more accurate judgments individually due to the G-I-transfer. Furthermore, the experimental design allows us to examine whether G-I-transfer is equally strong after multiple group interactions in comparison to just one, which could then be interpreted as evidence for its stability beyond the group context.

Method

Participants, design and task

One hundred eighty-three German or German-speaking students (112 women, 70 men, one participant did not report his or her gender) with an average age of 21.43 (*SD* = 3.28) years participated in the experiment, with three persons each forming a real or nominal group. The sample size was based on a previous relevant study (Schultze et al., 2012). Experiment 1 used a mixed design with group type (continuous-interaction, single-interaction, no interaction) as a between subjects variable and task trial (or, for some analyses, trial block) as a within subjects variable.

The participants worked on a set of distance estimations between different European capital cities. This is the same task that has been used by Schultze et al. (2012). It was chosen based on two pretests (*N* = 40 and *N* = 38) revealing that there were stable differences in participants' individual performance (mean Spearman's Rho = .28, *p* < .001 and mean Spearman's Rho = .35, *p* < .001, respectively), which is a prerequisite for learning processes (Schultze et al., 2012). We measured the estimation accuracy with the mean absolute percent error (MAPE). In group judgment research, the MAPE score is a common measure of accuracy (e.g., Sniezek & Henry, 1989, 1990) and indicates the average relative deviation of the participant's estimates from the true values. The average MAPE scores in the two pretests of Experiment 1 indicated that pretest participants' deviated roughly 60 and 50 percent from the true value (*M* = 62.22, *SD* = 63.55 and *M* = 48.31, *SD* = 21.14). Furthermore, we checked whether participants' estimates were evenly distributed around the true values or whether the task contains a systematic *population bias*, that is, they tended to over or underestimate the true value. For this purpose, we calculated participants' mean percent deviation from the true values (thereby allowing overestimations and underestimations to cancel out each other). Corresponding *t*-tests against zero revealed no significant differences (*M* = 16.31, *SD* = 79.72), *t*(39) = 1.29, *p* = .203, *d* = 0.29, and (*M* = -2.59, *SD* = 38.94), *t*(37) = -.41, *p* = .685, *d* = 0.09, respectively, indicating that this task contained no substantial population bias.

Procedure

In each experimental session, six to nine participants were invited and randomly guided to one of three lab rooms, where they were placed at separate tables. Participants were informed about the task and the procedure of the experiment. They were instructed that the experiment consisted of two phases with ten distance estimates each: an individual practice phase and a group phase. Hence, participants knew from the beginning that they were going to interact unless the number of participants showing up was not divisible by three. In this case, excess participants were assigned to the individual control condition. The specific distances that the participants should estimate were not identical between the two phases but were, as the pretests indicated, on average, of similar difficulty.[1] In the practice phase, participants were asked to work on ten trials individually, and they were told that the goal was to estimate the airline distances between cities in kilometers as accurately as possible. Furthermore, the experimenter asked them not to communicate or to exchange notes. There was no time limit, but participants usually took between ten and fifteen minutes to finish this

[1] Furthermore, participants were asked to rate their estimation confidence on a six-point Likert scale after every judgment. However, because this confidence measure revealed no relevant effects for our research question we refrain from reporting any analyses regarding this measure. Nevertheless, we will be glad to publish our data for further analyses.

phase. Once they were done, the experimenter collected the data and computed the MAPE for each participant. Afterwards, the participants were assigned to three-person-groups. Whenever possible, we aimed for some heterogeneity in group members' skill level, as a certain amount of heterogeneity is necessary for individual capability gains and differential weighting strategies. To this end, groups were composed so that MAPE scores of the most capable and medium group member differed by at least ten percentage points and of the medium and least capable group member by another ten percentage points. Hence, participants' practice phase MAPE influenced the assignment to the three conditions.[2]

The second phase of the experiment differed depending on the experimental condition. In the continuous-interaction and the single-interaction condition, three participants each formed a group and were asked to take a seat at a shared table. The groups received four questionnaires containing the estimation tasks, one for each group member to write down his or her individual estimates, and one for the group estimates. The difference between the two group conditions was the number of group interactions. In continuous-interaction groups, the group members first worked on a specific distance estimate individually and then discussed their individual estimates in order to come up with a consensus estimate. Afterwards, they proceeded with the next trial of the group phase in the same fashion. Participants were told that they were neither allowed to inspect their estimates of trials they had already worked on, nor to revise these previous estimates. Furthermore, they were reminded not to communicate or exchange notes when working on their individual estimates. The single-interaction groups only interacted on the first trial of the second phase. Again, before interacting as a group, each group member had to come to an individual estimation. After discussion of the first task

[2] The assignment to the three conditions did not affect accuracy differences between the conditions, $F(2, 58) = 0.05$, $p = .953$, $\eta_p^2 < .01$, that is, the average MAPE-scores of the individual training phase (prior to the experimental manipulation) were almost identical for the three conditions. Beyond that, there was no evidence for average group member accuracy variance differences between conditions, $F(2, 58) = 0.73$, $p = .487$, $\eta_p^2 = .02$.

and making a group judgment, the single-interaction groups were disbanded, and their members were placed at separate tables where they worked on the remaining nine trials independently and without further discussion. The individual judgments of the nine trials on which participants in the single-interaction condition worked on their own were later averaged to form hypothetical group judgments. In the nominal group condition, participants worked on all ten estimates of the second phase individually, that is, they worked at separate tables and were not allowed to communicate or exchange information. Subsequently, the judgments of the three nominal group members were averaged to create the nominal group judgments. Participants had no guidelines regarding how to work on a particular task and, again, there was no time limit.

In each condition, the experimenter explained that the accuracy of the estimates during the second phase would determine the amount of money the participants would receive for participating in the experiment. In addition to a show-up fee of 5 Euro, there was an accuracy-based bonus payment ranging from 0 to 5 Euros.[3] After completing phase 2, participants were asked to fill in a final questionnaire containing a suspicion check. In the meantime, the experimenter calculated the MAPE score of the second phase to determine the bonus payment. Before the participants were dismissed, they were thanked for their participation and debriefed.

[3] For the continuous-interaction groups, this bonus was based on the accuracy of the group judgments. For single-interaction groups, it was based on the group judgment of the first trial in the second phase and the average of their individual estimates for the remaining nine trials. For nominal group members, the bonus depended solely on their individual accuracy during the second phase. However, participants were only aware of a general performance bonus and were not informed about the details of how this bonus was computed in each of the experimental conditions (to prevent that individual task motivation might differ between the experimental conditions as a consequence of differences in bonus computation).

Results and discussion

Group-to-individual transfer

In order to test for individual performance enhancement in terms of G-I transfer, we analyzed whether group interaction led to improved subsequent individual estimations. For reasons of simplification, we compared the differences in individual MAPE scores between the individual practice phase and the group phase in the three experimental conditions and did not add the two phases as a within-subjects factor. The trial right before the first group interaction was treated as the last trial of the individual practice phase, since this trial still took place before any effects of group interaction could have occurred.[4] Positive values of the accuracy difference measure represent an increase in accuracy from phase one to phase two. We conducted a 3 (*group type*: continuous-interaction vs. single-interaction vs. no interaction) × 3 (*group member*: most capable vs. medium vs. least capable) ANOVA with experimental condition as between-subjects factor and group member as within-subjects factor. The analysis revealed a significant main effect of group type, $F(2, 58) = 4.84$, $p = .011$, $\eta_p^2 = .14$. LSD post hoc comparisons showed that the performance enhancements where roughly similar in the continuous-interaction and the single-interaction condition ($M = 19.28$, $SD = 20.68$ vs. $M = 16.13$, $SD = 17.09$), $p = .611$.[5] However, participants increased their performance significantly more in both the continuous-interaction groups and the single-interaction groups than in the non-interacting nominal groups ($M = 19.28$, $SD = 20.68$ vs. $M = 1.77$, $SD = 20.12$), $p = .005$, and ($M = 16.13$, $SD = 17.09$ vs. $M = 1.77$, $SD = 20.12$), $p = .023$, respectively. Furthermore, separate post hoc *t*-tests

[4] We calculated a paired-samples *t*-tests between participants' average individual accuracy on the ten trials of the practice phase and the first trial of the second phase for the two group conditions, to rule out any motivational effects because participants' anticipation of group interaction might increase the accountability they feel for their estimate. However, this analysis revealed no differences at all ($M = 48.44$, $SD = 17.02$ vs. $M = 47.18$, $SD = 33.32$), $t(39) = 0.30$, $p = .762$, $d = 0.05$.

[5] We used LSD post hoc comparisons because we aimed for the highest possible statistical power when comparing the two interacting group conditions. More conservative post hoc comparisons might have disguised significant differences. As a consequence of the higher statistical power, the absence of significant differences between the two interacting group conditions is an even more informative finding.

against zero revealed a significant performance enhancement in the continuous-interaction condition, $t(20) = 4.27$, $p < .001$, $d = 0.94$, as well as in the single-interaction condition, $t(18) = 4.11$, $p = .001$, $d = 0.95$. The nominal group condition, in contrast, showed no significant changes in the individual MAPE scores from the first to the second phase, $t(20) = 0.40$, $p = .692$, $d = 0.09$. Because participants in the nominal group condition did not improve their performance between the two phases, we can assume that there are no substantial individual practice effects in the task we used, which mirrors the findings of Schultze et al. (2012). Accordingly, in line with Hypothesis 1, the increase in judgment accuracy in the two group conditions should be the result of G-I transfer, supporting the idea that one group interaction is sufficient to increase individual estimation accuracy. The results further indicate that, in general, the performance enhancements were equally strong after one as after multiple interactions. In other words, groups working on the distance estimates were able to exchange all information necessary to induce the full amount of increase in individual accuracy already during their first group discussion, supporting Hypothesis 3a.[6]

The ANOVA further revealed a main effect of group member, $F(2, 57) = 26.14$, $p < .001$, $\eta_p^2 = .31$, which was qualified by an interaction of group member and group type, $F(4, 116) = 4.51$, $p = .002$, $\eta_p^2 = .13$, in line with the idea that differences in G-I transfer can only be observed in the two group conditions and not in the non-interacting control condition. Separate post hoc paired-samples *t*-tests for the continuous-interaction and the single-interaction condition indicated that the accuracy improvements differed between all levels of group members' judgment accuracy, all *t*s(20) > 3.58, all *p*s < .002, all *d*s > 0.78, for the

[6] Beyond that, participants' gender did not significantly affect the dependent variable, $t(180) = 1.90$, $p = .059$, $d = 0.29$, although, descriptively, performance changes were somewhat more pronounced for women than for men. However, there was no association between the condition and participant's gender, χ^2 (2) = 0.33, $p = .859$. In other words, female and male participants were roughly equally distributed across conditions. Moreover, there was also no significant relationship between participants' age and the dependent variable, $r < .01$, $p = .990$.

continuous-interaction condition and, all *t*s(18) > 3.12, all *p*s < .006, all *d*s > 0.71, for the single-interaction condition (for an overview of all individual performances changes, see Table 1). In contrast, and as expected, there were no significant differences in performance changes as a function of the particular group member's capability in the nominal-group condition, all *t*s(20) < 1.27, all *p*s > .222, all *d*s < 0.28. Additionally, post hoc *t*-tests against zero revealed that in both continuous-interaction and single-interaction groups only the medium and the least capable members improved their estimation accuracy from phase one to phase two, all *t*s(20) > 3.44, all *p*s < .003, all *d*s > 0.75, and all *t*s(18) > 2.95, all *p*s < .009, all *d*s > 0.67, respectively. In contrast, the most capable group members' performance remained, more or less, stable in both conditions, $t(20) = 0.18$, $p = .863$, $d = 0.04$, and $t(18) = 0.89$, $p = .384$, $d = 0.20$, respectively. These differences in increased accuracy depending on group members' judgment accuracy indicate that group members understand from whom to learn or at least whom to ignore. Apparently, this understanding already occurs during the very first within-group interaction. Superior group member seems to share task relevant knowledge that can and should be learned. Consequently, only inferior group members can benefit from G-I transfer. In contrast to this improved estimation accuracy, in the nominal-group condition, none of the group members significantly changed their performance between the two phases, all *t*s(20) < 1.68, all *p*s > .108, all *d*s < 0.37. Hence, the interaction effect of group member and condition is the result of stronger performance enhancements of inferior group members after interacting with superior group members. This finding further supports the idea of G-I transfer and indicates that the most capable group members were the source of the learning process. In contrast, the medium and least capable group members seemed to benefit from the most capable members, leading them to approximate their levels of individual accuracy.

In addition, we conducted a more detailed temporal analysis of the observed individual performance enhancements in the two interacting group conditions to determine when the major increase in individual accuracy occurs. For this purpose, we compared group members' averaged individual accuracy of the trials before the first group interaction (trial 1-11) with the trial immediately after the first group interaction (trial 12), and then the averaged individual accuracy of the remaining 8 trials (trial 13-20). Doing so allowed us to analyze whether the performance enhancement after the first group interaction is relatively stable over time. Accordingly, we conducted a 2 (*group type*: continuous-interaction vs. single-interaction) × 3 (*trial block*: practice phase vs. trial after first group interaction vs. remaining 8 trials) repeated measures ANOVA. This analysis revealed a main effect of trial block, $F(2, 37) = 35.57$, $p < .001$, $\eta_p^2 = .48$, and no main effect of group type, $F(1, 38) = 0.32$, $p = .574$, $\eta_p^2 < .01$, or interaction of group type and trial block, $F(2, 37) = 0.33$, $p = .724$, $\eta_p^2 < .01$. Post hoc paired-samples *t*-tests showed that in the continuous-interaction condition the average individual accuracy discontinuously increased after the first group interaction ($M = 48.63$, $SD = 19.97$ vs. $M = 26.30$, $SD = 14.17$), $t(20) = 6.58$, $p < .001$, $d = 1.44$, and remained (more or less) stable afterwards ($M = 26.30$, $SD = 14.17$ vs. $M = 29.73$, $SD = 7.01$), $t(20) = -1.01$, $p = .324$, $d = 0.22$. Moreover, the same was true for the single-interaction condition ($M = 47.99$, $SD = 15.01$ vs. $M = 29.69$, $SD = 11.66$), $t(18) = 5.05$, $p < .001$, $d = 1.16$, and ($M = 29.69$, $SD = 11.66$ vs. $M = 32.13$, $SD = 8.29$), $t(18) = -1.13$, $p = .274$, $d = 0.26$, respectively (see also Figure 1). This result suggests that a single group interaction is sufficient to ensure the stability of the observed G-I transfer, in line with Hypothesis 2a.[7]

[7] Additionally, we split the 20 trials in four blocks (trial 1-6, 7-11, 12-16, 17-20) to analyze more evenly sized blocks of trials. Corresponding post hoc paired-samples *t*-tests only revealed significant changes in group members' individual accuracy right after the first group interaction in both the continuous-interaction condition ($M = 49.86$, $SD = 22.39$ vs. $M = 31.13$, $SD = 9.92$), $t(20) = 3.84$, $p = .001$, $d = 0.88$, and the single-interaction condition ($M = 50.67$, $SD = 18.37$ vs. $M = 31.18$, $SD = 8.88$), $t(18) = 4.65$, $p < .001$, $d = 1.09$ (all other *t*s < 1.58; all other *p*s > .130). Furthermore, there was no significant linear trend towards an individual accuracy decrease or improvement after the first group interaction for the continuous-interaction nor the single-interaction condition (all *F*s < 1.48, all *p*s > .240). These analyses further support the assumption of immediate performance enhancements right after the first group interaction with no additional improvements on subsequent trials.

Changes in metric and mapping error

Beyond that, we were interested in what members of interacting groups actually learn. To this end, we calculated the mean overall deviation (MOD) (Brown & Siegler, 1993), which is a measure of metric property defined as the absolute difference between the median estimate across all items and the true overall median and is therefore less susceptible for outlier than the arithmetic mean. Accordingly, lower values indicate a lower metric error. However, the magnitude of participants' judgment errors covaried strongly with the respective true values. Hence, we worked with the percentage error instead of the absolute deviation from the true values. Nevertheless, the pattern of results remains unchanged when working with the median absolute error instead of the median absolute percentage error in both experiments. Similar to the analyses of group members' MAPE scores, we compared the differences of individual MOD scores between the individual practice phase and the group phase in the three experimental conditions. Thus, positive values indicate decreasing metric errors. Again, the trial right before the first group interaction was treated as the last trial of the individual practice phase and we averaged across group members for reasons of simplicity.[8] We calculated an ANOVA with group type (continuous-interaction vs. single-interaction vs. no interaction) as between-subjects factor, which showed significant differences, $F(2, 58) = 9.50$, $p < .001$, $\eta_p^2 = .25$. Whereas participants in the continuous-interaction condition and in the single-interaction condition decreased their metric errors ($M = 24.38$, $SD = 23.95$, and $M = 15.38$, $SD = 22.83$), this error even increased for participants in the nominal group condition (M = -3.52; SD = 15.91). The differences between the two interacting groups condition and the nominal group condition were significant, p < .001 and p = .007. In contrast, the difference between continuously interacting groups and single interaction groups fell short of significance (p =

[8] The pattern of results when adding group member as a within-subjects factor mirrors the general finding of stronger performance enhancements of inferior group members, such that metric error reductions were stronger for inferior group members as well in both experiments.

.184) although, descriptively, the metric error reduction was more pronounced among the former than among the latter. In sum, we found evidence that interacting with others reduces group members' metric error, which is line with Hypothesis 4.

Besides the metric error, we were also interested in possible changes in mapping errors, i.e. whether participants were able to put different target values in the correct order. Therefore, we computed rank-order correlations, which represent the correlation between the ranks of estimates with the ranks of true values (Brown & Siegler, 1993), and calculated the difference between group members' Fisher z-transformed averaged rank-order correlation coefficients (Spearman's rho) during the group phase and the individual practice phase. Accordingly, positive values indicate decreasing mapping errors. Similar to the metric error analysis, we calculated an ANOVA with group type (continuous-interaction vs. single-interaction vs. no interaction) as between-subjects factor, which revealed no significant differences, $F(2, 54) = 1.71$, $p = .191$, $\eta_p^2 = .06$. Hence, whether participants interacted with others or not had no effect on their mapping error.

Exploratory analyses: Group-level data

In an exploratory fashion, we investigated whether interacting groups outperformed nominal groups. For this purpose, we calculated the MAPE score for the 10 trials of the group phase. In the continuous-interaction condition, these MAPE scores were based on the groups' consensus estimates, whereas in the nominal-group condition these scores were calculated as the average of the three nominal-group members' individual estimates. In the single-interaction condition, the groups' average MAPE score was a composite measure of the group estimate in the first trial of the second phase and the averaged individual estimates in the remaining nine trials. Based on these calculations, our first analysis was an ANOVA with group type (continuous-interaction vs. single-interaction vs. no interaction) as a between-subjects factor. This analysis showed no significant effect of group type, $F(2, 58) =$

1.56, p = .219, η_p^2 = .05. Descriptively, the results indicate that both the continuous-interaction groups (M = 25.97, SD = 9.80) and the single-interaction groups (M = 25.49, SD = 11.07) performed somewhat better than the nominal groups (M = 32.05, SD = 17.46). However, due to the relatively high variances within the conditions, the superiority of interacting over non-interacting groups fell short of significance.

Furthermore, we tested whether the performance of continuously interacting groups exceeded the average model that was calculated on the basis of their members' individual estimates right before each of the group trials (i.e., the estimates that already benefitted from G-I transfer). If this were the case, it would indicate that groups differentially weight the proposals of superior members more strongly than the proposals of weaker members. We excluded the first trial of the second phase in order not to artificially penalize the average model and averaged across the remaining trials.[9] The corresponding paired samples t-tests showed that the actual group performance was slightly (but not significantly) inferior to the average model (M = 25.09, SD = 9.51 vs. M = 23.42, SD = 8.25), t(20) = 1.11, p = .282, d = 0.24. In general, our results indicate that groups in Experiment 1 were not able to outperform the average model.

However, we cannot yet rule out that the results of Experiment 1 might be task specific, because we used the same task as Schultze et al. (2012). For example, our participants might have had a more or less accurate representation of the map of Europe, which, in turn, could have facilitated learning processes when receiving an accurate point of reference. Beyond that, the task was characterized such that participants were as likely to underestimate as to overestimate the true values. Consequently, group members could have accomplished

[9] Individuals cannot yet benefit from G-I transfer on the first trial of the second phase. The group judgments, however, can, lead to an "unfair" advantage over the averaged previous individual judgments, which, in turn, would distort the results regarding whether or not groups engage in differential weighting strategies.

individual performance enhancements similar to G-I transfer, by simply centering their individual estimates. Therefore, the question is whether our findings would still hold if the task were more difficult and if participants would tend to over- or underestimate the true value. Furthermore, the sequence of trials in Experiment 1 – and also in both experiments by Schultze et al. (2012) – was in a fixed order. Therefore, we cannot rule out that differences in the difficulty of the different trials had an influence on the magnitude of the observed G-I transfer or changes in metric and mapping error. Hence, to validate our findings in terms of replicability and generalizability, we conducted a second experiment with a different task type and a randomized trial order.

Experiment 2

To generalize our findings, we conducted a second experiment with the same design but a different task type, namely estimating the weights of different objects. The task was considerably more difficult and characterized by a strong population bias (see section task and procedure). In spite of these differences, we expected to replicate the results of Experiment 1 with respect to individual performance enhancements. Particularly when taking into account that evidence on G-I transfer in quantitative estimation tasks is extremely scarce so far, we consider a replication of our results as being indispensable.

Method

Participants and design

A total of 252 German or German-speaking students (152 women, 100 men) with an average age of 23.25 (SD = 4.53) years participated in the experiment, with three persons each forming a (real or nominal) group. Experiment 2 used the same mixed factorial design as in Experiment 1, with the group type (continuous-interaction, single-interaction, no interaction) as a between subjects variable and task trial (or trial block) as a within subjects variable.

Task and procedure

The procedure of Experiment 2 was identical to Experiment 1, with the following exceptions: First, participants were asked to estimate the weight of different small items (e.g., hammer, dustpan, or umbrella) that were present in the room, without being allowed to touch or lift them. We chose this task based on two pretests (*N* = 30 and *N* = 29) revealing that there were stable differences in participants' individual performance (mean Spearman's Rho = .39, *p* = .031 and mean Spearman's Rho = .49, *p* = .008, respectively). Furthermore, this task was evidently more difficult than the task of Experiment 1: The average MAPE scores in the two pretests of Experiment 2 were markedly above the corresponding scores in the two pretests of Experiment 1 (*M* = 293.02, *SD* = 234.63 and *M* = 398.71, *SD* = 261.07 vs. *M* = 62.22, *SD* = 63.55 and *M* = 48.31, *SD* = 21.14). Beyond that, in contrast to the pretests of Experiment 1, participants had a strong tendency to overestimate the true values. When calculating participants' mean percent deviation from the true values, these average deviations were significantly greater than zero (*M* = 281.36, *SD* = 241.84), *t*(29) = 6.37, *p* < .001, *d* = 1.16, and (*M* = 395.80, *SD* = 263.79), *t*(29) = 8.08, *p* < .001, *d* = 1.50, respectively, indicating a large population bias. The second change was that we aimed to rule out that the results obtained in Experiment 1 were in any way due to the fixed order of trials. To this end, we randomly created four different trial orders in Experiment 2 by splitting the 20 trials into two task blocks of 10 trials each. Half of the participants worked on the first block in the individual practice phase and on the second block in the group phase, whereas this order was reversed for the other half. In each sequence, we additionally reversed the order of trials within the two blocks for half of the participants.

Results and Discussion

Group-to-individual transfer

Similar to Experiment 1, we started by testing for increased accuracy of group members' individual estimates consistent with G-I transfer. For this purpose, we again calculated individual performance enhancements by subtracting the individual MAPE scores of the group phase from those of the individual practice phase. Again, the first trial of the group phase (i.e., the trial right before the first group interaction) was counted as the last trial of the individual practice phase, since this trial could not, by definition, be affected by any group interaction. With the MAPE scores as dependent variable, we conducted a 3 (*group type*: continuous-interaction vs. single-interaction vs. no interaction) × 3 (*group member*: most capable vs. medium vs. least capable) ANOVA with group type as between-subjects factor and group member as within subjects factor.[10] This analysis revealed a main effect of group type, $F(2, 81) = 8.39$, $p < .001$, $\eta_p^2 = .17$. LSD post hoc comparisons showed that the performance enhancements in the continuous-interaction condition were somewhat stronger than those in the single-interaction condition, but the comparison did not reach conventional levels of significance ($M = 87.33$, $SD = 62.90$ vs. $M = 49.24$, $SD = 100.60$), $p = .077$. As the more detailed temporal analysis reported below will clarify, this descriptive difference is indeed most likely due to random variation. As in Experiment 1, participants in both the continuous-interaction groups as well as in the single-interaction groups increased their performance significantly more than participants in the nominal groups ($M = 87.33$, $SD = 62.90$ vs. $M = 0.79$, $SD = 70.48$), $p < .001$, and ($M = 49.24$, $SD = 100.60$ vs. $M = 0.79$, $SD = 70.48$), $p = .026$, respectively. Furthermore, post hoc *t*-tests against zero showed a significant increase in

[10] We conducted the same ANOVA with the four randomly created trial orders as an additional between-subjects factor. This analysis revealed no significant interaction of trial order and group type, $F(6, 72) = 0.46$, $p = .838$, $\eta_p^2 = .04$, indicating that differences between the three conditions cannot be attributed to differences in the order of trials. Therefore, we dropped the order of trials as a between-subjects factor in all analyses.

individual accuracy in the continuous-interaction condition, $t(27) = 7.35$, $p < .001$, $d = 1.39$, as well as in the single-interaction condition, $t(27) = 2.59$, $p = .015$, $d = 0.49$.[11] The nominal-group condition, in contrast, showed virtually no change in MAPE scores from the first to the second phase, $t(27) = 0.06$, $p = .953$, $d = 0.01$, indicating that, similar to Experiment 1, participants in the nominal-group condition did not improve their performance in terms of practice effects. Hence, the increases in individual accuracy in the other two conditions are the result of G-I transfer, in line with Hypothesis 1. Furthermore, participants did not manifest a significantly stronger G-I transfer after multiple group interactions as compared to a single group interaction, which supports Hypothesis 3a over Hypothesis 3b.[12]

Beyond that, the ANOVA revealed a main effect of group member, $F(2, 80) = 40.00$, $p < .001$, $\eta_p^2 = .31$, and an interaction of group member and group type, $F(4, 162) = 4.72$, $p = .001$, $\eta_p^2 = .10$. Separate post hoc paired-samples *t*-tests for the continuous-interaction and the single-interaction conditions showed that the accuracy improvements differed between all levels of group member expertise for the continuous-interaction condition, all $ts(27) > 4.87$,

[11] As a closer inspection of the data showed that a substantial part of the descriptive difference between performance enhancements in the continuous-interaction and the single-interaction condition derives from one participant in the latter condition whose performance dramatically decreased by 670 percentage points. When excluding this participant's group, the difference between the two group conditions decreases markedly ($M = 87.33$, $SD = 62.90$ vs. $M = 60.12$, $SD = 84.08$), $p = .170$, although some moderate difference remains. Furthermore, the effect size of the post hoc *t*-tests against zero in the single-interaction condition increases, $t(27) = 3.72$, $p < .001$, $d = 0.71$.

[12] Again, participants' gender didn't significantly affect the individual performance change from phase one to two, $t(250) = 0.75$, $p = .456$, $d = 0.01$. Furthermore, there was also no significant relationship between participants' age and the dependent variable, $r = .11$, $p = .073$, although, descriptively, younger participants increased their performance somewhat more strongly than older participants. Furthermore, the results revealed differences in participants' age between the conditions that fell short of significance, $F(2, 249) = 2.29$, $p = .057$, $\eta_p^2 = .02$. However, LSD post hoc comparisons showed that the participants were significantly older in the single-interaction condition than in the nominal-group condition, $p = .017$, whereas there were no differences between the continuous-group condition and the other two (all *p*s > .205). Accordingly, participants' age should, if at all, lead to somewhat smaller performance enhancements in the single-interaction condition than in the nominal-group condition and, thereby, work against our hypothesis. Hence, the weak relationship between performance changes and participants' age should not interfere with our results. Finally, there was no evidence for baseline accuracy differences between the conditions, $F(2, 81) = 0.29$, $p = .751$, $\eta_p^2 < .01$, or for average group member accuracy variance differences between condition, $F(2, 81) = 0.42$, $p = .658$, $\eta_p^2 = .01$.

all *p*s < .001, all *d*s > 0.92, and for the single-interaction condition, all *t*s(27) > 2.44, all *p*s < .022, all *d*s > 0.46. Again, post hoc *t*-tests against zero revealed significant performance increases for the medium and the least capable members in both interacting group conditions, all *t*s(27) > 2.22, all *p*s < .035, all *d*s > 0.41 (for an overview of all individual performance changes, see Table 2). In contrast, there was a tendency for the most capable group members' performance to slightly deteriorate in the continuous-interaction condition, *t*(27) = -1.87, *p* = .073, *d* = 0.35, and even more profoundly in the single-interaction condition, *t*(27) = -3.13, *p* = .004, *d* = 0.59. A similar analysis of the non-interacting nominal groups unexpectedly revealed a significant difference in performance changes between the most capable and the least capable group members, *t*(27) = 2.90, *p* = .007, *d* = 0.55, as well as marginal differences between the medium and least capable group members, *t*(27) = 2.03, *p* = .052, *d* = 0.38. However, these differences are unlikely to stem from systematic learning effects; instead, they are likely the result of regression to the mean. Specifically, the least capable group members significantly increased their performance between the two phases, *t*(27) = 2.30, *p* = .029, *d* = 0.44, whereas there were no significant performance changes for the most capable and medium group members, all *t*s(27) < 1.05, all *p*s > .307, all *d*s < 0.20. This finding also suggests that not all performance changes in the interacting group conditions can be necessarily attributed to social learning processes. At least a small part might also be ascribed to statistical regression. However, post hoc *t*-tests revealed that the medium and least capable group members in both the continuous-interaction and the single-interaction condition increased their estimation accuracy more strongly than the medium and least capable members of the nominal groups, all *t*s(54) > 2.08, all *p*s < .043, all *d*s > 0.55. Hence, the interaction effect of group member and condition mainly derives from stronger performance enhancements of inferior group members after interacting with others.

Similar to Experiment 1, we were interested in a temporal analysis of the individual performance enhancements in the two interacting group conditions. Therefore, we again compared individual accuracy (averaged across group members) before any group interaction had taken place (trial 1-11) with the trial after the first group interaction (trial 12), and with the averaged individual accuracy of the remaining 8 trials (trial 13-20). The 2 (*group type*: continuous-interaction vs. single-interaction) × 3 (*trial block*: practice phase vs. trial after first group interaction vs. remaining 8 trials) repeated measures ANOVA showed a main effect of trial block, $F(2, 53) = 21.28$, $p < .001$, $\eta_p^2 = .28$, and no main effect of group type, $F(1, 54) = 0.63$, $p = .432$, $\eta_p^2 = .01$, or interaction of group type and trial block, $F(2, 53) = 1.69$, $p = .190$, $\eta_p^2 = .03$. Post hoc paired samples *t*-tests revealed that the average individual accuracy in the continuous-interaction condition discontinuously increased after the first group interaction ($M = 237.04$, $SD = 81.49$ vs. $M = 148.29$, $SD = 110.84$), $t(27) = 5.29$, $p < .001$, $d = 1.00$, and remained (more or less) stable afterwards ($M = 148.29$, $SD = 110.84$ vs. $M = 149.89$, $SD = 85.76$), $t(27) = -0.09$, $p = .932$, $d = 0.02$. The same was true for the single-interaction condition, with a major performance enhancement directly after the single group interaction ($M = 229.48$, $SD = 98.59$ vs. $M = 180.04$, $SD = 111.37$), $t(27) = 2.33$, $p = .027$, $d = 0.44$, and virtually no further changes in individual estimation accuracy ($M = 180.04$, $SD = 111.37$ vs. $M = 180.26$, $SD = 112.08$), $t(27) = -0.02$, $p = .987$, $d < 0.01$. This result, illustrated in Figure 2, once more suggests that a single group interaction is sufficient to induce stable G-I transfer, which supports Hypothesis 2a.[13]

Figure 2 also indicates stronger individual performance enhancements in the

[13] Again, we split the 20 trials in four blocks (trial 1-6, 7-11, 12-16, 17-20) to analyze more evenly sized blocks of trials. Post hoc paired-samples *t*-tests revealed significant changes in group members' individual accuracy right after the first group interaction in both the continuous-interaction condition ($M = 249.91$, $SD = 110.13$ vs. $M = 153.24$, $SD = 90.39$), $t(27) = 7.13$, $p < .001$, $d = 1.35$, and the single-interaction condition ($M = 237.55$, $SD = 125.16$ vs. $M = 179.27$, $SD = 103.94$), $t(27) = 3.31$, $p = .003$, $d = 0.63$ (all other *t*s < 1.32, all other *p*s > .198). Beyond that, neither the continuous-interaction nor the single-interaction condition showed a significant linear trend towards an individual accuracy decrease or improvement after the first group interaction (all *F*s < 1.53, all *p*s > .227), which further supports the idea of a stable G-I transfer.

continuous-interaction condition as compared to the single-interaction condition. However, as this analysis shows, it is highly unlikely that this difference is due to the sustained group interaction in the former condition, because the full difference is already present after the first group trial – in other words, at a point in the experiment where the procedure has yet been identical for both conditions – and it remains stable afterwards. Hence, by chance, participants in the former condition seem to have reacted somewhat more strongly to the first group interaction than participants in the latter condition.

Changes in metric and mapping error

Similar to Experiment 1, we were interested in what participants learn when interacting with others. To this end, we calculated an ANOVA with group type (continuous-interaction vs. single-interaction vs. no interaction) as between-subjects factor and differences between the averaged group members' individual median percentage error between the individual practice phase and the group phase as dependent variable. This analysis revealed a significant main effect of group type, $F(2, 81) = 10.40$, $p < .001$, $\eta_p^2 = .20$. Additional LSD post hoc comparisons showed no significant differences in metric error reduction between the continuous-interaction and the single-interaction condition ($M = 68.47$, $SD = 60.50$ vs. $M = 41.43$, $SD = 82.25$), $p = .176$. In contrast, participants' reductions in metric error were significantly stronger in both the continuous-interaction groups and the single-interaction groups than in the non-interacting nominal groups ($M = 68.47$, $SD = 60.50$ vs. $M = -19.77$, $SD = 74.68$), $p < .001$, and ($M = 41.43$, $SD = 82.25$ vs. $M = -19.77$, $SD = 74.68$), $p = .003$, respectively. Hence, interacting with others reduced group members' metric error, which supports Hypothesis 4.

Furthermore, we analyzed changes in participants' mapping errors averaged across group members. To this end, we calculated an ANOVA with group type (continuous-interaction vs. single-interaction vs. no interaction) as between-subjects factor and the difference between participants' Fisher z-transformed rank-order correlation coefficients (Spearman's rho) during the group phase and the individual practice phase as dependent variable. This analysis revealed no significant differences between the group types, *F*(2, 81) = 1.12, *p* = .307, η_p^2 = .03. Hence, there were no systematic difference in participants' mapping error changes.

Exploratory analyses: Group-level data

Similar to Experiment 1, we also analyzed group performance for exploratory purposes and checked whether interacting groups' judgments were more accurate than those of nominal groups. Accordingly, we calculated the groups' average MAPE score for the 10 trials of the group phase in the same way as in Experiment 1. Hence, in the single-interaction condition, the groups' average MAPE score was a composite measure of the group estimate of the first trial of the second phase and the averaged individual estimates of the remaining nine trials. Afterwards, we ran an ANOVA with the group type (continuous-interaction vs. single-interaction vs. no interaction) as a between-subjects factor and (nominal or real) group performance as the dependent variable. This analysis revealed a significant effect of group type, *F*(2, 81) = 4.02, *p* = .022, η_p^2 = .09. LSD post hoc comparisons showed that the accuracy of both the continuous-interaction groups and the single-interaction groups were superior to the nominal groups (*M* = 142.30, *SD* = 82.06 vs. *M* = 220.21, *SD* = 110.29), *p* = .006, and (*M* = 175.08, *SD* = 114.42 vs. *M* = 220.21, *SD* = 110.29), *p* = .106, respectively, even though the latter comparison did not reach conventional levels of significance. Beyond that, continuous-interaction and single-interaction groups did not differ significantly with regard to accuracy (*M* = 142.30, *SD* = 82.06 vs. *M* = 175.08, *SD* = 114.42), *p* = .238.

Finally, we were interested in the possible occurrence of functional differential weighting strategies. To this end, and similar to Experiment 1, we tested whether interacting groups outperformed the average of their members' individual estimates after controlling for

G-I transfer. A paired samples *t*-tests showed that, on average, group estimates were significantly more accurate than the average model; the difference in accuracy was about 11 percentage points (*M* = 133.99, *SD* = 80.86 vs. *M* = 144.07, *SD* = 87.09) *t*(27) = -3.37, *p* = .002, *d* = 0.64. Hence, continuously interacting groups were apparently willing and able to assign different weights to their members' individual estimates, and they did so in an effective manner, allowing them to outperform the average of their members' estimates. Hence, Experiment 2 constitutes, to our knowledge, the first evidence for functional differential weighting after controlling for G-I transfer.

These findings raise an interesting question: Why did we find evidence for group members weighting their individual contributions differentially in Experiment 2, but not in Experiment 1, and also not in the study by Schultze et al. (2012)? Of course, at this point we can only speculate about this, but we find it, at least, plausible that parts or all of this divergence could be due to differences between the tasks that were used in these experiments: Whereas in Experiment 1 we used the same distance estimation task that had been used by Schultze et al. (2012), and with similar results (no differential weighting), in Experiment 2 we introduced a new weight estimation task. As already stated, this task was characterized by a large population bias, whereas the distance estimation task of Experiment 1 contained no such bias. Now, the presence vs. absence of a population bias should have consequences for whether or not the group members' individual values can be expected to bracket the true value: If there is no population bias, that is, if over- and underestimations cancel out on average, the group members' individual estimates should often bracket the true value. In contrast, if all group members systematically overestimate (or underestimate) the true value, bracketing is less likely to occur. In line with this, we found that the percentage of overall cases where the individual estimates bracketed the true value in Experiment 1 was far above the bracketing rate in Experiment 2 (*M* = 64.05, *SD* = 14.37 vs. *M* = 32.32, *SD* = 21.54).

Because more bracketing means that averaging more often leads to accurate results, differential weighting had better chances to pay off in Experiment 2 as compared to Experiment 1.

General discussion

In the present study, we tested whether group members cooperatively working on estimation tasks benefit from G-I transfer. Based on previous findings (Laughlin et al., 2008; Schultze et al., 2012), we expected individual performance enhancements due to group discussion. More specifically, we postulated a stable increase in group members' individual accuracy that persists even when the group is disbanded. Furthermore, we aimed to clarify whether a single group interaction is sufficient to produce G-I transfer, and whether further within-group interaction induces additional performance enhancements. Beyond that, we also wanted to shed some light on what exactly group members learn when interacting with others. We expected a transfer of metric knowledge that should lead to a better calibration of group members' estimates, but also tested the possibility that interaction improves group members' mapping knowledge. In an exploratory manner, we checked for the superiority of groups over an equivalent number of individuals, and for the occurrence of possible differential weighting strategies that improve the group judgments beyond the level of individual capability gains.

In line with our assumptions, the results of our experiments provide evidence for socially induced learning processes as a consequence of group interaction on quantitative estimation tasks. The estimates of superior group members served as a benchmark towards which the inferior group members adjusted their subsequent individual estimates. More precisely, group members reduced their metric but not their mapping error. Importantly, the individual increases in accuracy remained stable even if we disbanded groups after their first group discussion. Since participants in the nominal group condition were not able to enhance their performance over time, the aforementioned result can be interpreted as unequivocal evidence for G-I transfer. Furthermore, additional group interactions did not lead to further increases in individual accuracy, suggesting that groups were able to exchange all information necessary to induce this G-I transfer during their first group discussion. Beyond that, we found first evidence that after group members benefitted from G-I transfer groups are indeed able to assign more weight to more accurate judgments under certain circumstances. We will get back to this finding after having discussed our central results regarding G-I transfer.

Group-to-individual transfer

Our main aim was to test for the relevance of group interaction for the subsequent estimation accuracy of the individual group members. Our results allow us to draw several conclusions: First, one group interaction is sufficient to induce a stable G-I transfer, since in both experiments the performance remained on the improved level even after the group was disbanded. These results are in line with the assumptions of Schultze et al. (2012) and, to the best of our knowledge, constitute the first unambiguous evidence for the stability of this performance enhancement on estimation tasks, thereby mirroring similar findings from the field of problem-solving tasks (Laughlin et al., 2008). This finding also rules out an alternative explanation for the performance changes. The performance enhancement cannot be reduced to changes in group members' motivation because discussing their estimates increases the accountability they feel for their judgment. Otherwise, the individual performance should have dropped after the group was disbanded. Second, the replicability of the G-I-transfer with a different task speaks to the generalizability of this phenomenon at least when tasks have similar characteristics as the ones we used.

Third, the fact that we found strong metric error reductions, and that the performance enhancement already occurred after one group interaction on two different task types allows some speculations regarding what people learn when interacting with others on estimation tasks, and what information has to be exchanged in order to produce the observed G-I-transfer. In our opinion, the most likely explanation for the strong reduction of group members' metric error and the rapid increase in estimation accuracy is the exchange of reference values during the first group interaction. As we know from previous research, frames of reference play an important role when it comes to increases in estimation accuracy (e.g., Bonner & Baumann, 2008; Bonner et al., 2007; Laughlin et al., 1999; Laughlin et al., 2003). Such reference values are relatively easy to communicate and to retain, and they should have a beneficial effect on all subsequent judgments in the same domain. As Schultze et al. (2012) discuss, points of reference might provide a basis for better calibration and could enable group members to reduce their individual estimation bias. With this additional information, individuals might be able to improve their individual accuracy, even without any further benchmarks, on subsequent trials. For example, group members might communicate the length of Germany from north to south (approx. 900 km) as a reference value, which might help them when estimating the distance between London and Rome and will prevent very inaccurate estimates. In other words, accurate benchmarks could also serve as a source of error checking. This assumption might also explain the lack of G-I-transfer in an earlier study that used an experimental design somewhat similar to ours. In one condition of this study, Sniezek and Henry (1990) asked participants to estimate the prices of different automobile models individually before and after interacting with others. During this interaction, group members were allowed to exchange all information relevant to the task with exception of numeric estimates. In other words, they could not provide reference values and, therefore, not reduce their metric error. Nevertheless, future research should systematically investigate the exact nature of the information required to induce G-I-transfer and identify what group interactions provide beyond the beneficial effect of receiving accurate reference values.

Superiority of group judgments and differential weighting

Not only the individual group members but also the groups as a whole seem to benefit from G-I transfer. The results of our experiments generally support the idea that interacting groups outperform nominal groups in quantitative estimation tasks. Although the respective comparisons only reach conventional levels of significance for continuously interacting groups in Experiment 2, where groups were able to benefit from G-I transfer and differential weighting, descriptively groups were more accurate than nominal groups in both experiments. One reason why this comparison was not significant in Experiment 1 is the remarkably good estimation accuracy of the nominal groups in this experiment (*M* = 32.05, *SD* = 17.46). In contrast, the performance of nominal groups in the (in large parts similar) first experiment of Schultze et al. (2012), who found a significant superiority of interacting groups over nominal groups with the same task type, was notably lower (*M* = 39.54, *SD* = 38.08). Presumably, by chance, nominal groups in our first experiment might have consisted of individuals whose idiosyncratic biases cancelled each other out more frequently than in the experiment of Schultze et al. (2012).

The results regarding the occurrence of differential weighting differ between our two experiments. In Experiment 1, where simple averaging was a rather effective strategy due to the lack of a population bias and small remaining differences in individual accuracy, we replicated the results of Schultze et al. (2012) that groups do not outperform the average of their group members' contributions. However, the fact that we found no evidence for differential weighting in Experiment 1 does not mean that group members necessarily weighted their individual contributions equally. The lack of bias in participants' estimates only implies that differential weighting wouldn't improve the group judgments as strong and, as a consequence, wouldn't be detectable with a performance-based proxy assessment of weighting. Nevertheless, as the actual group performance was even slightly inferior to the average model, differential weighting was not beneficial in Experiment 1. In contrast, groups engaged in functional differential weighting in Experiment 2, which employed a task favoring weighting by expertise or accuracy over averaging, due to a strong population bias (e.g., Davis-Stober, Budescu, Dana, & Broomell, 2014; Einhorn et al., 1977). This allowed interacting groups in Experiment 2 to outperform the simple average of their members' individual estimates. Taken together, our results suggest that groups can – to some degree – engage in rather functional weighting strategies. Our findings, thus, provide an interesting basis for systematic research on the weighting strategies groups employ – for example, by investigating how various task and group characteristics relevant to the effectiveness of differential weighting influence the choice of the weighting strategy and its impact on group performance.

Limitations and directions for future research

Although our experiments provide evidence for stable G-I transfer in quantitative estimation tasks, we should also mention some limitations. Despite the fact that we used two different tasks with quite different characteristics that consistently produced G-I transfer, we still cannot exclude the possibility that other types of estimation tasks might yield different results. In the tasks we used, metric errors were rather common and – at times – extreme, as indicated by systematic idiosyncratic biases of group members in both experiments. This is particularly evident in Experiment 2, though, where participants systematically overestimated the weights of the small objects by an average factor of three. Mapping errors, on the other hand, might have been less pronounced, because most participants presumably had a rough recollection of the geographical location of the EU's member countries (if not necessarily the location of the capital cities within the countries), allowing them to distinguish long distances from short ones. Likewise, they could tell that a small plastic comb weighed less than a small metal hammer. Hence, the tasks we used had a great potential for metric error reductions whereas it impeded the occurrence of mapping error reductions as a consequence of interacting with others. Admittedly, this presumed combination of relatively low mapping and high metric errors might not generalize to all estimation tasks. For example, forecasting tasks, like predicting the return on a capital investment, are mainly characterized by mapping errors. In this case, the previous and current values of the variable that is to be predicted constitute rather reliable reference values that minimize the individual metric error. In contrast, there are several factors that should predominantly affect the mapping knowledge component. For example, when predicting the future market rate of a certain stock, one has to learn general market trends, as well as the previous prosperity and future plans of certain companies to reduce one's mapping error. All of these knowledge components and cues might be transferable through group discussion, quite similar to the exchange of reference values. However, in this case, G-I transfer should take more time to emerge, and also more time to fully develop. Therefore, it is crucial to replicate our findings with different types of quantitative estimation tasks, preferably tasks with a high ecological validity like forecasting tasks, or even in a real world setting. In general, our findings should be replicated with tasks of different complexity and different estimation biases to form an overall picture regarding the strength of G-I transfer on the one hand, and functional differential weighting strategies on the other hand, under different circumstances.

Furthermore, it remains an open question as to how groups manage to identify their group members' expertise or the accuracy of their judgments in order to know whom to learn from. Previous evidence regarding ad hoc groups' ability to recognize expertise is rather contradictory. On the one hand, some studies indicate that groups are capable of identifying their most capable members (e.g., Baumann & Bonner, 2004; Bonner et al., 2007; Henry et al., 1996; Libby, Trotman, & Zimmer, 1987). On the other hand, there is also evidence of groups failing to recognize the specific expertise of their members (e.g., Littlepage, Robinson, & Reddington, 1997, studies 1 and 2; Littlepage, Schmidt, Whisler, & Frost, 1995; Trotman, Yetton, & Zimmer, 1983). The fact that, in our experiments, the most capable members' performance remained largely stable, whereas the performance of the medium and least capable members considerably increased, speaks to the groups' ability to assess their group members expertise or at least the quality of their judgments. One possible determinant for the ability to recognize expertise might be the plausibility of individual estimations. As Yaniv and Kleinberger (2000) discuss, individuals might identify particularly poor estimates as out of the bounds of plausibility, even if people cannot generate correct estimates themselves. In others words, group members might have been reasonably good in realizing whom to ignore. This could also explain why there was no negative individual learning in our experiments. Nevertheless, further research should address the question of which circumstances facilitate the recognition of expertise or the accuracy of certain estimates, and what cues are relevant for groups to determine the relative expertise of their members.

Finally, we do not yet know whether group interaction is really *indispensable* to induce the phenomenon of G-I transfer. Since our results reveal strong learning effects after just one group interaction, this raises the question of whether similar processes might be possible even without any direct communication. As Farrell (2011) suggests, individual accuracy can be improved by knowing the estimates of other persons, without any form of group interaction (in terms of free information exchange). In other words, it is questionable whether discussing individual estimates with other people is crucial to individual learning, or whether the knowledge about others' judgments might be sufficient to achieve the same or at least a similar beneficial effect, at least in some tasks. Hence, a promising line of future research is to identify which factors are indispensable for individual learning effects and by which means group interaction might additionally strengthen these processes.

Conclusion

In accordance with the idea of G-I transfer, group members can learn relevant knowledge in quantitative estimation tasks by cooperatively working with others. One group interaction seems to be sufficient for an increase in metric knowledge that leads to more accurate individual judgments, whereas further group interaction does not foster additional capability gains. Furthermore, under certain circumstances, effective weighting strategies when combining those individual estimates with a group judgment might occur. Thus, we know now that a single group discussion can robustly improve group members' individual judgment accuracy and can also lead to an improved collaboration, although the specific mechanisms underlying these improvements are still an open topic for future research.

References

Baumann, M. R., & Bonner, B. L. (2004). The effects of variability and expectations on utilization of member expertise and group performance. *Organizational Behavior and Human Decision Processes, 93,* 89-101. http://dx.doi.org/10.1016/j.obhdp.2003.12.004

Bednarik, P., & Schultze, T. (2015). The effectiveness of imperfect weighting in advice taking. *Judgment and Decision Making, 10,* 265-276.

Bonner, B. L., & Baumann, M. R. (2008). Informational intra-group influence: the effects of time pressure and group size. *European Journal of Social Psychology, 38,* 46-66. http://dx.doi.org/10.1002/ejsp.400

Bonner, B. L., & Baumann, M. R. (2012). Leveraging member expertise to improve knowledge transfer and demonstrability in groups. *Journal of Personality and Social Psychology, 102*, 337-350. http://dx.doi.org/10.1037/a0025566

Bonner, B. L., Sillito, S. D., & Baumann, M. R. (2007). Collective estimation: Accuracy, expertise, and extroversion as sources of intra-group influence. *Organizational Behavior and Human Decision Processes, 103,* 121-133. http://dx.doi.org/10.1016/j.obhdp.2006.05.001

Brodbeck, F., & Greitemeyer, T. (2000a). A dynamic model of group performance: Considering the group members' capacity to learn. *Group Processes & Intergroup Relations, 3,* 159-182. https://doi.org/10.1177/1368430200003002004

Brodbeck, F. C., & Greitemeyer, T. (2000b). Effects of individual versus mixed individual and group experience in rule induction on group member learning and group performance. *Journal of Experimental Social Psychology, 36,* 621-648. http://dx.doi.org/10.1006/jesp.2000.1423

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 41, pp. 321–360). New York: Academic Press. http://dx.doi.org/10.1016/S0079-7421(02)80011-1

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*, 511-534. http://dx.doi.org/10.1037/0033-295X.100.3.511

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*, 79-101. http://dx.doi.org/10.1037/dec0000004

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*, 158-172. http://dx.doi.org/10.1037/0033-2909.84.1.158

Farrell, S. (2011). Social influence benefits the wisdom of individuals in the crowd. *Proceedings of the National Academy of Sciences*, *108*, E625. http://dx.doi.org/10.1073/pnas.1109947108

Henry, R. A. (1993). Group judgment accuracy: Reliability and validity of postdiscussion confidence judgments. *Organizational Behavior and Human Decision Processes*, *56*, 11-27. http://dx.doi.org/10.1006/obhd.1993.1043

Henry, R. A. (1995). Improving group judgment accuracy: Information sharing and determining the best member. *Organizational Behavior and Human Decision Processes, 62*, 190-197. http://dx.doi.org/10.1006/obhd.1995.1042

Henry, R. A., Strickland, O. J., Yorges, S. L., & Ladd, D. (1996). Helping groups determine their most accurate member: The role of outcome feedback. *Journal of Applied Social Psychology*, *26*, 1153-1170. http://dx.doi.org/10.1111/j.1559-1816.1996.tb02290.x

Laughlin, P. R., & Barth, J. M. (1981). Group-to-individual and individual-to-group problem-solving transfer. *Journal of Personality and Social Psychology*, *41*, 1087-1093. http://dx.doi.org/10.1037/0022-3514.41.6.1087

Laughlin, P. R., Bonner, B. L., Miner, A. G., & Carnevale, P. J. (1999). Frames of reference in quantity estimations by groups and individuals. *Organizational Behavior and Human Decision Processes*, *80*, 103-117. http://dx.doi.org/10.1006/obhd.1999.2848

Laughlin, P. R., Carey, H. R., & Kerr, N. L. (2008). Group-to-individual problem-solving transfer. *Group Processes & Intergroup Relations, 11*, 319-330. https://doi.org/10.1177/1368430208090645

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*, 177-189. http://dx.doi.org/10.1016/0022-1031(86)90022-3

Laughlin, P. R., Gonzalez, C. M., & Sommer, D. (2003). Quantity estimations by groups and individuals: Effects of known domain boundaries. *Group Dynamics: Theory, Research, and Practice*, *7,* 55-63. http://dx.doi.org/10.1037/1089-2699.7.1.55

Laughlin, P. R., & Jaccard, J. J. (1975). Social facilitation and observational learning of individuals and cooperative pairs. *Journal of Personality and Social Psychology*, *32,* 873-879. http://dx.doi.org/10.1037/0022-3514.32.5.873

Laughlin, P. R., & Sweeney, J. D. (1977). Individual-to-group and group-to-individual transfer in problem solving. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 246-254. http://dx.doi.org/10.1037/0278-7393.3.2.246

Libby, R., Trotman, K. T., & Zimmer, I. (1987). Member variation, recognition of expertise, and group performance. *Journal of Applied Psychology*, *72*, 81-87. http://dx.doi.org/10.1037/0021-9010.72.1.81

Littlepage, G., Robison, W., & Reddington, K. (1997). Effects of task experience and group experience on group performance, member ability, and recognition of expertise. *Organizational Behavior and Human Decision Processes, 69*, 133-147. http://dx.doi.org/10.1006/obhd.1997.2677

Littlepage, G. E., Schmidt, G. W., Whisler, E. W., & Frost, A. G. (1995). An input-process-output analysis of influence and performance in problem-solving groups. *Journal of Personality and Social Psychology, 69*, 877-889. http://dx.doi.org/10.1037/0022-3514.69.5.877

Peltokorpi, V. (2008). Transactive memory systems. *Review of General Psychology*, *12*, 378-394. http://dx.doi.org/10.1037/1089-2680.12.4.378

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative to differential weighting. *Organizational Behavior and Human Decision Processes*, *118*, 24-36. http://dx.doi.org/10.1016/j.obhdp.2011.12.006

Sherif, M. (1937). An experimental approach to the study of attitudes. *Sociometry*, *1*, 90-98. http://dx.doi.org/10.2307/2785261

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes, 43*, 1-28. http://dx.doi.org/10.1016/0749-5978(89)90055-1

Sniezek, J. A., & Henry, R. A. (1990). Revision, weighting, and commitment in consensus group judgment, *Organizational Behavior and Human Decision Processes, 45*, 66-84. http://dx.doi.org/10.1016/0749-5978(90)90005-T

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 780-805. http://dx.doi.org/10.1037/a0015145

Stasson, M. F., Kameda, T., Parks, C. D., Zimmerman, S. K., & Davis, J. H. (1991). Effects of assigned group consensus requirement on group problem solving and group members' learning. *Social Psychology Quarterly*, *54*, 25-35. http://dx.doi.org/10.2307/2786786

Trotman, K. T., Yetton, P. W., & Zimmer, I. R. (1983). Individual and group judgments of internal control systems. *Journal of Accounting Research*, *21*, 286-292. http://dx.doi.org/10.2307/2490948

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*, 260-281. http://dx.doi.org/10.1006/obhd.2000.2909

Figure 1. Mean absolute percent error (MAPE) of individual estimates by group type during Experiment 1. Lower scores indicate greater accuracy.

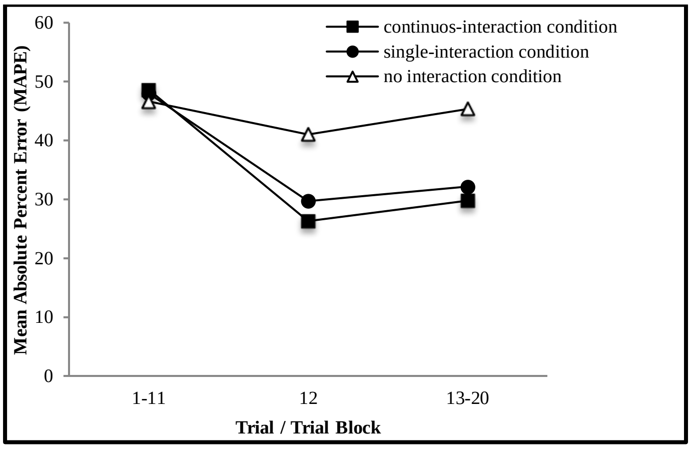

Figure 2. Mean absolute percent error (MAPE) of individual estimates by group type during Experiment 2. Lower scores indicate greater accuracy.

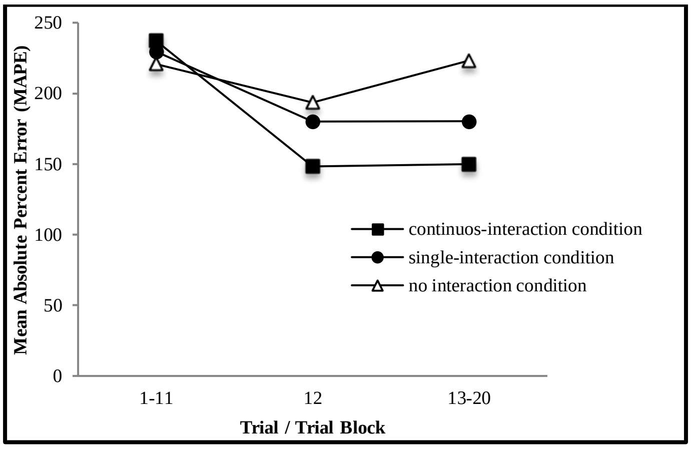

Table 1. *Group members' individual performance changes by group type in Experiment 1.*

	group member					
	most capable		medium		least capable	
group type	*M*	*SD*	*M*	*SD*	*M*	*SD*
continuous-interaction	-0.43	11.18	11.40	15.16	46.87	48.56
single-interaction	2.56	12.52	12.42	18.36	33.36	30.01
no interaction	-4.03	11.00	3.89	27.76	5.29	37.27

Table 2. *Group members' individual performance changes by group type in Experiment 2.*

	group member					
	most capable		medium		least capable	
group type	*M*	*SD*	*M*	*SD*	*M*	*SD*
continuous-interaction	-29.02	82.30	83.80	94.62	207.22	106.77
single-interaction	-58.72	99.41	45.89	109.36	160.34	239.04
no interaction	-11.88	71.25	-33.21	168.91	46.25	106.30

Anhang 2:

Social learning in the judge-advisor-system:
A neglected advantage of advice-taking

Social learning in the judge-advisor-system:

A neglected advantage of advice-taking

Alexander Stern, Thomas Schultze, & Stefan Schulz-Hardt,

University of Goettingen

Authors' Note

Alexander Stern, Thomas Schultze, and Stefan Schulz-Hardt, University of Goettingen, Institute of Psychology, Germany

Correspondence concerning this manuscript should be addressed to the first author: Alexander Stern, Institute of Psychology, Georg-August-University Göttingen, Waldweg 26, 37073 Göttingen, Germany.

E-mail: stern@psych.uni-goettingen.de

Telephone: +49 551 39 21114

Fax: +49 551 39 1821111

Abstract

Previous research in the judge-advisor paradigm has focused on how judges weight advice, and on the beneficial effect of receiving advice on judges' post-advice final judgments. However, a completely different possibility of how judges might benefit from advice has been overlooked so far: Social learning processes could improve the accuracy of judges' subsequent *initial* judgments as well. Hence, we test the assumption that advice can induce individual performance enhancements that differ as a function of the advisor's judgment accuracy. The results of three experiments support our hypothesis and indicate positive social learning particularly when participants receive high quality advice, which leads to diminished metric errors. Furthermore, we show that external feedback about the advisor's accuracy is not crucial for the occurrence of individual performance enhancements. In general, our results suggest that advice can have a positive effect on subsequent initial judgments in terms of social learning.

Keywords: judgment; decision making; advice taking; social learning; estimation accuracy

Social learning in the judge-advisor-system:
A neglected advantage of advice-taking

In recent years, both social and organizational psychological research has increasingly dealt with the process of using advice from others (e.g., Soll & Larrick, 2009; Yaniv & Kleinberger, 2000; for reviews, see Bonaccio & Dalal, 2006; Yaniv, 2004a). This topic is predominantly investigated in the so-called *judge advisor system* (JAS; Sniezek & Buckley, 1995). The JAS differentiates between an advisor, who provides information or recommendations, and a judge, who is responsible for the judgment. Usually, the judge first makes an initial estimate, receives a recommendation by the advisor, and then makes a final, possibly revised, estimate (e.g., Sniezek, Schrah, & Dalal, 2004; Yaniv, 2004b). In this final estimate, the judge combines his own initial judgment with the advice that he or she received – which, of course, can also mean that the judge fully sticks to his or her initial judgment, or completely takes the position advocated by the advisor.

Although research within this paradigm has revealed a variety of interesting results, a major part of judge advisor research concentrates on advice weighting and judgment accuracy, with three particularly pronounced and very robust findings. The first is that judges are sensitive to various cues of advice quality, leading them to heed better advice more (e.g., Harvey & Fischer, 1997; Soll & Larrick, 2009; Yaniv & Kleinberger, 2000). The second robust finding is *egocentric advice discounting*. Judges usually overweight their own opinion compared to the recommendation of the advisor (e.g., Yaniv, 2004b; Yaniv & Kleinberger, 2000). Finally, when the advisor provides an independent benevolent opinion, using advice leads to more accurate judgments and decisions compared to the uninfluenced initial judgments (e.g., Soll & Larrick, 2009; Sniezek et al., 2004). The reason is that aggregating independent opinions reduces unsystematic or even systematic errors (Soll & Larrick, 2009; Yaniv, 2004a).

However, even though the judge-advisor literature focuses on advice taking and the accuracy of post-advice judgments, there is also research concentrating on other effects of advice, like sharing the responsibility, minimizing the effort, or confirming the judge's initial opinion (Bonaccio & Van Swol, 2014). In line with this tradition, we want to examine another possible function of advice that has received little attention so far, namely social learning. In contrast to previous research investigating improvements in judges' *final* estimates, we aim to focus on changes in the accuracy of judges' *initial* estimates (after having received advice on previous occasions). We argue that, beyond improved *final* judgments, judges' ability to come up with accurate *initial* judgments might improve, too, due to social learning. In the following, we briefly review findings regarding social learning, and we then outline the possibility of performance enhancements of initial judgments in the JAS. Subsequently, we point out what exactly the judge can learn from the advisor in the standard JAS. Finally, we hypothesize about the potential moderating roles of advice quality and its salience.

Social learning in the JAS

Social learning theories deal with the influence of social information on people's behavior. For example, Bandura (1977) assumed that humans learn by means of observational and cognitive modeling processes. According to social learning theory, observational learning is governed by four processes: (a) attentional processes determine what is selectively observed, (b) retention processes involve transforming and restructuring information for memory representation, (c) production processes translate symbolic conceptions into appropriate courses of action, and (d) motivational processes emphasize different types of incentive motivators, which determine whether the observer actually executes the observed behavior. In line with this theory, many learning processes from direct experience can also be achieved vicariously by observing people's actions and their consequences (Bandura, 1986, Rosenthal & Zimmerman, 1978). Importantly, observational learning does not require direct social interaction to be effective, as shown in many different contexts such as mathematics (e.g., Schunk & Hanson, 1985), argumentative writing (Braaksma, Rijlaarsdam, & Van den Bergh, 2002; Couzijn, 1999; Raedts, Rijlaarsdam, Van Waes, & Daems, 2007), creative tasks (Groenendijk, Janssen, Rijlaarsdam, & Van den Bergh, 2013a, 2013b), and learning to collaborate (Rummel & Spada, 2005).

Previous research has already related advice to particular social learning processes. There is ample evidence that advice can improve the quality of subsequent decisions about the same issue (e.g., Biele, Rieskamp, & Gonzalez, 2009; Çelen, Kariv, & Schotter, 2010; Chaudhuri, Graziano, & Maitra, 2006; Kocher, Sutter, & Wakolbinger, 2014). For example, Biele et al. (2009) showed that one-time advice can have a positive effect on the upcoming individual performance on multiple trials of the same task. Even though independent decision makers improved their performance in the course of time, they did not catch up with those who received advice. Hence, advice can convey information that eliminates misperceptions and, thus, can help to find the right solution more quickly on one specific task. In line with these findings, advice taking in JAS experiments can be interpreted as a form of social learning as well. The advice leads to a reconsideration of one’s initial judgment, which, in turn, can improve subsequent final judgments (e.g., Farrell, 2010; Soll & Larrick, 2009; Sniezek et al., 2004). Both processes can be seen as *specific* social learning, because the benefit from the (social) information is limited to the subsequent individual performance on the *same* task. Hence, when using JAS terminology, they represent the beneficial effect of advice weighting. However, there is good reason to believe that the beneficial effect of receiving advice could also initiate *general* social learning, in terms of a *transfer* from one task to a different task of the same class (for a similar definition of specific and general transfer, see Stasson, Kameda, Parks, Zimmerman, & Davis, 1991). In other words, advice might not only affect judges' post-advice final judgments, but also improve their subsequent pre-advice initial estimates on subsequent tasks from the same domain.

But what exactly is the content of the learning process that allows judges to benefit on future related judgments? To answer this question, it is helpful to distinguish between two different sources of estimation error. In their framework about real-world quantitative estimations, Brown and Siegler (1993, see also Brown, 2002) argue that people depend on two types of information when generating such estimates: *metric knowledge* and *mapping knowledge*. Metric knowledge is a general understanding of the appropriate scaling; it represents one's calibration of judgments. In other words, metric knowledge determines whether people have an accurate representation of the correct upper and lower boundaries of the quantity that they have to estimate, that is, the range of values that is plausible. For example, knowing that distances have a natural zero and that the length of the equator is approximately 40,000 km informs us about plausible numerical values when estimating airline distances between different cities. In contrast, mapping knowledge involves ordinal relations among individual estimations of the domain, that is, it allows us to put different target values of the same kind in the correct order. For example, one might know that London is closer to Rome than New York, without having a good guess about the actual distances.

Previous research on quantitative judgments in groups suggests that frames of reference play an important role when it comes to increases in estimation accuracy (e.g., Bonner & Baumann, 2008; Bonner, Sillito, & Baumann, 2007; Laughlin, Bonner, Miner, & Carnevale, 1999; Laughlin, Gonzalez, & Sommer, 2003). For example, the knowledge about the length of Germany from north to south (approx. 900 kilometers) and from east to west (approx. 600 kilometers) should improve accuracy when estimating distances within Germany, and should prevent completely implausible judgments. Obviously, these benchmarks can be illustrated easily when people are allowed to communicate, and can explain the reasons for a particular reference value. For example, there is evidence for *group-to-individual transfer* (G-I transfer) on quantitative estimation tasks similar to those frequently used in the JAS (e.g., Schultze, Mojzisch, & Schulz-Hardt, 2012; Stern, Schultze, & Schulz-Hardt, 2017). G-I transfer denotes an increase of individual capabilities as a consequence of prior collaborative task experience and is, hence, a type of social learning (e.g., Laughlin & Barth, 1981; Laughlin & Sweeney, 1977). However, a simple piece of advice without further explanations, as is usually the case in JAS experiments, might also serve as a frame of reference and lead to similar increases in judgment accuracy. Even though the judge may attribute systematic differences in metrics to the advisor being biased, a phenomenon also referred to as naïve realism (e.g., Pronin, Gilovich, & Ross, 2004), we think there is good reason to believe in increases in initial judgment accuracy after receiving advice. The judge might recognize differences to the received advice, particularly in the *metric error*, and especially when the differences are very pronounced (i.e., poor judges receiving accurate advice). When the judge's estimates are always markedly above the advice, a process of recalibration might be triggered that leads to lower subsequent judgments. For example, when the judge provides estimates of distances between European cities that are always above 20,000 km, and then receives advice that never exceeds 3,000 km, he or she might recognize the own bias and adjust the own judgments towards the advice. This, in turn, should lead to more accurate judgments.

In contrast, the process of reducing one's *mapping error* should be much more difficult. It should be very hard to detect differences between the own and the advisor's mapping error, because one needs multiple pieces of advice with a small mapping error, and also has to remember the specific advice values over a longer period. For example, when asked whether the distance between London and Rome is greater than the distance between London and Madrid, one single point of reference is hardly helpful, because it does not facilitate putting different distances in the right order. It is only when the judge has multiple accurate benchmarks for different distances and precisely remembers them that mapping error improvements might be possible. Hence, the process of learning mapping knowledge should be very time-consuming and, if at all, only possible with detailed explanations or additional information about the advisor's recommendation. Consequently, in the JAS, advice should predominantly increase the judge's metric knowledge, while leaving his or her mapping knowledge largely unaffected.

Quality of advice and its salience

As learners mainly exhibit modeled behavior when they perceive that this behavior has a high functional value, the quality of advice should be an important determinant for the strength of learning. Those behaviors that seem to be effective for others are favored over those behaviors that produced negative outcomes (Bandura, 1986). In other words, observers should be more willing to adopt the behavior of a high performing, successful model (i.e., an accurate advisor) and not imitate the behavior of the model when they perceive it as useless (i.e., poor advice). Adopting the metric of a good advisor is more beneficial than adopting the metric of a poor advisor, because high quality advice provides more accurate reference values. Consequently, good advisors should enable stronger improvements in judges' subsequent initial judgment accuracy than poor advisors, in particular when the judge recognizes the advice quality.

Even though the judge usually has no access to the advisor's reasoning in JAS research (Bonaccio & Dalal, 2006), previous research suggests that judges are, at least to some extent, sensitive to advice quality even in the absence of any kind of advisor performance feedback (Biele et al., 2009; Yaniv & Kleinberger, 2000). As Yaniv and Kleinberger (2000) discuss, judges could perform plausibility checks. Particularly poor estimates might be recognized as implausible, even if the judge cannot generate a correct estimate on his own. Nevertheless, the learning effect should be stronger with feedback about the advisor's accuracy, which is in line with Yaniv and Kleinberger (2000), who showed that judges were better at assessing the quality of advice when receiving feedback about it. In sum, based on social learning theory, there is good reason to believe that improvements of initial judgment accuracy will occur in judge advisor experiments, and that this happens particularly when the quality of advice is high and salient.

In sum, we aim to examine the occurrence of socially induced learning from one task to a different task of the same class after receiving advice, and we want to analyze its connection to advice quality, and to the salience of the advisors' accuracy. To demonstrate this transfer, we will use quantitative estimation tasks, which are suitable to account for gradual changes in judgment accuracy.

Hypotheses and overview of the present research

In three experiments, we investigate learning processes in a prototypical JAS, with the focus on socially induced individual learning and its relation to advisors' judgment accuracy, as well as to the judges' knowledge about this advice quality. In general, we expect that advice quality affects not just post-advice final judgments, as shown in many previous studies, but also the accuracy of subsequent initial (i.e., pre-advice) judgments. Hence, our first aim is to clarify whether receiving advice of different quality systematically changes the accuracy of judges' initial estimations on subsequent trials. Because of the higher accuracy of better advisors and, potentially, a stronger willingness to learn from superior advisors, we postulate:

Hypothesis 1: *Judges' initial judgment accuracy will improve dependent on advice quality. Judges' estimation accuracy will increase more strongly when receiving high quality advice compared to low quality advice.*

We further aim to show *what* judges learn as a consequence of receiving high quality advice. Building on the idea that reference values are crucial for increases in individual estimation accuracy, we postulate that there is a transfer of metric knowledge. High quality advice, in particular, should reduce the judge's metric error, because differences in metrics should be recognized quite easily, and high quality advice should provide reference values with a relatively small metric error. In contrast, we do not expect transfer of mapping knowledge since simple advice without any additional explanations or information should be insufficient for such a learning process, even though high quality advice should be characterized by a small mapping error as well.

Hypothesis 2a: *The metric error of judges' initial estimates after receiving advice differs as a function of advice quality. Judges' metric error will be reduced more strongly after receiving high quality advice compared to low quality advice.*

Hypothesis 2b: *There will be no changes in the mapping error of judges' initial estimates as a consequence of receiving advice, independent of its quality.*

As we have stated before, feedback about the advisors' estimation accuracy should influence the differentiation of advice quality and, thereby, the strength of learning. However, even in the absence of feedback, judges are supposed to be sensitive to advice quality or should, at least, benefit from better calibrated advisors, which should allow an increase in initial judgment accuracy as well. Hence, we hypothesize:

Hypothesis 3: *The accuracy of judges' initial estimates after receiving advice differs as a function of advice quality with or without feedback. However, the effect of advice quality is stronger if feedback about advice quality is given.*

For exploratory purposes, we also address the question of how much the hypothesized social learning processes contribute to the finding of improved post-advice final estimates that we already know from the literature (e.g., Soll & Larrick, 2009; Sniezek et al., 2004). In other words, we expect that improvements in the accuracy of final estimates after receiving advice partially stem from already more accurate initial judgments as long as the task type remains stable and advice is presented sequentially on every trial (e.g., Harvey & Fischer, 1997). Hence, we want to differentiate between two beneficial effects of receiving advice: (a) increases in initial judgment accuracy as a consequence of social learning and (b) simple advice weighting that should improve judges' final estimates and put them in relation to each other.

Experiment 1

In Experiment 1, we tested whether receiving advice improves subsequent initial estimates, and whether the advisor's accuracy moderates the extent of these learning effects (Hypothesis 1). Furthermore, we aimed to investigate whether learning gains manifested as improved metric or mapping knowledge (Hypothesis 2a & 2b). To address these questions, participants worked on quantitative estimation tasks and received randomly drawn advice from a pool of 76 advisors of different expertise who had taken part in a pretest. Participants either received advice from one randomly drawn advisor throughout all trials, or from a different randomly drawn advisor on each trial. Previous judge advisor research used both constant advisors (e.g., Gino & Schweitzer, 2008; Soll & Mannes, 2011; Minson & Müller, 2012) and varying advisors (e.g., Harvey & Fischer, 1997; Gino, 2008; Schultze, Rakotoarisoa, & Schulz-Hardt, 2015). We compared these two types of advice since they might have different impact on potential learning gains. On the one hand, receiving advice from the same advisor on all trials could facilitate learning, because judges can assess their advisor's systematic deviations from their own estimates. On the other hand, being stuck with one advisor means that the potential for learning is limited by that advisor's accuracy. If the advisor's accuracy is low, there might even be the risk of negative learning if the judge adopts the advisor's inferior metric or mapping. In the case of varying advisors, judges cannot infer systematic discrepancies between their own estimates and those of a specific advisor. Instead, they need to integrate information over several advisors, for example, by inferring a consensus among their advisors, allowing them to adjust their own metric or mapping to this central tendency. However, varying advisors could also foster learning because this variation should diminish naïve realism. If a judge has only one advisor, the judge may attribute systematic differences in metrics to the advisor being biased. In contrast, when confronted with recommendations from a group of people that differ from one's opinion while being similar to each other, it is more likely to attribute these discrepancies to oneself being subject to bias. Finally, we compared the two advice conditions with a control condition in which participants received no advice at all. This procedure allowed us to control for practice effects and, thus, to attribute stronger increases in initial accuracy in the two advice conditions unequivocally to social learning.

Method

Participants and design

One hundred and ninety-seven German or German-speaking students (133 women, 61 men, 3 participants did not report their gender), with an average age of 23.81 years (*SD* = 5.53), participated in the experiment. Eight participants were excluded from all analyses: Six of them had already participated in a preceding study and, thus, were familiar with the estimation task, and two other participants were excluded because their initial estimates were unreasonably high (they overestimated the true values by more than 2,300 percent), raising doubts about whether they had taken the task seriously. Experiment 1 is based on a one-factorial design with experimental condition (constant advisor, variable advisor, no advice) as a between-subjects factor.

Task and procedure

In each experimental session, up to 12 participants were invited. Upon arrival in the laboratory, they were guided to a room with several computer workstations. They were seated at separate computers and were informed about the task and the procedure of the experiment. Participants had to estimate airline distances between different European capital cities, a task where prior studies could successfully show social learning in interacting groups (Schultze et al., 2012; Stern et al., 2017). The estimations of participants from a previous pretest (*N* = 76) served as advice in our first experiment. Participants were randomly assigned to one of the three conditions. The experiment consisted of two phases: an individual practice phase with 10 distance estimates, and a subsequent test phase with 20 distance estimates. The individual practice phase was identical in all three conditions. Participants were given the opportunity to get used to the task without any advice, and the individual accuracy during this practice phase served as a performance baseline. The second phase differed between conditions. In the constant advisor condition, each participant received advice from one specific, randomly drawn participant of the pretest. In the variable advisor condition, participants also received advice from randomly determined participants of the pretest. However, in this condition a new advisor was randomly drawn for each trial. In both conditions, participants were fully informed about the drawing procedure at the beginning of the experiment, that is, they knew whether they were dealing with a constant advisor or with varying advisors. Finally, in the control condition participants received no advice at all, which allowed us to calculate an uninfluenced performance baseline, and to analyze whether participants' judgments change in the course of time even without advice.

In the two advice conditions, the second phase followed the classic JAS procedure, with the respondents taking the role of the judge. After giving an initial estimate, the participants received advice on the following screen, without additional information about advice quality or advisor competence. Judges then made a final, and possibly revised, estimate. In order to hold the number of estimates constant between conditions, participants in the control condition also estimated each distance twice. Instead of receiving advice, they were instructed to think about their initial estimate and then estimate the same target again. If they did not want to change their estimate, they were asked to simply enter their initial estimate a second time. The sequence of distances was randomized by the computer. Finally, we asked participants to report whether they received advice and, if they did, whether it came from the same advisor or varying advisors. These questions served as an awareness check for the manipulation of the experimental condition. Upon completing the experiment, participants received their payment of €5[1], were thanked for their participation, and debriefed.

Results and discussion

Manipulation check and check for possible interfering effects

Prior to the main analyses, we checked whether our advice manipulation was recognized as intended. In the no advice condition, all participants reported that they did not receive advice at the end of the experiment. In the constant advisor condition, 84.1 percent of participants reported that they had always received advice from the same advisor. Finally, in the variable advisor condition, 85.9 percent reported that they received advice from varying advisors. Accordingly, there was a significant association between the true and perceived experimental condition, χ^2 (4) = 277.74, $p < .001$. Hence, most of the participants understood whether they received advice from a constant or from varying advisors.

[1] Previous research on advice taking either rewarded their participants with a flat payment (e.g., Gino & Schweitzer, 2008), or with (additional) financial incentives for accurate estimates (e.g., Soll & Larrick, 2009). Usually, we reward our participants for accurate final estimates. In the current experiments, however, we refrained from doing so because financial incentives for accurate final estimates might have shifted participants' concentration exclusively to those judgments, possibly placing less effort on the initial estimates that were the focus of our analyses. Hence, we decided to work with a flat payment.

As we outlined before, advice quality should be a crucial moderator for the effects of social learning. Hence, we calculated some preliminary analyses on this important determinant. Firstly, we checked whether judges' initial accuracy was similar in all three conditions, to rule out baseline performance differences. To this end, we calculated the mean absolute percent error (MAPE) score as our main dependent variable. The MAPE is a common measure for estimation accuracy in quantitative judgment research (e.g., Sniezek & Henry, 1989, 1990). As the MAPE is a deviance score, lower MAPE values indicate a greater accuracy. An ANOVA with experimental condition (constant advisor vs. variable advisor vs. no advice) as between-subjects factor and participants' MAPE during the practice phase as dependent variable showed no significant differences in baseline performance, $F(2, 186) = 0.20$, $p = .815$, $\eta_p^2 < .01$. Additionally, we analyzed whether judges' initial estimates and the advice they received were, on average, equally accurate. A paired-sample *t*-test that compared the judges' MAPE during the practice phase to the MAPE of advice in both advice conditions revealed that advisors were significantly more accurate than the judges' baseline accuracy ($M = 43.18$, $SD = 16.46$ vs. $M = 62.92$, $SD = 56.29$), $t(126) = -3.69$, $p < .001$, $d = 0.33$. Furthermore, there was less variance among advisors than among judges. In total, the advisors were superior to the judges in 56 percent of all cases and outperformed them on average by 20 percentage points ($SD = 60.29$). However, there were no significant differences in the MAPE of advice between the constant and the variable advisor condition (M = 43.04, SD = 21.89 vs. M = 43.32, SD = 8.40), t(125) = -0.09, p = .925, d = 0.02.[2]

Judges' accuracy of initial estimates

Next, we investigated our main research question, namely whether socially induced individual learning occurs in a prototypical JAS, and how it is related to advisors' judgment accuracy. To this end, we were interested in how the initial estimate accuracy changed from the first individual practice phase to the second experimental phase. We treated the initial estimate of the first trial of the second phase (trial 11) as part of the individual practice phase, since participants provided this estimate prior to receiving any advice and, thus, before any socially induced learning effects could have occurred. We ran a 3 (*experimental condition:* constant advisor vs. variable advisor vs. no advice) × 2 (*trial block*: practice phase vs. experimental phase) repeated measures ANOVA with participants' initial accuracy during the two phases as dependent variable. This analysis revealed no main effect of experimental condition, F(2, 186) = 1.30, p = .274, η_p^2 = .01, but a significant main effect of trial block, F(2, 186) = 7.23, p = .008, η_p^2 = .04, that was qualified by an interaction of experimental condition and trial block, F(2, 186) = 5.06, p = .007, η_p^2 = .05. Additional simple effects analyses showed that there were no differences between the experimental conditions during the practice phase, F(2, 186) = 0.20, p = .815, η_p^2 < .01. In contrast, during the experimental phase there were significant differences between the experimental conditions, F(2, 186) =

[2] We also analyzed the strength of advice taking (AT), which is defined as (initial estimate – final estimate) / (initial estimate – advice), and is a common measure of advice weighting (e.g., Harvey & Fischer, 1997; Sniezek et al., 2004; Bonaccio & Dalal, 2006). We conducted a t-test to compare the AT in the constant advisor and the variable advisor condition. This analysis revealed no significant differences between the two conditions (M = 26.91, SD = 19.16 vs. M = 29.86, SD = 22.01), t(125) = -.81, p = .421, d = .14. In other words, judges weighted constant and variable advisors more or less equally. Furthermore, they only shifted between 26 and 30 percent towards the advice, which mirrors the usual amount of egocentric advice discounting reported in the literature (e.g., Bonaccio & Dalal, 2006). Experiments 2 and 3 replicate this general pattern of results with stronger advice taking of high quality advisors. However, we refrain from reporting further results regarding advice taking because it is not central to our research question.

4.23, $p = .016$, $\eta_p^2 = .04$ (see Table 1). Pairwise comparisons revealed that participants in both the constant and variable advice condition were significantly more accurate than participants who received no advice, $p = .020$, and, $p = .008$, respectively. The two advice conditions were not significantly different from each other, $p = .741$. Beyond that, the simple effects of trial block showed no accuracy changes for the no advice condition between blocks, $F(1, 186) = 0.75$, $p = .388$, $\eta_p^2 < .01$. In contrast, both in the constant and the variable advice condition participants' initial judgment accuracy increased after receiving advice, $F(1, 186) = 3.79$, $p = .053$, $\eta_p^2 = .02$, and, $F(1, 186) = 13.05$, $p < .001$, $\eta_p^2 = .07$, respectively, even though the former comparison did not reach conventional levels of significance. Hence, participants' average initial judgment accuracy only improved when they received advice. Figure 1 shows the distribution of gains and losses across participants and conditions. In the two advice conditions, participants' accuracy increased in 64 percent of the cases (81 of 127) and decreased in 36 percent of the cases (46 of 127). In the no advice condition, increases (34 of 62, or 55%) and decreases (28 of 62, or 45%) in participants' accuracy more or less balanced each other out.

Beyond that, we wanted to clarify the role of advice quality and judges' baseline accuracy on the strength of learning. To this end, we predicted judges' accuracy gains from the advisors' MAPE (for the variable advice condition we averaged the advisors' accuracy) and the judges' MAPE during the practice phase in a multiple regression. We computed accuracy gains as the difference between judges' MAPEs in the first and second phase of the experiment. Hence, positive values indicated a performance enhancement, and negative values a decrease in accuracy from the practice phase to the test phase. As the results of Experiment 1 revealed no significant differences between the constant and the variable advice condition at all, we collapsed across the two conditions. Therefore, we averaged the advisors' accuracy throughout the trials for the variable advice condition (the pattern of results remains unchanged when calculating two separate multiple regressions). The analysis revealed that the two predictors explained 79.3% of the variance, $R^2 = .79$, $F(2, 124) = 237.91$, $p < .001$. Judges' accuracy during the training phase significantly predicted subsequent changes in initial estimate accuracy ($\beta = .89$, $p <.001$), whereas advice quality did not ($\beta = -.02$, $p =.704$). In other words, judges with a low estimation accuracy benefitted the most from receiving advice. In sum, judges' initial accuracy increased after receiving advice, but instead of the quality of advice, only judges' own baseline accuracy moderated the strength of learning. Hence, the results of Experiment 1 do not support Hypothesis 1 to its full extent, but clearly show that receiving advice can enable social learning processes that increase the accuracy on subsequent initial estimates.

Changes in judges' metric and mapping error

In line with the assumption that advice offers a frame of reference that can reduce the individual metric error, and not the mapping error, we started by calculating the mean overall deviation (MOD) (Brown & Siegler, 1993). The MOD is a measure of metric property and represents the absolute discrepancy between, on the one hand, the subject's median estimate across all items and, on the other hand, the true overall median, with lower values indicating a more central estimation tendency. However, since the magnitude of participants' judgment errors covaried strongly with the respective true values, we worked with the percentage error instead of the absolute deviation from the true values.[3] We calculated a 3 (*experimental condition:* constant advisor vs. variable advisor vs. no advice) × 2 (*trial block*: practice phase vs. experimental phase) repeated measures ANOVA with participants' initial absolute median percentage error during the two phases as dependent variable. This analysis showed no main effect of experimental condition, $F(2, 186) = 1.11$, $p = .322$, $\eta_p^2 = .01$, but a significant main

[3] The pretest data revealed a strong correlation between the target values and the corresponding absolute errors ($r = .96$). However, the pattern of results remains the same when working with the median absolute error instead of the median absolute percentage error in all three experiments.

effect of trial block, $F(2, 186) = 7.49$, $p = .007$, $\eta_p^2 = .04$, that was qualified by an interaction of experimental condition and trial block, $F(2, 186) = 3.22$, $p = .042$, $\eta_p^2 = .03$. Simple effects analyses revealed that there were no differences between the three experimental conditions during the practice phase, $F(2, 186) = 0.03$, $p = .968$, $\eta_p^2 < .01$, whereas there were significant differences during the experimental phase, $F(2, 186) = 3.81$, $p = .024$, $\eta_p^2 = .04$ (see Table 1). Pairwise comparisons showed that participants in both the constant and variable advice condition had a significantly lower metric error than participants who received no advice during the experimental phase, $p = .029$, and, $p = .012$, respectively. In contrast, the two advice conditions were not different from each other, $p = .735$. Simple effects analyses comparing metric errors between the practice and the experimental phase revealed no changes in the control condition, $F(1, 186) = 0.21$, $p = .649$, $\eta_p^2 < .01$. However, in both the constant and the variable advice condition, metric errors decreased after receiving advice, $F(1, 186) = 5.09$, $p = .025$, $\eta_p^2 = .03$, and, $F(1, 186) = 8.80$, $p = .003$, $\eta_p^2 = .05$, respectively. Accordingly, receiving advice improved judges' metric error on subsequent initial estimates.

In contrast to metric errors, we did not expect systematic changes in mapping errors to occur. We operationalized mapping errors as rank-order correlations of participants' estimates and the true values (Brown & Siegler, 1993). In order to compare mapping errors between phases and conditions, we subjected them to a Fisher *z*-transformation. Similar to the metric errors, we calculated a 3 (*experimental condition:* constant advisor vs. variable advisor vs. no advice) × 2 (*trial block*: practice phase vs. experimental phase) repeated measures ANOVA with participants' Fisher *z*-transformed rank-order correlation coefficients during the two phases as dependent variable. The analysis revealed neither a main effect of experimental condition, $F(2, 186) = 0.12$, $p = .887$, $\eta_p^2 < .01$, nor main effect of trial block, $F(2, 186) = 0.14$, $p = .906$, $\eta_p^2 < .01$, nor interaction of experimental condition and trial block, $F(2, 186) = 0.48$,

p = .620, η_p^2 < .01 (see Table 1). In other words, there were no systematic changes in participants' mapping error.

In sum, receiving advice affected participants' metric but not their mapping error.[4] This supports the idea that accurate frames of reference have a beneficial effect on judges' initial estimates. However, it was the judge's baseline accuracy that mainly predicted metric error reductions after receiving advice, and not the quality of advice. Accordingly, the results of Experiment 1 are not fully in line with Hypothesis 2a. Furthermore, as there were no systematic changes in participants' mapping error, the results support Hypothesis 2b.

Exploratory analyses

Finally, to put the accuracy gains of initial judgments after receiving advice in perspective, we compared them to the accuracy gains resulting from advice taking when coming to the final judgment. To this end, we calculated whether participants' final estimates in the second phase (trials 12-30) were more accurate after receiving advice (averaged over both advice conditions) compared to the no advice control condition. Again, we only report one analysis for both advice conditions. However, the pattern of results remains unchanged when comparing the no advice condition with the constant advice or variable advice condition separately. Judges in the two advice conditions outperformed participants who did not receive advice by 27 percentage points (M = 42.83, SD = 23.38 vs. M = 70.25, SD = 80.86), t(187) = 3.54, p = .001, d = 0.46, thereby mirroring the finding of increased accuracy of judges' final decisions after receiving advice (e.g., Gardner & Berry, 1995; Gino & Schweitzer, 2008; Sniezek et al., 2004). Furthermore, we ran the same analysis with

[4] We found a significant relationship between advice quality and the advisor's metric error, r(127) = .71, p < .001, as well as advice quality and advisor's mapping error, r(127) = -.19, p = .029, indicating that high quality advice was characterized by a small metric and mapping error. Hence, the finding that superior advice only reduces judges' metric error cannot stem from systematic differences between the metric and mapping error of high quality advice. Furthermore, participants' Fisher z-transformed rank-order correlation coefficients suggest that there was still space for reductions in their mapping error (r = .68). Hence, we can rule out that ceiling effects might have prevented changes in mapping error from occurring.

participants' initial judgments in the second phase, to see how much of this performance advantage was due to transfer from one task to another. This *t*-test showed that judges' initial estimates were more accurate with than without advice by about 23 percentage points (*M* = 47.45, *SD* = 25.77 vs. *M* = 70.25, *SD* = 80.86), *t*(187) = 2.90, *p* = .004, *d* = 0.38. The accuracy gains due to adjusting the initial estimates towards the advice accounted for the remaining 4 percentage points. In other words, social learning accounted for 83 percent of the total beneficial effect of receiving advice, whereas advice weighting (i.e., the integration of advice into one's already improved initial judgments) only accounted for 17 percent.

Conclusions

In Experiment 1, we found evidence for socially induced learning in a prototypical JAS. Receiving advice improved judges' subsequent initial accuracy on estimation tasks, no matter whether it came from the same or varying advisors. The most likely explanation for this phenomenon is that judges adjusted their own metric towards that of the advisor(s), since advice affected participants' subsequent metric error, but not their mapping error. This adjustment in metrics accounted for the major part of the total beneficial effect of receiving advice. Contrary to our expectations, the magnitude of the improvements in initial judgment accuracy was not systematically related to the quality of advice. This might have to do with characteristics of the pretest participants who served as advisors in Experiment 1. As we have shown in the preliminary analyses, the random advice judges received was, in general, relatively accurate and had less variance than the judges' baseline accuracy. Hence, it might have been difficult to detect the moderating influence of advice quality.

Experiment 2

To substantiate the findings of social learning after receiving advice found in Experiment 1, it is crucial to manipulate the quality of advice. Therefore, we conducted a second experiment with some modifications, most importantly, an experimental manipulation leading to a more pronounced variation of advice quality, which is quite common in JAS research (e.g., Yaniv & Kleinberger, 2000). Participants received advice of high quality, moderate quality, or low quality, or they received no advice at all. Additionally, we wanted to be sure that effects of advice quality would be detected if they existed. Therefore, participants received veridical feedback about their corresponding advisor's accuracy. Beyond that, we worked with a different estimation task to improve the generalizability of our findings.

Method

Participants, design and task

One hundred and thirty-two German or German-speaking students (87 women, 45 men), with an average age of 23.59 (*SD* = 5.16) years, participated in the experiment. Experiment 2 used a one-factorial design with the experimental condition (high advice quality vs. moderate advice quality vs. low advice quality vs. no advice) as a between-subjects variable. Participants estimated the weight of different physical items (e.g., hammer, dustpan, umbrella) that were present in the room, without being allowed to touch or lift them.

Procedure

The procedure of Experiment 2 was similar to that of Experiment 1, with the following exceptions. First, we manipulated whether judges received advice from a good, moderate, or poor advisor in Experiment 2 and compared them to a control condition without advice. Again, we used participants of a pretest (*N* = 61) as advisors. We selected the participant with the best average performance, the participant whose performance marked the median of the sample, and the participant with the worst performance (see Yaniv & Kleinberger, 2000, for a similar manipulation of advice quality). The advisors' respective MAPE scores were 33 (high advice quality), 150 (moderate advice quality), and 523 (low advice quality). Beyond that, high quality advice was characterized by a smaller metric and mapping error than the medium quality advice and the low quality advice (metric error: 2.32

vs. 31.97 vs. 426.52; mapping error: .72 vs. .59 vs. .42). Second, participants received accurate feedback about their advisor's performance rank during the pretest (1st vs. 31st vs. 61st of 61). In the control condition, participants received no advice at all. Furthermore, we dropped the individual practice phase, because the control condition is sufficient to analyze changes in participants' judgment accuracy without advice. Hence, comparisons with the no advice control condition are sufficient to examine social learning processes after receiving advice. Consequently, in contrast to Experiment 1, the number of trials was reduced to 20, which made the experiment shorter overall. Due to the reduced duration of the experiment, participants only received a compensation of €4. After estimating all distances, participants were asked to rate the accuracy of their corresponding advisor to see whether our feedback manipulation was successful. To this end, participants estimated their advisors' MAPE.

Results and discussion

Manipulation checks and check for possible interfering effects

We analyzed whether participants drew the correct conclusions about the quality of the advice they received. To this end, we calculated an ANOVA with the three advice conditions (high quality vs. moderate quality vs. low quality) as between-subjects factor and the judges' estimation of the advisors' MAPE as a dependent variable. This analysis revealed significant differences between the advice conditions, $F(2, 96) = 8.79$, $p < .001$, $\eta_p^2 = .15$. Additional Tukey post hoc tests showed that the high quality advice was not rated significantly more accurate than the moderate quality advice ($M = 23.09$, $SD = 16.34$ vs. $M = 43.78$, $SD = 43.38$), $p = .751$, although the means were in the predicted direction. Beyond that, the high and moderate quality advice was rated significantly more accurate than the low quality advice ($M = 23.09$, $SD = 16.34$ vs. $M = 134.12$, $SD = 191.57$), $p < .001$, and ($M = 43.78$, $SD = 43.38$ vs. $M = 134.12$, $SD = 191.57$), $p = .006$, respectively. Hence, participants were able to distinguish the low quality of the poor advice, but did not perceive the good advisor to be significantly more accurate than the moderate advisor. As a consequence, we expect that the moderating effect of advice quality will be restricted to differences between low quality advice, on the one hand, and moderate and high quality advice, on the other hand.

Judges' accuracy of initial estimates

We first conducted an ANOVA with experimental condition (high advice quality vs. moderate advice quality vs. low advice quality vs. no advice) as a between-subjects factor and the MAPE of initial estimates as a dependent variable. For this analysis, we eliminated the first trial from the calculations, because on that trial participants had not yet received any advice. This analysis is equivalent to analyzing initial accuracy in the second phase of Experiment 1. We found a significant effect of experimental condition, $F(3, 128) = 11.44$, $p < .001$, $\eta_p^2 = .21$. Tukey post hoc tests showed that the MAPE of initial estimates was not significantly different in the high and in the moderate advice quality condition, $p = .630$, although, descriptively, judges in the high quality advice condition were more accurate by about 35 percentage points (see Table 2). Initial estimates were more accurate in the high and moderate than in the low advice quality condition, $p < .001$, and, $p = .001$, respectively. Both, receiving advice from a highly accurate and from a moderately accurate source led to significantly more accurate initial estimates than receiving no advice at all, $p = .001$, and, $p = .040$, respectively. Finally, participants in the low advice quality condition were (descriptively) somewhat inferior to participants without advice, but this difference was not significant, $p = .588$. Figure 2 shows the accuracy distribution across participants and conditions, with a clear tendency in the high and moderate advice quality condition for more accurate estimates. In total, these results indicate that participants' subsequent initial estimation accuracy increases as a function of advice quality, supporting Hypothesis 1. However, the fact that there were no differences between the high and moderate advice quality condition supports the idea that judges can benefit from advice as long as it is not particularly poor.

Changes in judges' metric and mapping error

To test for possible differences in participants' metric errors, we conducted an ANOVA with experimental condition (high advice quality vs. moderate advice quality vs. low advice quality vs. no advice) as a between-subjects factor and the absolute median percentage error of initial estimates as a dependent variable. Again, we eliminated the first trial from these calculations. This analysis showed a significant effect of experimental condition, $F(3, 128) = 13.29$, $p < .001$, $\eta_p^2 = .24$. Tukey post hoc tests revealed that judges' metric errors were not significantly different in the high and in the moderate advice quality condition, $p = .713$ (see Table 2). However, metric errors were significantly lower in the high and moderate than in the low advice quality condition, $p < .001$, and, $p < .001$, respectively. Judges' metric errors in the high and moderate advice quality condition were also lower than in the no advice condition, $p < .001$, and, $p = .018$, respectively. Finally, there were no significant differences between the low advice quality and the no advice condition, $p = .468$. In sum, these analyses show that participants had a lower metric error after they received high or moderate quality advice compared to the low advice quality and the no advice condition, thereby mirroring the corresponding findings with regard to overall judgment accuracy.

We tested for possible differences in participants' mapping error in an ANOVA with experimental condition (high advice quality vs. moderate advice quality vs. low advice quality vs. no advice) as a between-subjects and the participants' Fisher *z*-transformed rank-order correlations (trials 2 to 20) as a dependent variable. This analysis revealed no significant differences between the conditions, $F(3, 128) = 0.25$, $p = .864$, $\eta_p^2 = .06$ (see Table 2). In sum, we found evidence for differences in judges' metric errors as a function of advice quality with no differences in judges' mapping error, thereby supporting Hypothesis 2a and 2b.

Exploratory analyses

Finally, in line with Experiment 1, we put the accuracy gains of initial judgments after receiving advice in perspective and compared them to the accuracy gains in participants' final judgments, resulting from advice taking. To this end, we calculated an ANOVA with experimental condition (high advice quality vs. moderate advice quality vs. low advice quality vs. no advice) as a between-subjects factor and the MAPE of final estimates after the first advice (trials 2-20) as a dependent variable. This analysis revealed a significant effect of the experimental condition, $F(3, 128) = 16.70$, $p < .001$, $\eta_p{}^2 = .28$. Tukey post hoc tests showed that the MAPE of final estimates was significantly more accurate when receiving high quality advice or moderate quality advice as compared to no advice ($M = 62.14$, $SD = 45.16$ vs. $M = 201.11$, $SD = 174.94$), $p < .001$, and ($M = 114.64$, $SD = 47.47$ vs. $M = 201.11$, $SD = 174.94$), $p = .020$, respectively. In the low advice quality condition, the final estimate accuracy was not significantly different from the control condition ($M = 248.30$, $SD = 143.38$ vs. $M = 201.11$, $SD = 174.94$), $p = .365$. Hence, participants who received high and medium quality advice outperformed those without advice by 139 and 86 percentage point, whereas low quality advice led to a statistically insignificant inferiority of 47 percentage points.

To see what part of the benefit of advice stemmed from social learning, we compared these results to the results of the same analysis with the initial judgment accuracy as dependent variable (see section *Judges' accuracy of initial estimates*). This analysis revealed that judges who received high quality advice outperformed no advice control participants by 114 percentage points, and judges who received moderate quality advice outperformed the control individuals by 79 percentage points. In contrast, participants in the low advice quality condition were somewhat (but not significantly) inferior to those in the control condition (namely by 37 percentage points). Hence, in the high advice quality condition social learning accounted for 82 percent and in the moderate advice quality condition for even 92 percent of the total beneficial effect of receiving advice. Accordingly, these results indicate that the major part of differences between participants who received high or moderate quality advice and those who didn't receive advice already manifested in their initial judgments (i.e., due to social learning processes), with minor subsequent changes as a consequence of advice weighting, thereby mirroring the findings of Experiment 1.

Conclusion

Summarizing Experiment 2, we replicated the findings of the first experiment that advice can affect the accuracy of judges' subsequent initial estimates. Judges benefitted from high and moderate quality advice, which improved their initial judgment accuracy. The increase in estimation accuracy seems to be independent of whether participants have the opportunity to get used to the task during an individual practice phase, which further substantiates our findings and speaks to the generalizability of the phenomenon. Surprisingly, moderate quality and high quality advice had about the same effect. The differences between these two conditions were, albeit in the predicted direction, weak and statistically insignificant. Hence, as long as it is sufficiently reasonable, advice seems to have a positive effect on the judges' initial estimate accuracy. However, one has to take into account that in our advice quality manipulation the high and moderate quality advice was much more similar than, for example, the moderate and the low quality advice. This, in turn, could also explain the insignificant difference when it comes to improvements in subsequent initial judgments between participants that received high compared to moderate quality advice. Low quality advice, in contrast, did not improve these judgments, but neither did it significantly harm judges initial estimate accuracy, supporting the idea that judges seem to be rather sensitive to the quality of advice, which prevents them from adjusting their subsequent initial judgments towards the poor advice. Beyond that, high and moderate quality advice reduced participants' metric error. Furthermore, the results support the idea of two beneficial effects of high quality advice, namely social learning processes that strongly improve the initial judgment accuracy and, in addition to it, adjustments towards the advice that further improve the accuracy of final judgments. Nevertheless, the major increase in accuracy seems to derive from the transfer from one task to another.

Experiment 3

Experiment 2 provided evidence that the advice quality moderates improvements in subsequent initial judgment accuracy. However, participants received veridical feedback about the quality of advice and, hence, were in a somewhat "ideal" situation of perfect advice quality salience. In real-world advice situations, this degree of salience is rather uncommon. Accordingly, in Experiment 3, we wanted to find out whether we could still find a moderating effect of advice quality without such feedback, and how strong it would be in comparison to a situation with feedback. For this reason, we implemented some differences compared to Experiment 2. First, we worked with the same task as in Experiment 1, in order to substantiate the general findings of Experiment 2 with a different estimation task. Second, as we found no significant differences in individual learning between judges who received high quality compared to moderate quality advice in Experiment 2, we dropped the latter condition in Experiment 3. Judges received advice from either a very accurate or an inaccurate advisor. Moreover, we were interested in how feedback about the advisors' estimation accuracy influences the strength of learning. In everyday life, people often get no reliable information about their advisor's competence, which raises the question of how judges react to the different quality of advice in the absence of this meta-knowledge. To address this issue, we compared judges without such information to judges who received veridical feedback about whether they received good or poor advice.

Method

Participants and design

One hundred and sixty-four German or German-speaking students (104 women, 59 men, one participant did not report the gender), with an average age of 23.98 (*SD* = 4.23) years, participated in the experiment. Four participants were excluded because of unreasonably high estimates either during the practice phase or during the last trials of the experimental phase (overestimating the true values by more than 1,300 percent), indicating that these persons had not taken the experimental task seriously. Experiment 3 is based on a 2 (advice quality: high vs. low) × 2 (feedback about advice quality: yes vs. no) factorial design.

Task and procedure

In general, the task and basic procedure was the same as in Experiments 1 and 2. Therefore, we focus on reporting the changes made compared to the previous experiments. First, we manipulated whether judges received advice from a good or poor advisor in Experiment 3. In the high advice quality condition, subjects received advice from the most capable of 76 participants of a pretest (the same pretest that we referred to in Experiment 1) with an average MAPE-score of 22. In the low advice quality condition, the advisor was the least accurate pretest participant, with a MAPE-score of 146. Furthermore, high quality advice was also characterized by a smaller metric and mapping error than the low quality advice (metric error: 11.98 vs. 121.60; mapping error: .72 vs. .50). Second, half of the participants received accurate feedback about their advisor's performance rank during the pretest (1st vs. 76th), whereas the other half solely received the advice, without any additional information about its quality. Finally, we re-established the individual practice phase, comparing judges' accuracy in the practice phase to that of the initial estimates in the test phase. Since Experiment 1 showed that there were no accuracy gains in the absence of advice, we can attribute all changes between the two phases as effects of receiving advice.

This allowed us to drop the no advice control condition, resulting in a straightforward two-factorial design (instead of a design featuring a non-factorial control group). As in Experiment 1, participants received a compensation of €5.

Results and discussion

Manipulation checks and check for possible interfering effects

We first checked whether there were systematic differences in judges' baseline accuracy between the conditions. To this end, we calculated an ANOVA with advice quality (high vs. low) and feedback (yes vs. no) as between-subjects factors and participants' MAPE-scores during the practice phase as a dependent variable. This analysis revealed no significant main or interaction effect, all *F*s < 0.68, all *p*s > .413.

To analyze whether participants were able to assess the quality of advice, we calculated an ANOVA with advice quality (high vs. low) and feedback (yes vs. no) as between-subjects factors, and the judges' estimation of the advisors' MAPE as dependent variable. This analysis revealed a significant main effect of advice quality, $F(1, 156) = 68.14$, $p < .001$, $\eta_p^2 = .30$, that was qualified by an interaction of advice quality and feedback, $F(1, 156) = 16.43$, $p < .001$, $\eta_p^2 = .10$. The main effect of feedback was not significant, $F(1, 156) = 1.01$, $p = .318$, $\eta_p^2 = .01$. Simple effects analyses revealed that high quality advice was rated as more accurate with than without feedback ($M = 22.40$, $SD = 18.12$ vs. $M = 37.02$, $SD = 15.15$), $F(1, 156) = 12.94$, $p < .001$, $\eta_p^2 = .08$, whereas the low quality advice was rated as less accurate if feedback was given ($M = 58.00$, $SD = 21.44$ vs. $M = 49.18$, $SD = 18.07$), $F(1, 156) = 4.60$, $p = .034$, $\eta_p^2 = .03$. Consequently, our feedback manipulation was successful and led to a higher salience of advice quality. As expected, participants rated high quality advice as more accurate than low quality advice when they received veridical feedback, $F(1, 156) = 74.81$, $p < .001$, $\eta_p^2 = .32$. However, even without such feedback, the high quality advice was rated as being more accurate than the low quality advice, $F(1, 156) = 8.94$, $p = .003$, $\eta_p^2 =$

.05, indicating that the judges were still sensitive to the quality of advice even when they received no information about the advisor's accuracy.

Judges' accuracy of initial estimates

First, we conducted a 2 (*advice quality:* high vs. low) × 2 (*feedback*: yes vs. no) × 2 (*trial block*: practice phase vs. experimental phase) repeated measures ANOVA with participants' initial accuracy as the dependent variable. This analysis showed a main effect of advice quality, $F(1, 156) = 4.76$, $p = .031$, $\eta_p^2 = .02$, and a main effect of trial block, $F(1, 156) = 6.22$, $p = .014$, $\eta_p^2 = .04$. Both were qualified by an interaction of advice quality and trial block, $F(1, 156) = 14.08$, $p < .001$, $\eta_p^2 = .08$. Furthermore, there was no main effect of feedback nor interactions with feedback, all *F*s < 1.59; all *p*s > .210.[5] Simple effects analyses revealed that there were no differences between the advice quality conditions during the practice phase, $F(1, 156) = 0.02$, $p = .880$, $\eta_p^2 < .01$, whereas participants receiving high quality advice were significantly more accurate than those receiving low quality advice during the experimental phase, $F(1, 156) = 30.46$, $p < .001$, $\eta_p^2 = .16$ (see Table 3). Furthermore, participants receiving high quality advice significantly improved in accuracy from the practice to the experimental phase, $F(1, 156) = 19.76$, $p < .001$, $\eta_p^2 = .11$, whereas participants who received low quality advice showed no significant changes in accuracy, $F(1, 156) = 0.78$, $p = .378$, $\eta_p^2 < .01$. Figure 3 shows the distribution of gains and losses across participants and conditions. In the high advice quality conditions 84 percent of participants

[5] In addition, we tested whether the interaction of advice quality and trial block was significant even without feedback (i.e., to check whether advice quality affects accuracy changes even if the judge is not informed about the quality of the advice). Separate repeated measures ANOVAs for the two feedback conditions revealed that judges' initial accuracy improved more strongly after receiving high compared to low quality advice, both with feedback, $F(1, 77) = 5.82$, $p = .018$, $\eta_p^2 = .07$, and without feedback, $F(1, 79) = 12.33$, $p = .001$, $\eta_p^2 = .13$. To analyze whether the increase in accuracy was stronger with feedback, we also calculated separate repeated measures ANOVAs for the two advice quality conditions. However, the interaction of feedback and trial block was insignificant, both with high quality advice, $F(1, 79) = 1.03$, $p = .314$, $\eta_p^2 = .01$, and with low quality advice, $F(1, 77) = 0.76$, $p = .387$, $\eta_p^2 = .01$. Accordingly, even when receiving high quality advice, feedback by itself had no significant additional effect on the strength of learning.

(68 of 81) showed increases and 16 percent (13 of 81) decreases in their accuracy. In contrast, in the low advice conditions the number of participants that showed increases (42 of 79, or 53%) and decreases (37 of 79, or 47%) in their accuracy were more or less equal. In sum, judges' initial estimation accuracy improved after receiving good advice, with the result that they outperformed judges receiving poor advice, in line with Hypothesis 1. High quality advice led to higher initial judgment accuracy even without feedback, in line with the first part of Hypothesis 3. However, in contrast to the second part of Hypothesis 3, this performance enhancement was not significantly stronger when participants received feedback about the quality of advice, as we found no interaction with feedback whatsoever.

Changes in judges' metric and mapping error

Similar to Experiment 1, we analyzed changes in judges' metric error. To this end, we conducted a 2 (*advice quality:* high vs. low) × 2 (*feedback*: yes vs. no) × 2 (*trial block*: practice phase vs. experimental phase) repeated measures ANOVA with participants' absolute median percentage error of initial estimates as the dependent variable. This analysis revealed main effects of advice quality, $F(1, 156) = 5.89$, $p = .016$, $\eta_p^2 = .04$, and trial block, $F(1, 156) = 7.41$, $p = .007$, $\eta_p^2 = .05$, that were qualified by an interaction of advice quality and trial block, $F(1, 156) = 11.34$, $p < .001$, $\eta_p^2 < .07$. Again, the main effect of feedback as well as all interactions with feedback were insignificant, all Fs < 2.56; all ps > .112. Simple effects analyses showed no differences between the advice quality conditions during the practice phase, $F(1, 156) < 0.01$, $p = .943$, $\eta_p^2 < .01$. However, during the experimental phase, high quality advice led to a significantly lower metric error, $F(1, 156) = 37.16$, $p < .001$, $\eta_p^2 = .19$ (see Table 3). Beyond that, receiving high quality advice significantly reduced participants' metric error, $F(1, 156) = 18.78$, $p < .001$, $\eta_p^2 = .11$, whereas there were no changes in participants' metric error when they received low quality advice, $F(1, 156) = 0.21$, $p = .651$, $\eta_p^2 < .01$.

Furthermore, we also tested for possible changes in participants' mapping errors. Therefore, we calculated a 2 (*advice quality:* high vs. low) × 2 (*feedback*: yes vs. no) × 2 (*trial block*: practice phase vs. experimental phase) repeated measures ANOVA with participants' Fisher *z*-transformed rank-order correlations of initial estimates as the dependent variable. This analysis revealed neither main effects of advice quality, feedback or trial block, nor interaction effects of either advice quality or feedback and trial block, all *F*s < 2.34; all *p*s > .128 (see Table 3).[6] In summary, when judges received high quality advice, they were able to reduce their metric error, whereas there were no significant changes in judges' metric error when they received low quality advice. Beyond that, there were no systematic changes in judges' mapping errors at all. These results are in line with Hypothesis 2a and 2b.

Exploratory analyses

Similar to Experiment 1 and 2, we compared the accuracy gains of initial judgments after receiving advice to the accuracy gains resulting from advice taking when coming to the final judgment. Consequently, we calculated whether participants' final estimates in the second phase (trials 12-30) were more accurate after receiving advice compared to the uninfluenced training phase (because we had no control condition in Experiment 3, we worked with the training phase as a performance baseline). The corresponding paired-sample *t*-test revealed that the post advice final judgments were significantly more accurate than the training phase estimates when receiving high quality advice (M = 26.09, SD = 14.70 vs. M = 62.02, SD = 75.77), $t(80) = -4.39$, $p < .001$, $d = 0.49$, but not when receiving low quality advice (M = 65.76, SD = 44.35 vs. M = 60.54, SD = 49.02), $t(78) = 1.53$, $p = .131$, $d = 0.17$.

[6] The analyses showed a non-significant interaction of advice quality, feedback and trial block, $F(1, 156) = 3.80$, $p = .053$, $\eta_p^2 = .02$. However, separate paired sample *t*-tests for each condition only revealed significantly decreasing mapping errors when receiving advice of low quality with feedback, $t(38) = 2.87$, $p = .007$, $d = 0.46$, with no changes in the other three condition pairs, all *t*s < 1.02, all *p*s > .318. In our opinion, there is no theoretical basis as to why judges who receive particularly poor advice should increase their mapping knowledge, whereas judges receiving high quality advice should not. Hence, we think that this finding is a coincidence rather than a systematic interaction of advice quality, feedback, and trial block.

Furthermore, we compared participants' initial judgment accuracy in the second phase with the training phase to see how much of the accuracy gain can be attributed to the transfer from one task to another. The paired sample *t*-test showed that the initial judgments were significantly more accurate after receiving high quality advice than during the training phase (*M* = 34.31, *SD* = 17.91 vs. *M* = 62.02, *SD* = 75.77), *t*(80) = -3.40, *p* = .001, *d* = 0.39, whereas, again, there were no significant differences when receiving low quality advice (*M* = 66.17 *SD* = 48.44 vs. *M* = 60.54, *SD* = 49.02), *t*(78) = 1.67, *p* = .098, *d* = 0.19. Hence, receiving high quality advice increased subsequent initial judgments by almost 28 percentage points and post-advice final judgments by another 8 percentage points. Accordingly, social learning accounted for 77 percent of the total beneficial effect of receiving high quality advice, whereas adjustments toward the advice accounted for 23 percent.

Conclusions

The findings of Experiment 3 indicate that good advice leads to stronger improvements in subsequent initial judgments than poor advice, even when the advice quality is not made salient. High quality advice without feedback led to higher initial judgment accuracy and smaller metric errors, without significant additional performance enhancement when feedback about the advice quality was given. Interestingly, low quality advice did not lead to significant negative learning with or without feedback, which supports the idea that judges can identify low quality advice even without such feedback (Yaniv & Kleinberger, 2000). But why did feedback not have an additional beneficial effect in the high quality advice condition? This seems odd given that participants rated the accuracy of the high-quality advice as more accurate when receiving feedback about his or her competency. One explanation is that judges adjust their own metrics towards that of their advisor as long as they perceive the advice as sufficiently accurate, without substantial differences in the strength of this adjustment. This would be in line with the findings of Experiment 2. Finally, when receiving high quality advice, judges' post-advice final judgments were superior to the improved initial judgments. However, the major performance enhancement seems to derive from social learning and not advice taking, thereby mirroring the general findings of Experiment 1 and 2.

General discussion

In the present study, we tested for general social learning processes as a consequence of receiving advice, in terms of a transfer from one task to another. More precisely, we were interested in whether advice would affect not just the accuracy of post-advice final judgments, as shown in many previous studies (e.g., Gardner & Berry, 1995; Gino & Schweitzer, 2008; Sniezek et al., 2004), but also the accuracy of the judge's subsequent *initial* judgments. In particular, we expected individual performance enhancements especially when receiving advice of high quality, because of the superior accuracy of better advisors. We expected high quality advice to provide an accurate point of reference that should mainly reduce judges' metric error with no systematic changes in their mapping knowledge. Beyond that, we postulated that feedback about the advisors' accuracy should influence the judges' ability to infer advice quality and, thereby, the strength of social learning. However, even in the absence of feedback, we expected judges receiving high quality advice to outperform judges receiving low quality advice, because judges should be somewhat sensitive to advice quality or because they should, at least, benefit from the superior calibration of their advisors. In an exploratory manner, we also differentiated between the performance enhancements of initial judgments and the beneficial effect of combining the own initial estimate with the advice when coming to a final judgment. To this end, we compared the accuracy of post-advice final judgments to the, assumedly improved, initial judgments.

In line with our hypotheses, we found evidence that moderate or high quality advice led to social learning that manifested as improved accuracy of subsequent initial judgments in two different estimation tasks. In contrast, if at all, poor advice only seemed to mildly harm judges' subsequent initial judgments. In other words, the beneficial effect of receiving high quality advice markedly exceeded the possible detrimental effect of receiving low quality advice. Our results provide first evidence for social learning in terms of a generalized transfer in the JAS. As our results further indicate, high quality advice reduced judges' individual metric errors, whereas no systematic changes occurred in their mapping knowledge. Surprisingly, feedback about the quality of advice had no significant additional positive impact on the strength of learning. Even without feedback, participants were sensitive to advice quality. Participants receiving high quality advice who received no feedback benefited equally to those who did. In the case of low quality advice, participants' accuracy did not decrease substantially, and this was equally true in the absence and in the presence of feedback about the advisor's accuracy. Furthermore, judges' final estimates were still more accurate than their initial judgments when receiving recommendations from a good advisor, which supports the idea of two distinct beneficial effects of receiving high quality advice. On the one hand, social learning processes lead to improved subsequent initial judgments and, on the other hand, advice weighting (i.e., integrating the advice into one's final judgment) improves the accuracy of final post-advice judgments – with the former process being more pronounced than the latter. In the following section, we discuss these results against the backdrop of previous research, point out limitations of our experiments, and illustrate directions for future research.

Learning from advice

Previous research already dealt with the question of whether advice enables some kind of social learning processes, with the robust result that advice helps to find the right solution more quickly when repeatedly working on one specific task in prototypical decision making experiments (e.g., Biele et al., 2009; Çelen et al., 2010; Chaudhuri et al., 2006;

Kocher et al., 2014). For example, Biele et al. (2009) found that a single piece of advice can improve the performance on a repeated choice task, such that the decision maker identifies the recommended correct option more quickly and, consequently, chooses this option more often over the course of the experiment. In other words, people seem to learn from specific advice, with the result of improved post advice decisions on the same task – which is also mirrored by classic judge advisor experiments showing that advice leads to improved final judgments (e.g., Soll & Larrick, 2009; Sniezek et al., 2004). Hence, one can conclude that there is ample evidence for *specific* social learning that improves the quality of post advice final judgments or decisions. Beyond this positive effect of receiving advice, we found strong evidence that advice can also have a more general beneficial effect on subsequent related tasks. In other words, advice does not only contain information that can increase one's performance on the same task, but can also initiate a transfer to different tasks from the same domain. Altogether, these findings suggest that *general* social learning should be added to the list of positive effects of receiving advice, and should be addressed more often in future research.

Furthermore, our study focused on the moderating effect of advice quality on the strength of learning, with the common result of stronger learning after receiving high quality advice. Judges seem to understand the quality of advice in particular when receiving low quality advice. Accordingly, in Experiment 2, we found differences in the strength of learning between the high and low advice quality conditions as well as the moderate and low advice quality conditions, whereas there were no significant differences between the high and moderate advice quality conditions where the advice was less divergent. Beyond that, in Experiment 3, feedback had no significant additional effect on the strength of performance changes. Hence, to a certain extent, judges are sensitive to the quality of advice even without any external cues, which is in line with previous research (Biele et al., 2009; Yaniv &

Kleinberger, 2000). As Yaniv & Kleinberger discuss, judges might recognize particularly poor estimates as out of bounds, even though the judge cannot generate a correct estimate on his or her own. For example, it should be rather difficult to determine whether the distance between London and Rome is rather 1,500 or 2,000 km. However, when the advice suggests that this distance is 30,000 km, the judge has a very good chance to understand its low quality and, hence, refrain from adjusting towards this poor point of reference, even without explicit feedback about the advice quality. However, in the study of Yaniv and Kleinberger, the judges had another cue for the advice quality. Along with the advice, judges received the advisors' lower and upper boundaries in which the true value should be included with a probability of 0.95. This range might have been used as a predictor for accuracy in such a way that closer intervals, indicating higher advisor confidence, could hint at the high quality of the advice. The fact that such an indirect cue was missing in our experiments makes it even more astonishing that our participants were able to somehow recognize the advice quality.

As our results further indicate, the individual performance enhancements mainly derive from diminished individual metric errors after receiving accurate advice, no matter whether its quality is made salient or not. As we know from previous research, frames of reference play an important role when it comes to increases in estimation accuracy (e.g., Bonner & Baumann, 2008; Bonner et al., 2007; Laughlin et al., 1999; Laughlin et al., 2003). One might have a correct representation of Europe or of weight proportions between different items, while lacking an adequate benchmark for accurate judgments. Even without fully trusting the advisor's recommendation, this information can be obtained by high quality advice, and judges might improve their metric knowledge by receiving such well-calibrated points of reference. In our opinion, judges recognize systematic differences between their own and the advisor's metric, and they seek to reconcile this discrepancy by adjusting their metric towards that of the advisor. However, this only seems to be true in cases where the judge perceives the advice as being of high functional value.

In general, our findings suggest two sources of using advice: (a) social learning processes that improve the initial judgment accuracy, in particular, when receiving high quality advice, and (b) advice weighting in terms of an adjustment towards the specific advice when coming to a final estimate. One interesting implication of the former, so far overlooked, learning process is that the true amount of advice use might have been underestimated in previous JAS research, at least when the task type remains stable and advice is presented sequentially on every trial (e.g., Harvey & Fischer, 1997). Our results imply that the major beneficial effect of receiving advice derives from social learning, with an additional minor but still significant effect of simple advice weighting. Hence, future research should always take into account these performance enhancements as a consequence of receiving advice to better assess the full benefit of receiving advice.

Limitations and directions for future research

There are some limitations of our current study that should be taken into account. Firstly, in Experiment 2 and 3, we employed a rather strong manipulation of advice quality. Hence, it is debatable how accurately people can judge their advisor's competency when it is less evident, and how this affects the learning processes that we investigated. On the one hand, judges might also negatively learn from low quality advice as long as this advice contains a certain degree of plausibility. On the other hand, poor but plausible advice could even have a beneficial effect as long as the advisor's and judge's metric errors are on opposite sides of the target value, with the result that adjusting one's judgment towards the advice leads to more accurate estimates.

Secondly, we only used two different types of estimation tasks. Although we found structurally similar patterns with both types, our results should be replicated with additional types of tasks. For example, a more complex task might, on the one hand, affect the amount of time needed until the learning process is completed, or might even eliminate the performance enhancements in initial estimates, because the exchange of well calibrated numeric information might not be sufficient to induce a general social learning process. On the other hand, on very difficult tasks, even negative learning might occur. When the task specific knowledge of the judge is low, it should be more difficult to assess the quality of advice. Consequently, even recommendations of weak advisors might be taken into account more strongly, which, in turn, should lead to a loss in estimation accuracy.

Thirdly, the fact that we found no evidence for changes in participants' mapping error might have to do with the JAS as an experimental paradigm. When people receive advice without any additional explanation, as is usually the case in the JAS, we think it is highly reasonable that judges only reduce their metric error, because reducing one's mapping error should be more complex and should require additional information. However, in a real-world advice situation with communication, the advisor can explain his or her advice, and even correct the misconceptions of a judge. Hence, in our opinion there is good reason to believe that advice has the power to also reduce mapping errors, but only in a more interactive situation than provided by the classic JAS.

Finally, we currently cannot say whether the learning process that we demonstrated in our experiments is stable over time. To be sure about this stability, future studies could provide the judges with only one or two advice trials and analyze how their judgment accuracy develops on subsequent trials without advice. In the somewhat related field of group-to-individual (G-I) transfer in group judgment research, there is evidence for stable socially-induced improvements of individual accuracy after the group is dissolved (Stern et al., 2017). Because of the high structural resemblance in the utilized tasks and experimental procedures between these group experiments and our current judge-advisor study, there is good reason to expect that the learning process that we found in the present study will also turn out to be rather stable over time.

Conclusion

In three experiments, we have shown that advice of moderate to high quality affects accuracy on subsequent initial judgments. More precisely, judges' individual metric errors decreased, most likely as a consequence of adjusting their own opinion towards the reference values provided by well-calibrated advisors. Furthermore, feedback about the advice quality had no additional beneficial effect. Even without advice quality salience, the accuracy of judges' subsequent initial estimates increased after receiving high quality advice, and these increases in initial judgment accuracy accounted for the lion's share of the beneficial effects of advice on final judgment accuracy that have been shown in numerous previous studies. Hence, general social learning in terms of a transfer from one task to another in the same domain should be added to the beneficial effects of advice.

References

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational models for the combination of advice and individual learning. *Cognitive Science, 33,* 206-242. http://dx.doi.org/10.1111/j.1551-6709.2009.01010.x

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*, 127-151. http://dx.doi.org/10.1016/j.obhdp.2006.07.001

Bonaccio, S., & Van Swol, L. (2014). Combining information and judgments. In S. Highhouse, R. S. Dalal & E. Salas (Eds.), *Judgment and decision making at work* (pp. 178-198). New York, NY: Routledge.

Bonner, B. L., & Baumann, M. R. (2008). Informational intra-group influence: the effects of time pressure and group size. *European Journal of Social Psychology, 38*, 46-66. http://dx.doi.org/10.1002/ejsp.400

Bonner, B. L., Sillito, S. D., & Baumann, M. R. (2007). Collective estimation: Accuracy, expertise, and extroversion as sources of intra-group influence. *Organizational Behavior and Human Decision Processes, 103,* 121-133. http://dx.doi.org/10.1016/j.obhdp.2006.05.001

Braaksma, M. A., Rijlaarsdam, G., & Van den Bergh, H. (2002). Observational learning and the effects of model-observer similarity. *Journal of Educational Psychology*, *94*, 405-415. http://dx.doi.org/10.1037/0022-0663.94.2.405

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100*, 511-534. http://dx.doi.org/10.1037/0033-295X.100.3.511

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 41, pp. 321–360). New York: Academic Press. http://dx.doi.org/10.1016/S0079-7421(02)80011-1

Çelen, B., Kariv, S., & Schotter, A. (2010). An experimental test of advice and social learning. *Management Science, 56,* 1687-1701. http://dx.doi.org/10.1287/mnsc.1100.1228

Chaudhuri, A., Graziano, S., & Maitra, P. (2006). Social learning and norms in a public goods experiment with inter-generational advice. *The Review of Economic Studies*, *73,* 357-380. https://doi.org/10.1111/j.1467-937X.2006.0379.x

Couzijn, M. (1999). Learning to write by observation of writing and reading processes: Effects on learning and transfer. *Learning and Instruction*, *9,* 109-142. http://dx.doi.org/10.1016/S0959-4752(98)00040-1

Farrell, S. (2011). Social influence benefits the wisdom of individuals in the crowd. *Proceedings of the National Academy of Sciences*, *108*, E625. http://dx.doi.org/10.1073/pnas.1109947108

Gardner, P. H., & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology*, *9,* 55-79. http://dx.doi.org/10.1002/acp.2350090706

Gino, F. (2008). Do we listen to advice just because we paid for it? The impact of advice cost on its use. *Organizational Behavior and Human Decision Processes*, *107,* 234-245. http://dx.doi.org/10.1016/j.obhdp.2008.03.001

Gino, F., & Schweitzer, M. E. (2008). Blinded by anger or feeling the love: how emotions influence advice taking. *Journal of Applied Psychology, 93*, 1165-1173. http://dx.doi.org/0.1037/0021-9010.93.5.1165

Groenendijk, T., Janssen, T., Rijlaarsdam, G., & van den Bergh, H. (2013a). The effect of observational learning on students' performance, processes, and motivation in two creative domains. *British Journal of Educational Psychology*, *83,* 3-28. http://dx.doi.org/10.1111/j.2044-8279.2011.02052.x

Groenendijk, T., Janssen, T., Rijlaarsdam, G., & van den Bergh, H. (2013b). Learning to be creative. The effects of observational learning on students' design products and processes. *Learning and Instruction*, *28*, 35-47. http://dx.doi.org/10.1016/j.learninstruc.2013.05.001

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*, 117-133. http://dx.doi.org/10.1006/obhd.1997.2697

Kocher, M., Sutter, M., & Wakolbinger, F. (2014). Social Learning in Beauty-Contest Games. *Southern Economic Journal*, *80*, 586-613. http://dx.doi.org/10.4284/0038-4038-2010.150

Laughlin, P. R., & Barth, J. M. (1981). Group-to-individual and individual-to-group problem-solving transfer. *Journal of Personality and Social Psychology, 41*, 1087-1093. http://dx.doi.org/10.1037/0022-3514.41.6.1087

Laughlin, P. R., Bonner, B. L., Miner, A. G., & Carnevale, P. J. (1999). Frames of reference in quantity estimations by groups and individuals. *Organizational Behavior and Human Decision Processes*, *80,* 103-117. http://dx.doi.org/10.1006/obhd.1999.2848

Laughlin, P. R., Gonzalez, C. M., & Sommer, D. (2003). Quantity estimations by groups and individuals: Effects of known domain boundaries. *Group Dynamics: Theory, Research, and Practice*, *7,* 55-63. http://dx.doi.org/10.1037/1089-2699.7.1.55

Laughlin, P. R., & Sweeney, J. D. (1977). Individual-to-group and group-to-individual transfer in problem solving. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 246-254. http://dx.doi.org/10.1037/0278-7393.3.2.246

Minson, J. A., & Mueller, J. S. (2012). The Cost of Collaboration Why Joint Decision Making Exacerbates Rejection of Outside Information. *Psychological Science*, *23*, 219-224. http://dx.doi.org/0.1177/0956797611429132

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. *Psychological Review*, *111*, 781-799. http://dx.doi.org/10.1037/0033-295X.111.3.781

Raedts, M., Rijlaarsdam, G., Van Waes, L., & Daems, F. (2007). Observational learning through video-based models: impact on students' accuracy of self-efficacy beliefs, task knowledge and writing performances. In G. Rijlaarsdam (Series Ed.) & P. Boscolo & S. Hidi (Vol. Eds.), Studies in writing, Vol. 19, writing and motivation (pp. 219-238). Oxford, England: Elsevier. http://dx.doi.org/10.1163/9781849508216_013

Rosenthal, T. L., & Zimmerman, B. J. (1978). *Social learning and cognition*. New York: Academic Press.

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *The Journal of the Learning Sciences*, *14*, 201-241. http://dx.doi.org/10.1207/s15327809jls1402_2

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative to differential weighting. *Organizational Behavior and Human Decision Processes*, *118*, 24-36. http://dx.doi.org/10.1016/j.obhdp.2011.12.006

Schultze, T., Rakotoarisoa, A. F., & Schulz-Hardt, S. (2015). Effects of distance between initial estimates and advice on advice utilization. *Judgment and Decision Making*, *10*, 144-171.

Schunk, D. H., & Hanson, A. R. (1985). Peer models: Influence on children's self-efficacy and achievement. *Journal of Educational Psychology*, *77*, 313-322. http://dx.doi.org/10.1037/0022-0663.77.3.313

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*, 159-174. http://dx.doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, *43*, 1-28. http://dx.doi.org/10.1016/0749-5978(89)90055-1

Sniezek, J. A., & Henry, R. A. (1990). Revision, weighting, and commitment in consensus group judgment. *Organizational Behavior and Human Decision Processes*, *45*, 66-84. http://dx.doi.org/10.1016/0749-5978(90)90005-T

Sniezek, J. A., Schrah, G. E., & Dalal, R. S. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, *17*, 173-190. http://dx.doi.org/10.1002/bdm.468

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 780-805. http://dx.doi.org/10.1037/a0015145

Soll, J. B., & Mannes, A. E. (2011). Judgmental aggregation strategies depend on whether the self is involved. *International Journal of Forecasting*, *27*, 81-102. http://dx.doi.org/10.1016/j.ijforecast.2010.05.003

Stasson, M. F., Kameda, T., Parks, C. D., Zimmerman, S. K., & Davis, J. H. (1991). Effects of assigned group consensus requirement on group problem solving and group members' learning. *Social Psychology Quarterly, 54*, 25-35. http://dx.doi.org/10.2307/2786786

Stern, A., Schultze, T., & Schulz-Hardt, S. (2017). *How much group is necessary? Group-to-individual transfer in estimation tasks.* Unpublished manuscript.

Yaniv, I. (2004a). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*, 75-78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I. (2004b). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*, 1-13. http://dx.doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83*, 260-281. http://dx.doi.org/10.1006/obhd.2000.2909

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*, 104-120. http://dx.doi.org/10.1016/j.obhdp.2006.05.006

Tables and Figures

Figure 1. *Distribution of gains and losses in initial estimate MAPEs across participants in Experiment 1. Each bar represents a participant's gain or loss (sorted in descending order)*

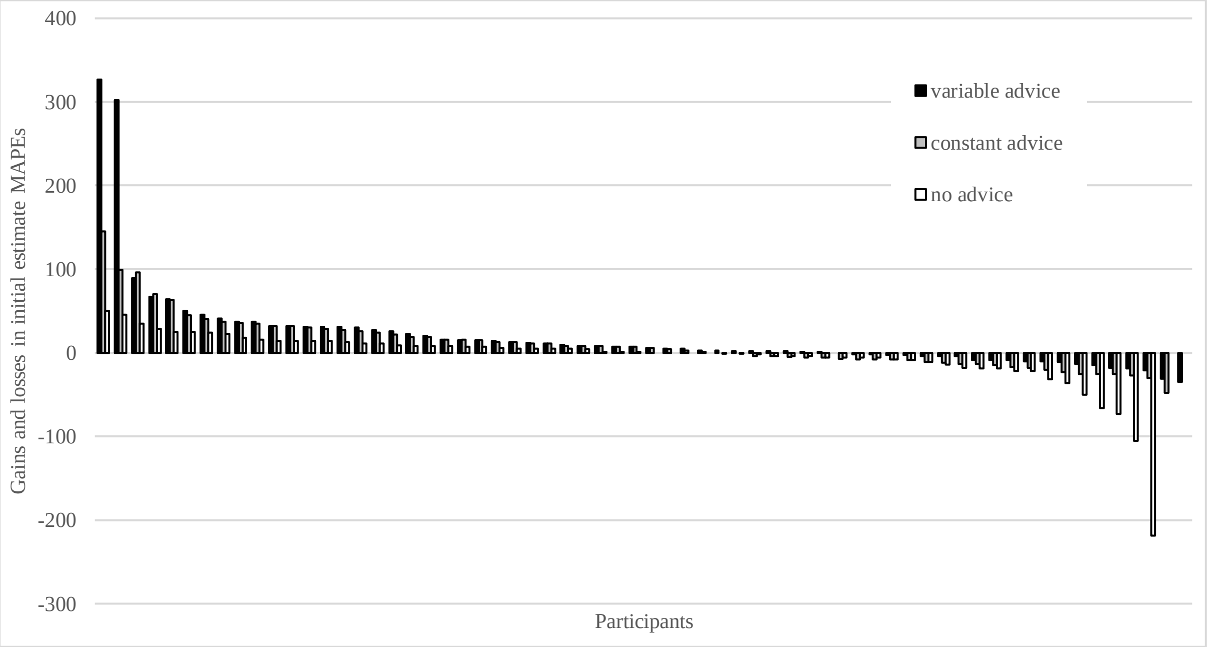

Figure 2. *Distribution of initial estimate MAPEs across participants in Experiment 2. Each bar represents a participant's MAPE (sorted in ascending order)*

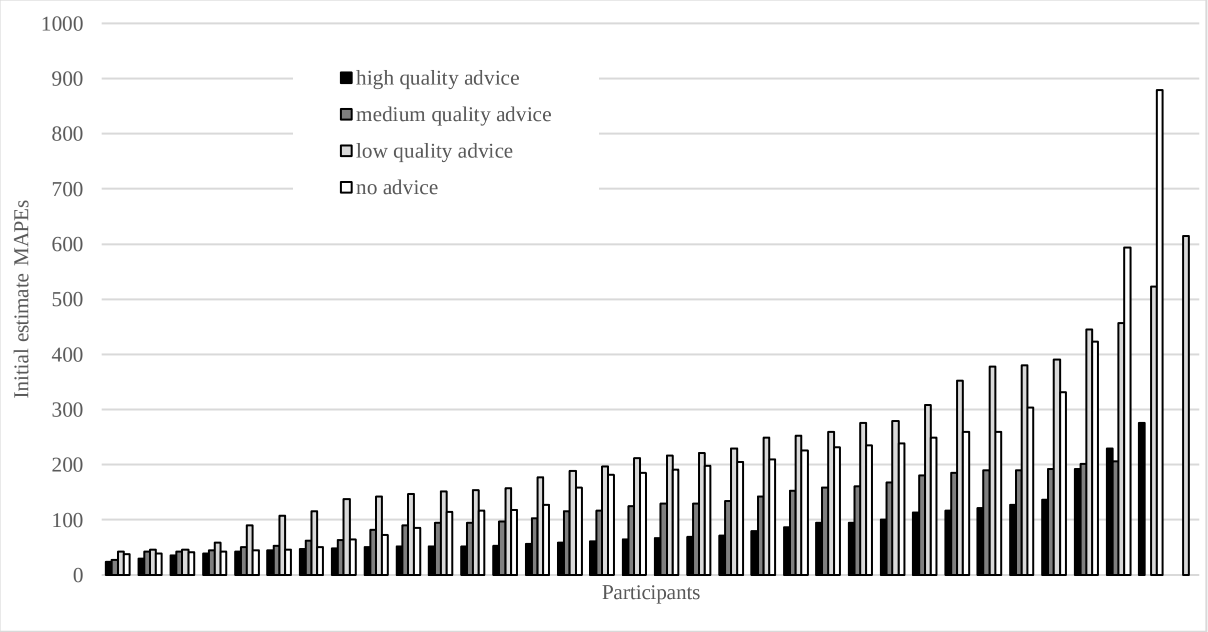

Figure 3. *Distribution of gains and losses in initial estimate MAPEs across participants in Experiment 3. Each bar represents a participant's gain or loss (sorted in descending order)*

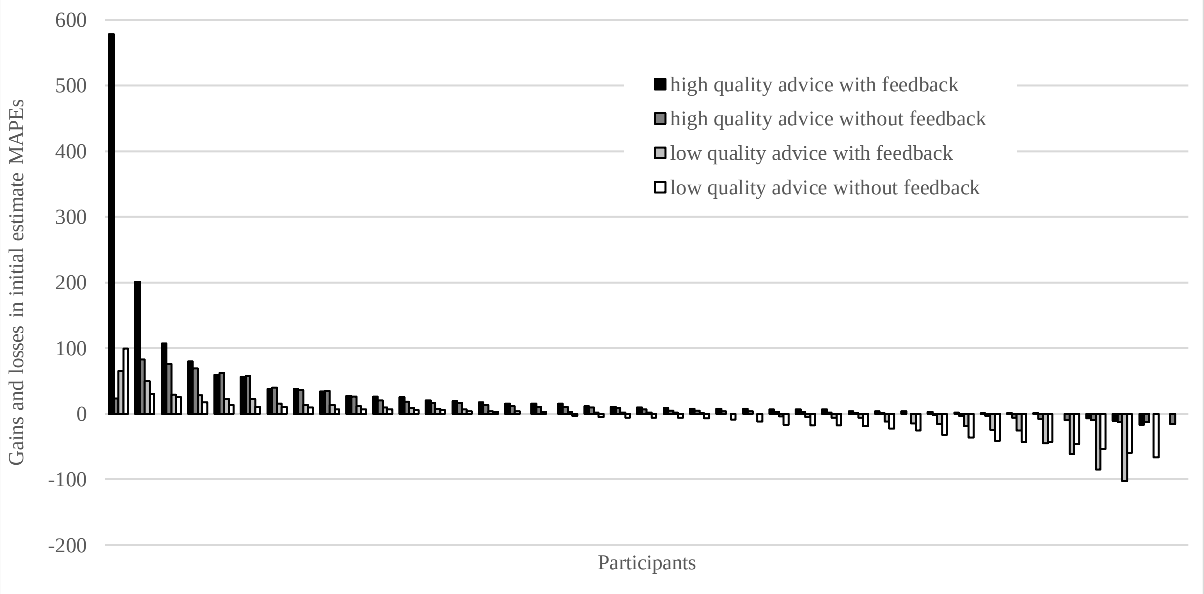

Table 1. *Judges' estimation accuracy, metric error and mapping error by advice condition in Experiment 1*

	advice condition					
	constant advisor		variable advisor		no advice	
	phase 1	phase 2	phase 1	phase 2	phase 1	phase 2
variable	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
accuracy (MAPE)	59.83 (45.66)	48.95 (27.66)	65.97 (65.33)	45.96 (23.88)	65.38 (65.49)	70.25 (80.86)
metric error	47.76 (46.57)	35.10 (24.99)	48.90 (53.99)	32.39 (26.40)	50.25 (62.99)	52.82 (69.34)
mapping error	.69 (.46)	.65 (.34)	.69 (.47)	.70 (.33)	.66 (.41)	.70 (.28)

Table 2. *Judges' estimation accuracy, metric error and mapping error by advice condition in Experiment 2.*

	advice condition			
	high quality	moderate quality	low quality	no advice
variable	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
accuracy (MAPE)	84.40 (56.64)[a]	119.64 (54.10)[a]	235.47 (140.82)[b]	198.80 (172.96)[b]
metric error	44.48 (46.76)[a]	75.21 (49.61)[a]	202.88 (131.18)[b]	161.43 (179.10)[b]
mapping error	.60 (.21)	.61 (.16)	.61 (.24)	.64 (.20)

Groups with different letters significantly differ on the corresponding dependent variable (Tukey post hoc tests)

Table 3. *Judges' estimation accuracy, metric error and mapping error by advice condition and feedback in Experiment 3*

		advice condition			
		high quality		low quality	
		phase 1	phase 2	phase 1	phase 2
variable	feedback	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
accuracy (MAPE)	yes	68.03 (98.01)	31.95 (15.05)	62.96 (51.45)	65.63 (53.90)
	no	56.16 (45.22)	36.61 (20.24)	58.18 (47.06)	66.71 (43.14)
metric error	yes	57.53 (112.99)	16.29 (17.95)	55.32 (55.46)	54.05 (49.83)
	no	39.21 (37.61)	20.54 (25.13)	42.98 (43.71)	50.59 (39.50)
mapping error	yes	.70 (.41)	.71 (.45)	.65 (.41)	.83 (.41)
	no	.74 (.46)	.78 (.36)	.72 (.43)	.66 (.33)

Anhang 3:

Lebenslauf

Lebenslauf

Name: Alexander Stern

Geburtstag: 14. August 1983

Geburtsort: Norderney

Berufserfahrung

Seit 04/2015	wissenschaftlicher Mitarbeiter Abteilung für Pädagogische Psychologie, Georg-August-Universität Göttingen
08/2010 – 12/2015	wissenschaftlicher Mitarbeiter Abteilung für Wirtschafts- und Sozialpsychologie, Georg-August-Universität Göttingen
08/2010 – 09/2015	wissenschaftlicher Mitarbeiter im DFG-Projekt „Koordinationsgewinne durch Gruppenlernen bei diskretionären Aufgaben“ Abteilung für Wirtschafts- und Sozialpsychologie, Georg-August-Universität Göttingen
04/2009 – 06/2010	studentische Hilfskraft Abteilung für Wirtschafts- und Sozialpsychologie, Georg-August-Universität Göttingen
11/2003 – 08/2004	Zivildienst Universitätsklinikum Göttingen (Station für Urologie)

Schule und Studium

08/2010 – heute	Promotion in Psychologie (Dr. rer. nat.) Abteilung für Wirtschafts- und Sozialpsychologie, Universität Göttingen, Thema: Sozial vermittelte Lernprozesse bei quantitativen Schätzaufgaben unter der Betreuung von Prof. Dr. Stefan Schulz-Hardt
10/2004 – 08/2010	Diplom in Sozialwissenschaften (Dipl.-Sozw.) Schwerpunktbereiche: Wirtschafts- und Sozialpsychologie, Politik, Betriebswirtschaftslehre & Arbeitsrecht Georg-August-Universität Göttingen
09/2007 – 02/2008	Auslandssemester

	Universiteit van Amsterdam
08/2000 – 06/2003	Gymnasium am Mühlenweg, Wilhelmshaven
08/1994 – 07/2000	Hans Sachs Gymnasium, Nürnberg
08/1990 – 07/1994	Grundschule Thoner Espan, Nürnberg

<u>Lehrveranstaltungen (Seminare)</u>

SoSe 2017	Seminar „Leistungsbeurteilung, Diagnostik und Intervention in der Schule“ (3 Kurse)
	Universität Göttingen, Master of Education
WiSe 2016/2017	Seminar „Psychologie des Lehrens und Lernens“ (3 Kurse)
	Universität Göttingen, Master of Education
SoSe 2016	Seminar „Psychologie des Lehrens und Lernens“
	Universität Göttingen, Master of Education
WiSe 2015/2016	Seminar „Pädagogische Psychologie II“
	Universität Göttingen, BSc Psychologie
SoSe 2015	Seminar „Pädagogische Psychologie I“
	Universität Göttingen, BSc Psychologie
SoSe 2014	Seminar „Gruppenlernen“
	Universität Göttingen, MSc Psychologie
SoSe 2013	Seminar „Soziale Kognition“
	Universität Göttingen, BSc Soziologie & Ethnologie
	Seminar „Gruppenurteile, Gruppenentscheidungen und Gruppenleistung“
	Universität Göttingen, MSc Psychologie
WiSe 2012/2013	Seminar „Kooperatives Lernen und Lehren“
	Universität Göttingen, BSc Psychologie
	Seminar „Sozialer Einfluss“
	Universität Göttingen, MSc Psychologie
SoSe 2012	Seminar „Kooperatives Lernen und Lehren“
	Universität Göttingen, BSc Psychologie

WiSe 2011/2012	Seminar „Gruppenurteile, Gruppenentscheidungen und Gruppenleistung“ Universität Göttingen, MSc Psychologie

Forschung

Vorträge und Poster

Stern, A., Schultze, T. & Schulz-Hardt, S. (2014). Köhler-Effect without a group? Individual motivation gains through the comparison with the own previous performance. Poster: 17th General Meeting of the European Association of Social Psychology, Amsterdam.

Stern, A., Schultze, T. & Schulz-Hardt, S. (2013). Köhler-Effect without a group? Individual motivation gains through the comparison with the own previous performance. Vortrag: 16th Congress of the European Association of Work and Organizational Psychology, Münster.

Stern, A., Schultze, T. & Schulz-Hardt, S. (2012). Lerneffekte und Gewichtungsstrategien bei Gruppenurteilen. Vortrag: 48. Kongress der Deutschen Gesellschaft für Psychologie, Bielefeld.

Stern, A., Schultze, T. & Schulz-Hardt, S. (2011). Wieviel Gruppe ist nötig? Individuelle Lerneffekte durch Gruppeninteraktion. Vortrag: 13. Tagung der Fachgruppe Sozialpsychologie der DGPs, Hamburg.

Publikationen und Manuskripte

Stern, A., Drewes, S. & Schulz-Hardt, S. (2017). Gruppenleistung. In D. Frey & H.-W.-Bierhoff (Hrsg.) Enzyklopädie Sozialpsychologie. Göttingen: Hogrefe.

Stern, A., Schultze, T. & Schulz-Hardt, S. (2017a). How much group is necessary? Group-to-individual transfer in judgment tasks. Unpublished manuscript.

Stern, A., Schultze, T. & Schulz-Hardt, S. (2017b). Social learning in the judge-advisor-system. A neglected advantage of advice-taking. Manuscript submitted for publication.

Unterschrift