Methods and Designs

Florian Schmiedek

Contents

Abstract Cognitive training research faces a number of methodological challenges. Some of these are general to evaluation studies of behavioral interventions, like selection effects that confound the comparison of treatment and control groups with preexisting differences in participants' characteristics. Some challenges are also specific to cognitive training research, like the difficulty to tell improvements in general cognitive abilities from improvements in rather task-specific skills. Here, an overview of the most important challenges is provided along an established typology of different kinds of validity (statistical conclusion, internal, external, and construct validity) that serve as the central criteria for evaluating intervention studies. Besides standard approaches to ensure validity, like using randomized assignment to experimental conditions, emphasis is put on design elements that can help to raise the construct validity of the treatment (like adding active control groups) and of the outcome measures (like using latent factors based on measurement models). These considerations regarding study design are complemented with an overview of data-analytical approaches based on structural equation modeling, which have a number of advantages in comparison to the still predominant approaches based on analysis of variance.

F. Schmiedek (✉)
DIPF | Leibniz Institute for Research and Information in Education,
Frankfurt am Main, Germany
e-mail: schmiedek@dipf.de

© Springer Nature Switzerland AG 2021

T. Strobach, J. Karbach (eds.), *Cognitive Training*,
https://doi.org/10.1007/978-3-030-39292-5_2

Introduction

Researchers who aim to investigate the effectiveness of cognitive trainings can draw on the well-established methodology for the evaluation of behavioral interventions in psychology and education (Murnane and Willett 2011; Shadish et al. 2002). Doing so, they face a long list of potential issues that can be characterized as threats to different types of the validity of findings. Here, the most common and relevant threats, as well as possible methodological approaches and study design elements to reduce or rule out these threats in the context of cognitive training studies, will be discussed.

The commonly preferred design for investigating cognitive training interventions is one with random assignment of a sample of participants to training and control groups with pre- and posttest assessments of a selection of tasks chosen to represent one or more cognitive abilities that the training might potentially improve. Significantly larger average improvements on such outcome measures in the training than in a control group are taken as evidence that the training benefits cognition. Such a design indeed clears out a number of potential issues. Certain problems that arise when evaluating cognitive trainings, however, require solutions that go beyond, or modify, commonly used of-the-shelf study design elements. For example, the inclusion of no-treatment control groups for ruling out threats to internal validity and the use of single tasks as outcome measures of transfer effects are associated with certain deficits. In the following, methodological problems and challenges will be discussed along the established typology of statistical conclusion validity, internal and external validity, as well as construct validity (Shadish et al. 2002).

Statistical Conclusion Validity

Statistical conclusion validity refers to whether the association between the treatment and the outcome can be reliably demonstrated. Such demonstration is based on inferential statistics, which can provide evidence that observed differences between experimental groups in posttest scores, or in pretest-to-posttest changes, are unlikely to be due to sampling error (i.e., one group having higher scores simply by chance). Given that existing training studies mostly have relatively small sample sizes (with experimental groups of more than 30–40 participants being rare exceptions), the statistical power to do so often is low, and the findings are in danger of being difficult to replicate and being unduly influenced by outliers and violations of statistical assumptions.

Furthermore, and in light of recent discussions about the replicability of findings and deficient scientific standards in psychological research (e.g., Maxwell et al. 2015), there is the problem that low power might increase researchers' propensity to lapse into fishing-for-effect strategies. Given that (a) the researchers' desired hypothesis often will be that a training has a positive effect, (b) that training studies

are resource-intensive, and (c) that the nonregistered analysis of data allows for a number of choices of how exactly to be conducted (Fiedler 2011), it has to be considered a danger that such choices (like choosing subsamples or subsets of outcome tasks) are made post hoc in favor of "finding" significant effects and thereby invalidate the results of inferential test statistics. In combination with publication biases that favor statistically significant over nonsignificant results, such practices in a field with typically low power could lead to a distorted picture of training effectiveness, even in meta-analyses. A general skepticism should therefore be in place regarding all findings that have not been replicated by independent research groups. Regarding the danger of fishing-for-effects practices, preregistration of training studies, including the specific hypotheses and details of data preparation and analysis, is a possible solution, which is well established in the context of clinical trials and gaining acceptance, support, and utilization in science in general (Nosek et al. 2018). In general, effort should be invested to increase statistical power and precision of effect size estimates. Besides large enough sample sizes, this also includes ensuring high reliability of outcome measures and of treatment implementation.

As an alternative to null hypothesis significance testing, which still dominates most of the cognitive training research, the use of a Bayesian inference framework should also be considered (Wagenmakers et al. 2018). A dedicated implementation of such a framework would require the use of knowledge and expectations regarding the distribution of effect sizes as priors in the analyses. Even without consent to such a fully Bayesian perspective, however, the use of Bayes factors offers a useful and sensible alternative to null hypothesis significance testing (Dienes 2016). Particularly when it is not clear whether a training program has any notable effect, and therefore the null hypothesis of no effect is a viable alternative, Bayes factors have the advantage that they allow quantifying evidence for the null hypothesis as well as for the hypothesis of an effect being present. When studies have sufficient statistical power, such analyses can result in strong and conclusive evidence for the null hypothesis, and thereby allow for a sobering acceptance of a certain training not producing the desired effects – something null hypothesis testing cannot provide (see von Bastian et al. 2020, for an evaluation of working memory training studies using Bayes factors).

Internal Validity

Internal validity, that is, a study's ability to unambiguously demonstrate that the treatment has a causal effect on the outcome(s), deserves getting a strong weight when judging the quality of intervention studies. It involves ruling out alternative explanations for within-group changes (including practice effects, maturation, or statistical regression to the mean from pretest to posttest) and/or between-group differences (e.g., systematic selection effects into the treatment condition). Common reactions to these problems are requests to (a) use a control group that allows to estimate the size of the effects due to alternative explanations and to (b) randomly

assign participants into the different groups. While intact random assignment assures that the mean differences between groups can be unbiased estimates of the *average causal effect* of the treatment (Holland 1986), several cautionary notes are at place regarding this "gold standard" of intervention studies.

First, the unbiasedness of the estimate refers to the expected value. This does not rule out that single studies (particularly if sample sizes are small) have groups that are not well comparable regarding baseline ability or other person characteristics that might interact with the effectiveness of the training. Therefore, the amount of trust in effect size estimates should only be high for studies with large samples or for replicated (meta-analytic) findings. For single studies with smaller samples, matching techniques based on pretest scores can help to reduce random differences between groups that have an effect on estimates of training effects.

Second, the benefits of randomization get lost if the assignment is not "intact," that is, if participants do not participate in the conditions they are assigned to or do not show up for the posttest. Such lack of treatment integrity or test participation can be associated with selection effects that turn an experiment into a quasi-experiment – with all the potential problems of confounding variables that can affect the estimate of outcome differences. In such cases of originally randomized, but later on nonintact experiments, instrumental variable estimation (using the randomized assignment as an instrument for the realized treatment variable) can be used to still get unbiased estimates of the causal effect of the treatment for the subpopulation of participants who comply with the treatment assignment (Angrist et al. 1996). Instrumental variable estimation requires larger samples, however, than those available in many cognitive training studies.

Third, formal analysis of causal inference based on randomized treatment assignment (Holland 1986) shows that the interpretation of mean group differences as average causal effects is only valid if participants do not interact with each other in ways that make individual outcomes dependent on whether or not particular other participants are assigned to the treatment or the control condition. While this is unlikely to pose a problem if training is applied individually, it could be an issue that has received too little attention in studies with group-based interventions – where interactions among participants might, for example, influence motivation. In such cases, a viable solution is to conduct a cluster-randomized experiment and randomize whole groups of participants into the experimental conditions. If groups systematically differ in outcome levels before the training, however, the power of such a study can be considerably lower than it would be if the same number of participants would be assigned individually to experimental conditions. To achieve sufficient power, often much larger total sample sizes and a careful choice of covariates at the different levels of analysis (i.e., individuals and groups) will be necessary (Raudenbush et al. 2007).

Whenever treatment assignment cannot be random, due to practical or ethical considerations, or when randomization breaks down during the course of the study, careful investigation of potential selection effects is required. This necessitates the availability of an as-complete-as-possible battery of potential confounding variables at pretest. If analyses of such variables indicate group differences, findings cannot unambiguously be attributed to the treatment. Attempts to remedy such

group differences with statistical control techniques are associated with strong conceptual (i.e., exhaustiveness of the available information regarding selection effects and correctness of the assumed causal model) and statistical assumptions (e.g., linearity of the relation with the outcome) and should therefore be regarded with great caution. An alternative to regression-based control techniques is post hoc matching and subsample selection based on propensity score analyses (Guo and Fraser 2014). This requires sample sizes that are typically not available in cognitive training research, however. Beneficial alternative design approaches for dealing with situations in which randomization is not possible, or likely to not stay intact, are available, like regression discontinuity designs or instrumental variable approaches (Murnane and Willett 2011), but have received little attention in cognitive training research so far.

Construct Validity

While the demonstration of causal effects of the treatment undoubtedly is a necessity when evaluating cognitive trainings, a strong focus on internal validity and randomization should not distract from equally important aspects of construct validity. Addressing the question of whether the investigated variables really represent the theoretical constructs of interest, construct validity is relevant for both, the treatment as well as the outcome measures.

Regarding the treatment, high internal validity does only assure that one or more aspects that differentiate the treatment from the control condition causally influence the outcome. It does not tell which aspect of the treatment it is, however. Given the complexity of many cognitive training programs and the potential involvement of cognitive processes as well as processes related to motivation, self-concept, test anxiety, and other psychological variables in producing improvements in performance, the comparison to so-called *no-contact control conditions* typically cannot exclude a number of potential alternative explanations of why an effect has occurred. In the extreme case, being in a no-contact control condition and still having to redo the assessment of outcome variables at posttest is so demotivating that performance in the control group declines from pre- to posttest. Such a pattern has been observed in several cognitive training studies and renders the interpretation of significant interactions of groups (training vs. control) and occasions (pretest vs. posttest) as indicating improved cognitive ability very difficult to entertain (Redick 2015). As from a basic science perspective, the main interest is in effects that represent plastic changes of the cognitive system; “active” control conditions therefore need to be designed, which are able to produce the same nonfocal effects, but do not contain the cognitive training ingredient of interest. This is a great challenge, however, given the number and complexity of cognitive mechanisms that potentially are involved in processing of, for example, working memory tasks and that can be affected by training (Von Bastian and Oberauer 2014; Könen et al., this volume). For many of these mechanisms, like the use of certain strategies, practice-related improvements are possible, but would have to be considered exploitations of

existing behavioral flexibility, rather than extensions of the range of such behavioral flexibility (Lövdén et al. 2010). If motivational effects are partly due to the joy of being challenged by complex tasks, it also will be difficult to invent tasks of comparably joyful complexity but little demand on working memory. In addition to inventive and meticulous creation of control conditions, it is therefore necessary to assess participants' expectations, task-related motivation, and noncognitive outcomes, before, during, and after the intervention (see also Cochrane and Green, Katz et al., this volume).

Regarding the outcome variables, construct validity needs to be discussed in light of the issue of transfer distance and the distinction between skills and abilities. When the desired outcome of a training is the improvement of a specific skill or the acquisition of a strategy tailored to support performing a particular kind of task, the assessment of outcomes is relatively straightforward – it suffices to measure the trained task itself reliably at pre- and posttest. As the goal of cognitive trainings typically is to improve an underlying broad ability, like fluid intelligence or episodic memory, demonstrating improvements on the practiced tasks is not sufficient, however, as those confound potential changes in ability with performance improvements due to the acquisition of task-specific skills or strategies. It is therefore common practice to employ transfer tasks that represent the target ability but are different from the trained tasks. The question of how different such transfer tasks are from the trained ones is often answered using arguments of face validity and classifications as "near" and "far" that are open to criticism and difficult to compare across studies. What seems far transfer to one researcher might be considered near transfer by another one. Particularly if only single tasks are used as outcome measure for a cognitive ability, it is difficult to rule out alternative explanations that explain improvements with a task-specific *skill*, rather than with improvements in the underlying *ability* (see, e.g., Hayes et al. 2015, or Moody 2009).

The likelihood of such potential alternative explanations can be reduced if the abilities that a training is thought to improve are operationalized with several heterogeneous tasks that all have little overlap with the trained tasks and are dissimilar from each other in terms of paradigm and task content. The analysis of effects can then be conducted on the shared variance of these tasks, preferably using confirmatory factor models. This allows to analyze transfer at the level of latent factors that represent the breadth of the ability construct, replacing the arbitrary classification of "near vs. far" with one that defines "narrow" or "broad" abilities by referring to well-established structural models of cognitive abilities (Noack et al. 2009). If transfer effects can be shown for such latent factors, this renders task-specific explanations less likely.

External Validity

External validity encompasses the generalizability of a study's results to other samples, as well as to other contexts, variations of the intervention's setting, and different outcome variables. As few training studies are based on samples that are representative for broad populations, mostly little is known regarding

generalizability to different samples. Furthermore, as findings for certain training programs are only rarely replicated by independent research groups, we only have very limited evidence so far regarding the impact of variations of the context, setting, and of the exact implementation of cognitive trainings. As one rare exception, the Cogmed working memory training (http://www.cogmed.com/) has been evaluated in a number of studies by different research groups and with diverse samples. This has resulted in a pattern of failed and successful replications of effects that has been reviewed as providing little support for the claims that have been raised for the program (Shipstead et al. 2012a, b).

Similarly, generalizations of effects for certain transfer tasks to real-life cognitive outcomes, like everyday competencies and educational or occupational achievement, are not warranted, unless shown with direct measures of these outcomes. Even if transfer tasks are known to have strong predictive validity for certain outcomes, this does not ensure that *changes* in transfer task performance show equally strong relations to *changes* in the outcomes (Rode et al. 2014). Finally, relatively little is known about maintenance and long-term effects of cognitive trainings. Here, the combination of training interventions and longitudinal studies would be desirable. In sum, there is a need for studies that reach beyond the typically used convenience samples and laboratory-based short-term outcomes, as well as beyond research groups' common practice of investigating their own pet training programs – to explore the scope, long-term effects, and boundary conditions of cognitive trainings in a systematic way.

Types of Studies

Trying to optimize the different kinds of validity often leads to conflicts because limited resources prohibit maximization of all aspects simultaneously. Furthermore, certain decisions regarding research design may need be to made against the background of direct conflicts among validity aspects. Maximizing statistical conclusion validity by running an experiment in strictly controlled laboratory conditions, for example, may reduce external validity. Balancing the different kinds of validity when planning studies requires to acknowledge that intervention studies may serve quite different purposes. Green et al. (2019) differentiate *feasibility studies*, *mechanistic studies*, *efficacy studies*, and *effectiveness studies* and discuss important differences between these regarding the study methodology, some of which shall be briefly summarized here (see also Cochrane and Green, this volume).

Feasibility studies serve to probe, for example, the viability of new approaches, the practicality of technological innovations, or the applicability of a training program to a certain population. They are typically implemented before moving to one or more of the other kinds of studies. In feasibility studies, the samples may be small in size, but carefully drawn from the target population to, for example, identify potential implementation problems early on. Control groups may often not be necessary, as the focus is not on demonstrating a causal effect yet. Outcome

variables may also be more varied and include aspects like compliance rates or subjective ratings of aspects of the training program.

Mechanistic studies test specific hypotheses deducted from a theoretical framework with the aim of identifying the causally mediating mechanisms and moderating factors underlying training-related performance improvements. As such, they provide the basic research fundamentals on which interventions with applied aims can be built. Furthermore, cognitive intervention studies may also serve to answer general questions about cognitive development and the range of its malleability, as for example in the testing-the-limits paradigm (Lindenberger and Baltes 1995), without the goal of generating available training programs. Trying to confirm or explore specific mechanisms of training-related cognitive changes, mechanistic studies will often require different kinds of training and control conditions (to generate the appropriate experimental contrasts) than efficacy and effectiveness studies, which are rather interested in the combined effect of all cognitive change processes involved. Similarly, the outcome variables of mechanistic studies may rather serve to identify a specific cognitive process than to demonstrate broad transfer effects of practical relevance.

Efficacy studies aim at establishing a causal effect of an intervention in comparison to some placebo or other standard control conditions and at thereby answering the question *"Does the paradigm produce the anticipated outcome in the exact and carefully controlled population of interest when the paradigm is used precisely as intended by the researchers?"* (Green et al. 2019, p. 6). Here, ensuring internal validity is of critical importance, as is construct validity of treatment and outcomes and the consideration of sufficient statistical power.

Finally, effectiveness studies aim at evaluating the outcomes of an intervention when implemented in real-world settings. Because such deployment and scaling up of interventions typically is associated with less control over the sampling of participants and fidelity of the dosage and quality of the intervention; the weighting of prime criteria shifts from internal validity to external validity. Control conditions typically will be the "business-as-usual" that is present without an intervention and a relatively stronger focus will lie on evaluating real-life outcome criteria, unwanted side effects, and long-term maintenance of training gains (Green et al. 2019).

Data Analysis

The standard data-analytical approach to the pretest–posttest control-group design in most studies still is a repeated measures ANOVA with *group* (training vs. control) as a between- and *occasion* (pretest vs. posttest) as a within-subject factor, and with a significant interaction of the two factors taken as evidence that observed larger improvements in the training than in the control group indicate a reliable effect of treatment. If there is interest in individual differences in training effects (Katz et al., this volume), either subgroups or interactions of the within-factor with covariates are analyzed. This approach comes with a number of limitations, however.

First, the associated statistical assumptions of sphericity and homogeneity of (co)variances across groups might not be met. For example, when a follow-up occasion (months or years after training) is added, sphericity is unlikely to hold across the unequally spaced time intervals. When the training increases individual differences in performance more than the control condition, homogeneity of variances might not be provided. Second, participants with missing data on the posttest occasions have to be deleted listwise (i.e., they are completely removed from the analysis). Third, analyses have to be conducted on a single-task level. This means that unreliability of transfer tasks can bias results and that, if several transfer tasks for the same ability are available, analyses have to be conducted either one by one or on some composite score. Fourth, when comparability of experimental groups is not ensured by randomized assignment to conditions, the prominent use of ANCOVA, using the pretest as a covariate to adjust for potential pretreatment group differences in the outcome, can be associated with further problems. Regarding causal inference, controlling for pretest scores will only lead to an unbiased estimate of the causal effect of the treatment if the pretest (plus other observed confounders entered as additional covariates) can be assumed to sufficiently control for all confounding that is due to unmeasured variables (Kim and Steiner in press). If this assumption cannot be made with confidence, but instead the assumptions that unmeasured confounders do influence pretest and posttest scores to the same degree (i.e., that confounding variables are time-invariant trait-like characteristics of the participants) and that the pretest does not influence the treatment assignment are likely to hold, then the use of analyses based on gain scores may be preferable over ANCOVA (Kim and Steiner in press).

The first three potential problems mentioned above can be cleared out by basing analyses on a structural equation modeling framework and using latent change score models (McArdle 2009; see also Könen and Auerswald, this volume). Provided large enough samples, multigroup extensions of these models (Fig. 1) allow testing all the general hypotheses typically addressed with repeated measures ANOVA – and more – while having several advantages: First, assumptions of sphericity and homogeneity of (co)variances are not necessary, as (co)variances are allowed to vary across groups and/or occasions. Second, parameter estimation based on full information maximum likelihood allows for missing data. If there are participants who took part in the pretest but dropped out from the study and did not participate in the posttest, their pretest score can still be included in the analysis and help to reduce bias of effect size estimates due to selective dropout (Schafer and Graham 2002). Third, change can be analyzed using latent factors. This has the advantage that effects can be investigated with factors that (a) capture what is common to a set of tasks that measure the same underlying cognitive ability and (b) are free of measurement error. This provides estimates of training effects that are not biased by unreliability of tasks. It also allows investigating individual differences in change in a way that is superior to the use of individual difference scores, which are known to often lack reliability. For example, the latent change score factor for a cognitive outcome could be predicted by individual differences in motivation, be used to

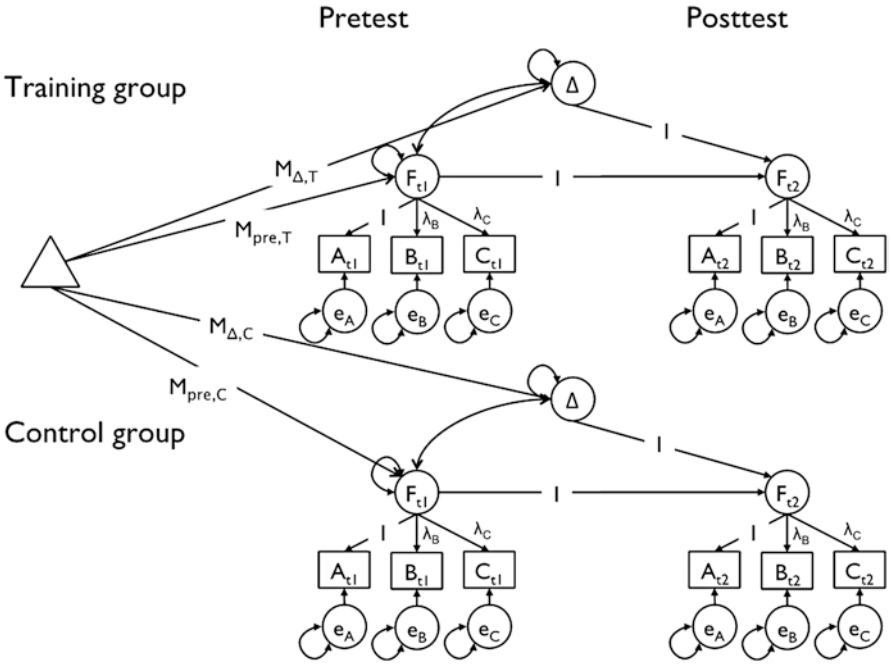

Fig. 1 Two-group latent change score model for pretest–posttest changes in a cognitive training study. Changes are operationalized as the latent difference (Δ) between latent factors at pretest (F_{t1}) and posttest (F_{t2}). These factors capture the common variance of a set of indicator tasks (A, B, and C). Ideally, factor loadings (λ), variances of the residual terms (e), and task intercepts (not shown) are constrained to be equal across groups and occasions (i.e., strict measurement invariance). Based on this model, hypotheses regarding group differences in pretest mean levels (M_{Pre}) and mean changes from pre- to posttest ($M_Δ$) can be investigated, as well as hypotheses regarding the variance and covariance of individual differences in pretest levels and changes (double-headed curved arrows on latent factors)

predict other outcomes (e.g., wellbeing), or be correlated with latent changes in other trained or transfer tasks (e.g., McArdle and Prindle 2008).

Regarding the fourth potential problem of potentially biased estimates in experiments with nonrandom assignment to conditions, latent change score models also allow for a choice between both general options – either analyzing (latent) gain scores or conducting ANCOVA-like adjustments for pretest scores – depending on which assumptions are thought to be more likely to hold.

Furthermore, these models can be extended using the full repertoire of options available in advanced structural equation models. These include multilevel analysis (e.g., to account for the clustering of participants in school classes), latent class analysis (e.g., to explore the presence of different patterns of improvements on a set of tasks), item response models (e.g., to model training-related changes at the level of responses to single items), and more.

Besides a lack of awareness of these advantages, three requirements of latent change score models might explain why they have been used relatively little in cognitive training research so far (Noack et al. 2014). First, these models typically require larger sample sizes than those available in many training studies. When analyzed in a multigroup model with parameter constraints across groups, however, it may be sufficient to have smaller sample sizes in each group than those typically requested for structural equation modeling with single groups. Second, the models require measurement models for the outcome variables of the training. As argued above, operationalizing outcomes as latent variables with heterogeneous task indicators also has conceptual advantages. If only single tasks are available, it still might be feasible to create a latent factor using parallel versions of the task (e.g., based on odd and even trials) as indicator variables. Third, these measurement models need to be invariant across groups and occasions to allow for unequivocal interpretation of mean changes and individual differences therein at the latent factor level (Vandenberg and Lance 2000; see also Könen and Auerswald, this volume). This includes equal loadings, intercepts, and preferably also residual variances of indicator variables. While substantial deviations from measurement invariance can prohibit latent change score analyses, they at the same time can be highly informative, as they can indicate the presence of task-specific effects.

Summary and Outlook

The field of cognitive training research is likely to stay active, due to the demands from societies with growing populations of older adults and attempts to improve the fundamentals of successful education and lifelong learning. As reviewed along the different validity types, this research faces a list of challenges, to which still more could be added (for other methodological reviews and recently discussed issues, see Boot and Simons 2012; Green et al. 2014; Strobach and Schubert 2012; Shipstead et al. 2012a, b; Tidwell et al. 2013). At the same time, awareness of the methodological issues seems to be increasing so that there is a reason to be optimistic that evaluation criteria for commercial training programs (like preregistration of studies) will be established, methodological standards regarding research design will rise, and available advanced statistical methods and new technological developments (like ambulatory assessment methods to assess outcomes in real-life contexts) will be used. Together with basic experimental and neuroscience research on the mechanisms underlying plastic changes in cognition (Wenger and Kühn, this volume), this should lead to better understanding of whether, how, and under which conditions different cognitive training interventions produce desirable effects.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444–455.

Boot, W. R., & Simons, D. J. (2012). Advances in video game methods and reporting practices (but still room for improvement): A commentary on Strobach, Frensch, and Schubert (2012). *Acta Psychologica, 141*, 276–277.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72*, 78–89.

Fiedler, K. (2011). Voodoo correlations are everywhere – not only in neuroscience. *Perspectives on Psychological Science, 6*, 163–171.

Green, C. S., Strobach, T., & Schubert, T. (2014). On methodological standards in training and transfer experiments. *Psychological Research, 78*, 756–772.

Green, C. S., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., ... & Facoetti, A. (2019). Improving methodological standards in behavioral interventions for cognitive enhancement. *Journal of Cognitive Enhancement, 3*, 2–29.

Guo, S., & Fraser, W. M. (2014). *Propensity Score Analysis: Statistical Methods and Applications* (2nd ed.). Thousand Oaks: Sage Publications, Inc..

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence, 48*, 1–15.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.

Kim, Y., & Steiner, P. M. (in press). Gain scores revisited: A graphical models perspective. *Sociological Methods & Research*. https://doi.org/10.1177/0049124119826155.

Lindenberger, U., & Baltes, P. B. (1995). Testing-the-limits and experimental simulation: Two methods to explicate the role of learning in development. *Human Development, 38*, 349–360.

Lövdén, M., Bäckman, L., Lindenberger, U., Schaefer, S., & Schmiedek, F. (2010). A theoretical framework for the study of adult cognitive plasticity. *Psychological Bulletin, 136*, 659–676.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 6*, 487–498.

McArdle, J. J. (2009). Latent variable modelling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605.

McArdle, J. J., & Prindle, J. J. (2008). A latent change score analysis of a randomized clinical trial in reasoning training. *Psychology and Aging, 23*, 702–719.

Moody, D. E. (2009). Can intelligence be increased by training on a task of working memory? *Intelligence, 37*, 327–328.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford: Oxford University Press.

Noack, H., Lövdén, M., Schmiedek, F., & Lindenberger, U. (2009). Cognitive plasticity in adulthood and old age: Gauging the generality of cognitive intervention effects. *Restorative Neurology and Neuroscience, 27*, 435–453.

Noack, H., Lövdén, M., & Schmiedek, F. (2014). On the validity and generality of transfer effects in cognitive training research. *Psychological Research, 78*, 773–789.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *PNAS, 115*, 2600–2606.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*, 5–29.

Redick, T. S. (2015). Working memory training and interpreting interactions in intelligence interventions. *Intelligence, 50*, 14–20.

Rode, C., Robson, R., Purviance, A., Geary, D. C., & Mayr, U. (2014). Is working memory training effective? A study in a school setting. *PLoS One, 9*, e104796.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for causal inference*. Boston: Houghton Mifflin.

Shipstead, Z., Hicks, K. L., & Engle, R. W. (2012a). Cogmed working memory training: Does the evidence support the claims? *Journal of Applied Research in Memory and Cognition, 1*, 185–193.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012b). Is working memory training effective? *Psychological Bulletin, 138*, 628–654.

Strobach, T., & Schubert, T. (2012). Video game experience and optimized cognitive control skills–On false positives and false negatives: Reply to Boot and Simons (2012). *Acta Psychologica, 141*, 278–280.

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. R., Thomas, R. P., & Mendoza, J. L. (2013). What counts as evidence for working memory training? Problems with correlated gains and dichotomization. *Psychonomic Bulletin & Review, 21*, 620–628.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70.

Von Bastian, C. C., & Oberauer, K. (2014). Effects and mechanisms of working memory training: A review. *Psychological Research, 78*, 803–820.

von Bastian, C. C., Guye, S., & de Simoni, C. (2020). How strong is the evidence for the effectiveness of working memory training. In J. M. Novick, M. F. Bunting, M. R. Dougherty, & R. W. Engle (Eds.), *Cognitive and working memory training: Perspectives from psychology, neuroscience, and human development* (pp. 58–78). New York: Oxford University Press.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35–57.