Small-sample adjustments for multiple-contrast hypothesis tests of meta-regressions using robust variance estimation

James E. Pustejovsky
The University of Texas at Austin

Elizabeth Tipton
Teachers College, Columbia University

July 8, 2015
Society for Research Synthesis Methods
Vanderbilt University, Nashville, TN

Robust variance estimation (RVE)

- Robust variance estimation (RVE) is a method for constructing **asymptotically valid** SEs, hypothesis tests, and CIs when the variance or dependence structure of a regression model is **unknown** or **mis-specified**.

- In meta-analysis/meta-regression, RVE is useful for:
  - Univariate meta-analysis (Sidik & Jonkman, 2006), if sampling variances are inaccurate.
  - Meta-regression with dependent effect sizes (Hedges, Tipton, & Johnson, 2010), if correlations between effect size estimates are not available.

RVE in large samples

- RVE standard errors are **asymptotically valid,** i.e., when the number of independent studies (*m*) is sufficiently large.
  - But standard errors tend to be **too small** when *m* is small.

- For testing single meta-regression coefficients, the z-statistic (estimate / robust SE) is normally distributed if *m* is sufficiently large.
  - But z-test has **inflated Type I error** when *m* is small.

- For testing hypotheses involving multiple coefficients, a Wald statistic will follow a chi-squared distribution if *m* is sufficiently large.
  - But Wald test has **severely inflated Type-I error** if *m* is not "large enough."

Example: Wilson, Lipsey, Tanner-Smith, Huang, & Steinka-Fry (2011)

- Meta-analysis of dropout prevention/intervention programs
  - Primary outcomes: school completion, school dropout
  - m = 152 studies
  - N = 385 effect size estimates
  - Many studies provided effect size estimates for multiple outcome measures based on the same sample of participants.
- Original analysis used RVE (without small-sample correction)
- Meta-regression model including five categorical moderators.
  - E.g., Study design: randomized experiment, matched, uncontrolled

Small-sample adjustments

- Tipton (In Press) devised a small-sample correction for single-coefficient tests, involving
  - Adjustments to the RVE formula
  - Estimated degrees-of-freedom based on a Satterthwaite approximation
- Our work provides small-sample corrections for multiple-contrast hypothesis tests.
  - Tests of equality of several levels of a moderator variable
  - Tests of overall model fit

Small-sample F-test

- Linear hypothesis with q contrasts: $\mathbf{C}\boldsymbol{\beta} = \mathbf{c}$.
- We consider adjustments to the Wald statistic

$$Q = \left(\mathbf{Cb} - \mathbf{c}\right)^T \left[\mathbf{C}\mathbf{V}^R\mathbf{C}^T\right]^{-1}\left(\mathbf{Cb} - \mathbf{c}\right)$$

  where **b** is the vector of coefficient estimates and $\mathbf{V}^R$ is the robust variance estimator.

- A two-part adjustment:
  1. Following Tipton (In Press), adjust $\mathbf{V}^R$ using the McCaffrey, Bell, & Botts (2001) "bias reduced linearization" approach.
  2. Approximate the distribution of Q using an F-distribution with estimated degrees-of-freedom.
- We investigated a wide variety of different degrees-of-freedom approximations.

The Winner: AHZ

- Match mean and **total variance** of $\mathbf{CV}^R\mathbf{C}^T$ to a Wishart distribution with η degrees of freedom.

- Approximate the distribution of Q by Hotelling's T^2 distribution:

$$\left(\frac{\eta - q + 1}{\eta q}\right) \times Q \mathrel{\dot\sim} \mathrm{F}\left(\mathrm{q}, \eta - q + 1\right)$$

- In an extensive set of simulations, we found that AHZ:
  - Nearly always had Type I error less than or equal to the nominal α
  - More accurate than the other level-α corrections
  - Tended to be conservative (Type I error < α) in small samples.

Example: Wilson et al. (2011)

		χ^2 test		AHZ test		
Moderator	q	Q	p-val	F	d.f.	p-val
Study design (3 levels)	2	0.46	.796	0.22	43	.801
Outcome measure (4 levels)	3	2.74	.436	0.84	22	.489
Evaluator independence (4 levels)	3	9.33	.029	2.78	17	.073
Implementation quality (3 levels)	2	28.31	<.001	13.78	37	<.001
Program format (4 levels)	3	11.54	.011	3.65	38	.021

- Based on m = 152 studies, N = 385 effect sizes.
- Weights based on "hierarchical" model proposed by Hedges et al. (2010).

Conclusions and future work

- Small-sample corrections **should always be used** in practice.
  - The performance of the large-sample test depends on **features of the covariates** (e.g., balance, leverage), not just sample size.
  - Consequently, it is hard to say what constitutes a "large enough" sample.
- Single- and multiple-contrast hypothesis tests implemented in R package **clubSandwich**
  - Works with **metafor** (Viechtbauer, 2010) and **robumeta** (Fisher & Tipton, 2015)
  - Currently available on Github (https://github.com/jepusto/clubSandwich)
- Interested in helping us implement in Stata?

Questions?

- Working paper available upon request
- Ask about our simulation results!

James E. Pustejovsky – pusto@austin.utexas.edu

Elizabeth Tipton – tipton@tc.columbia.edu

References

Fisher, Z., & Tipton, E. (2014). robumeta: An R-package for robust variance estimation in meta-analysis. *Journal of Statistical Software*.

Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65.

McCaffrey, D. F., Bell, R. M., & Botts, C. H. (2001). Generalizations of biased reduced linearization. In *Proceedings of the Annual Meeting of the American Statistical Association*.

Sidik, K., & Jonkman, J. N. (2006). Robust variance estimation for random effects meta-analysis. *Computational Statistics & Data Analysis*, *50*(12), 3681–3701.

Tipton, E. (In Press). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

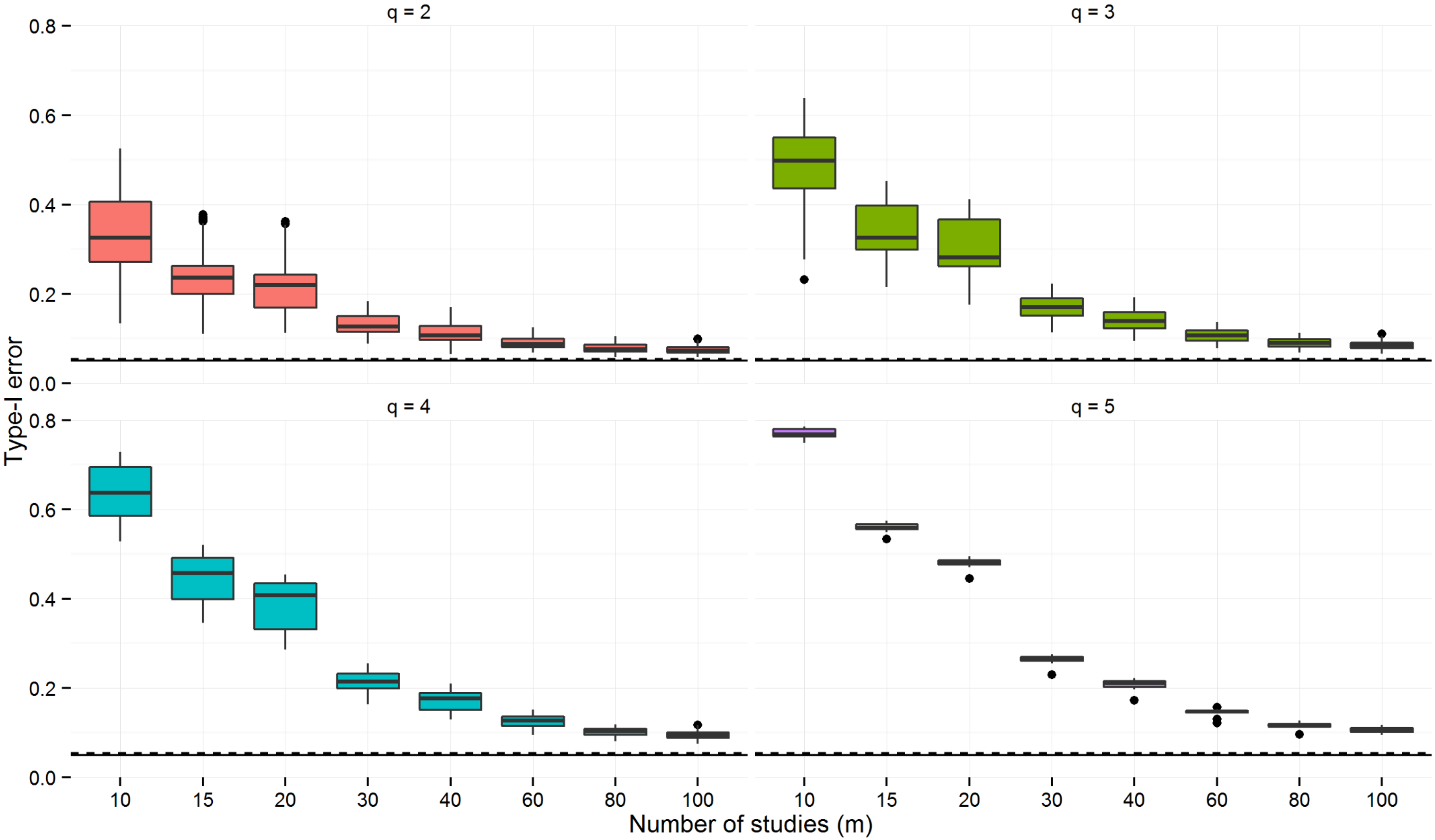
Wilson, S. J., Lipsey, M. W., Tanner-Smith, E., Huang, C. H., & Steinka-Fry, K. T. (2011). Dropout prevention and intervention programs: Effects on school completion and dropout among school-aged children and youth: A systematic review. *Campbell Systematic Reviews*, *7*(8).

Simulated Type-I error

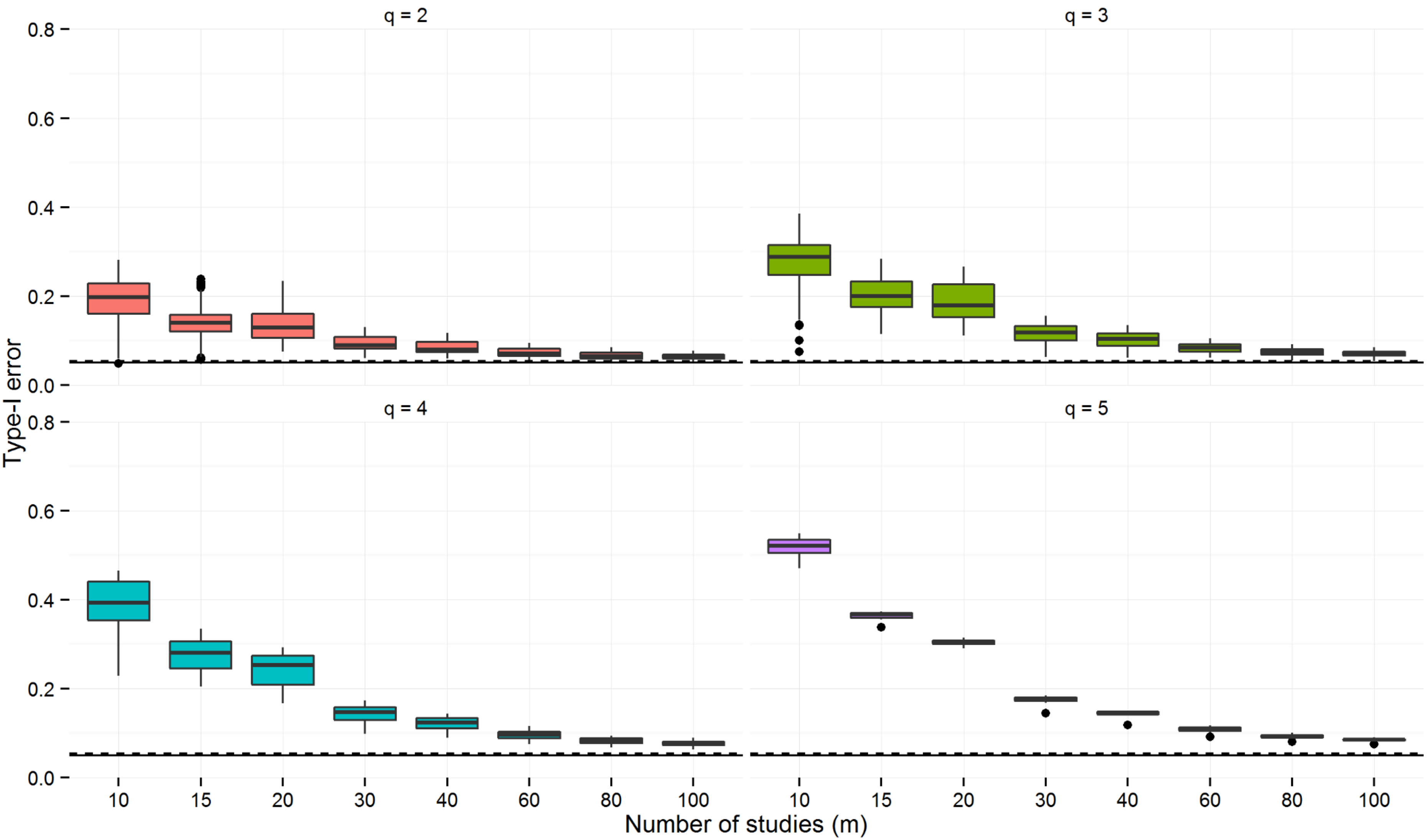
Moderator	q	χ^2 test		AHZ test	
		m = 32	*m* = 152	*m* = 32	*m* = 152
Study design (3 levels)	2	.278	.145	.073	.075
Outcome measure (4 levels)	3	.261	.155	.023	.046
Evaluator independence (4 levels)	3	.396	.175	.012	.051
Implementation quality (3 levels)	2	.248	.142	.048	.065
Program format (4 levels)	3	.383	.179	.044	.074

- Simulations based on design matrix of Wilson et al. (2011).
- *m* = 32 is the subset of 32 studies that report 3 or more effect sizes.
- Weights based on “hierarchical” model proposed by Hedges et al. (2010).
- 5000 replications.

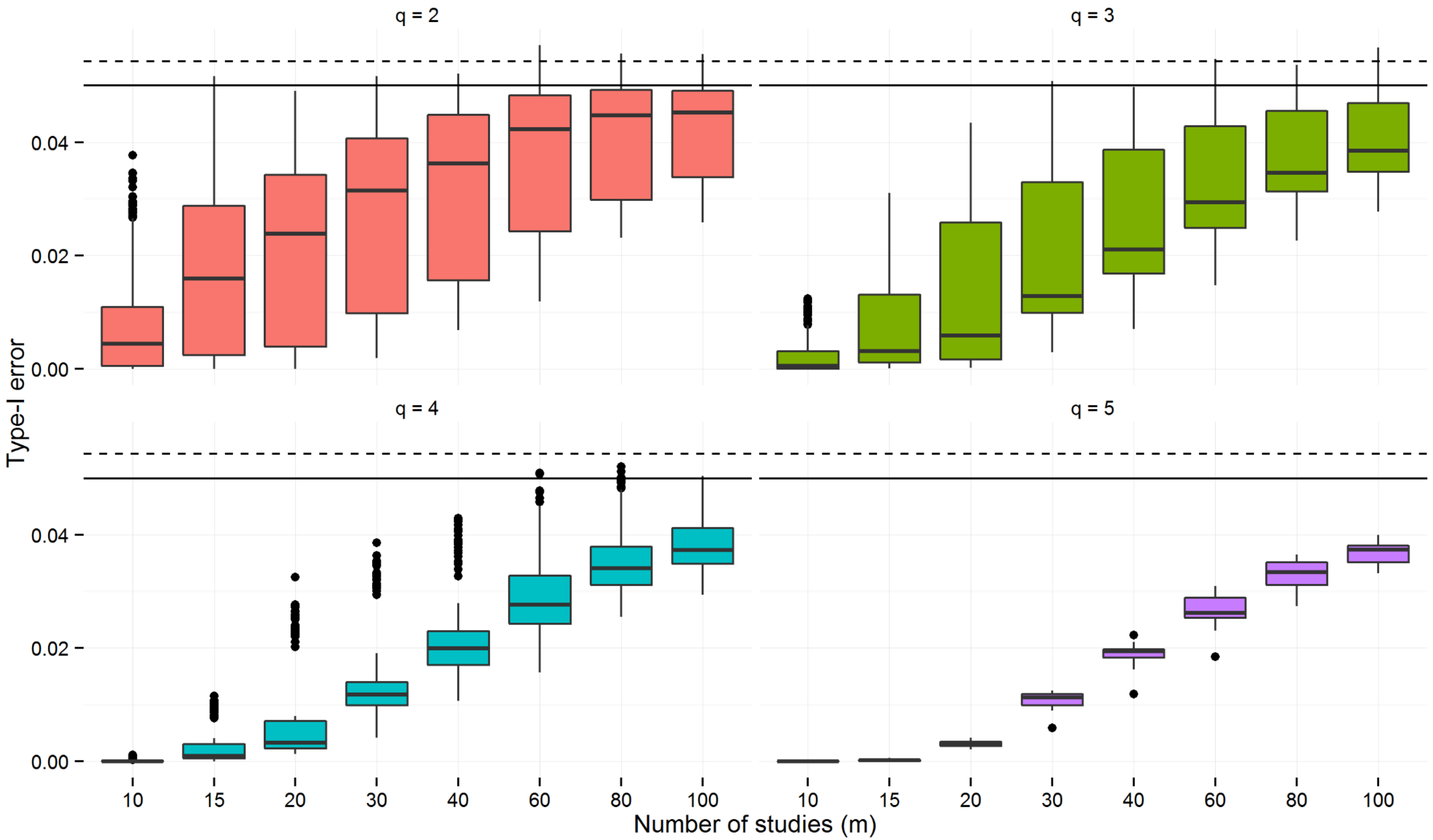
Simulated Type-I error of χ^2 test



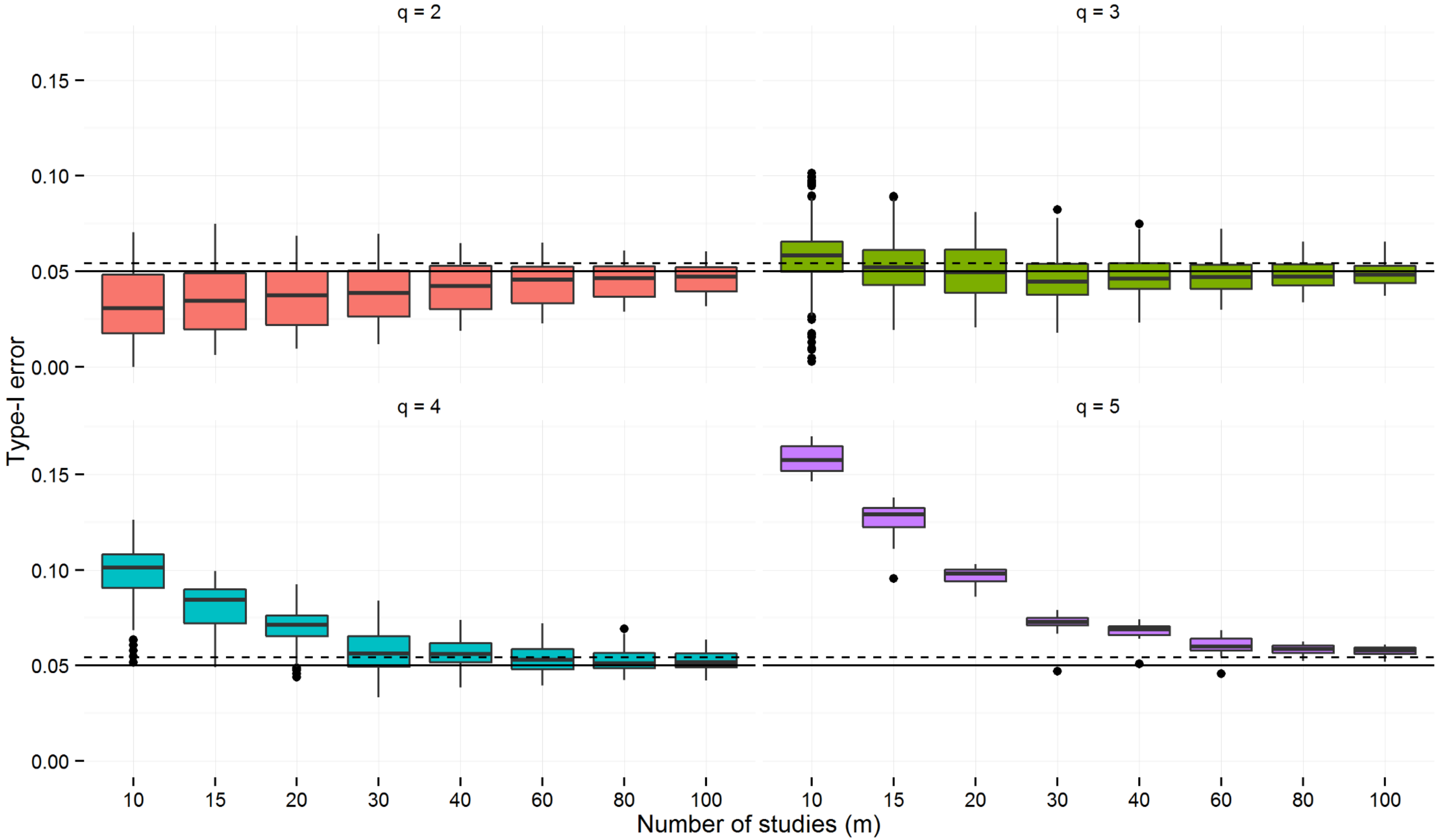
Simulated Type-I error of χ^2 test with bias-reduced linearization adjustment



Simulated Type-I error of AHZ test



Simulated Type-I error of EDT test



Bias-reduced linearization estimator

- McCaffrey, Bell, & Botts (2001) proposed a correction to $\mathbf{V}^R$ based on a working model for the error covariance structure.

- Suppose that weights are chosen to be inverse-variance under the working model. Then

$$\text{Var}(\mathbf{b}) = \left(\sum_{j=1}^{m} \mathbf{X}_j^T \mathbf{W}_j \mathbf{X}_j\right)^{-1} = \mathbf{M}.$$

- The corrected RVE is

$$\mathbf{V}^R = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{X}_j^T \mathbf{W}_j \mathbf{A}_j \mathbf{e}_j \mathbf{e}_j^T \mathbf{A}_j^T \mathbf{W}_j \mathbf{X}_j\right)\mathbf{M}$$

where the adjustment matrices $\mathbf{A}_1,\ldots,\mathbf{A}_m$ are chosen so that $E(\mathbf{V}^R) = \mathbf{M}$ when the working model is correct.